THE UNIVERSITY OF HULL

Investigating and Extending the Methods in Automated Opinion Analysis through Improvements in Phrase Based Analysis

being a Thesis submitted for the Degree of

Doctor of Philosophy

in the University of Hull

by

Amna Asmi,

MSc Computer Graphics Programming (University of Hull)

January 2015

Table of Contents

1 Chapter 1 – Introduction 1
1.1 Research Problem 4
1.2 Research Question 8
1.3 Aim and Objectives 8
1.4 Research Methodology and Thesis Structure 9
1.5 Research Contribution 12
2 Chapter 2 – Background and Literature Review 14
2.1 Information Overload 15
2.2 Information Retrieval (IR) 15
2.3 Web Mining 17
2.4 Text Mining 18
2.5 Opinion Analysis A Brief Introduction 22
2.6 Fact/Opinion 23
2.7 Opinion Analysis 24
2.7.1 Identification and Extraction of Opinion Oriented Data 25
2.7.2 Opinion Polarity and Degree of Polarity 27
2.7.3 Structure of Opinion in Textual Data 29
2.8 Level of Granularity 32
2.8.1 Document Level Opinion Analysis 32
2.8.2 Sentence Level Opinion Analysis 34
2.8.3 Clause Level Opinion Analysis 37
2.8.4 Phrase and Word Level Opinion Analysis 38
2.9 Existing Techniques for Opinion Analysis 39
2.9.1 Bag of Words (BoW) 40
2.9.2 Rules Based 42
2.9.3 Syntax 43
2.9.4 Machine Learning 43
2.9.5 Structural Opinion Extraction 44
2.10 Limitations in State-of-the-Art 46
2.11 Proposed Technique 48

2.12 Summary 49

3 Chapter 3 – Proposed Opinion Analysis Approach 51

3.1 Opinion Analysis Technique 51

3.2 Methodology 55

3.3 Analysis of Identified Research 55

3.3.1 Wilson et al. (2004) 55

3.3.2 Ding et al. (2008) 57

3.3.3 Thet et al. (2010) 58

3.4 Development of a Novel Opinion Analysis Approach 59

3.4.1 Research Design 60

3.4.2 Structure of Opinion 60

3.4.3 Proposed Opinion Analysis Approach 61

3.5 In Depth Analysis of Different Components and Decisions Made For Opinion Analysis 66

3.5.1 Granularity 66

3.5.2 Parsing 70

3.5.3 How to Resolve Clauses 75

3.5.4 The Scoring Technique 77

3.5.5 Generalized Scoring Method 86

3.5.6 Sentence Level Opinion Score Calculation 89

3.5.7 Other Issues 91

3.5.8 Limitations and Strengths of Proposed Opinion Analysis Approach 92

3.6 Summary 94

4 Chapter 4 – Background for Existing Resources and Frameworks 96

4.1 Existing Resources (NLP) (Based on Linguistic Analysis) 97

4.2 Lexical Resources 99

4.2.1 Automated/Semi-Automated Approaches for Generation of Lexical Resources 100

4.3 Existing Frameworks 104

4.3.1 Jin and Ho (2009) 105

4.3.2 Lloret et al. (2012) 108

4.3.3 Consoli et al. (2008) 111

4.4 Justification for a New Framework Development 112

5 Chapter 5 – Framework 114
5.1 Proposed Framework 114
5.1.1 Information Retrieval (IR) Stage 115
5.1.2 Pre-Processing 116
5.1.3 Opinion Analysis Stage 118
5.1.4 Corpus Generation 119
5.2 Summary 124
6 Chapter 6 – Evaluation Plan 126
6.1 Evaluation Process Followed in the State-of-the-Art 126
6.1.1 SentiFul 126
6.1.2 SentiWordNet 128
6.1.3 Thet et al. (2010) Clause Level Opinion Analysis 129
6.1.4 Critique 131
6.2 Evaluation Goals 136
6.3 Evaluation Plan 137
6.4 Core Components of the Evaluation Process 140
6.4.1 A Proof of Concept Prototype System 140
6.4.2 Dataset 142
6.4.3 Gold Standard Dataset 143
6.4.4 Participants 144
6.4.5 Ethics Permission 145
6.5 Evaluation Test Bed 145
6.6 Measurements 151
6.7 Limitations of Evaluation Plan 155
6.8 Summary 156
7 Chapter 7 – Evaluation 158
7.1 Evaluation Process 158
7.2 Analysis of Results 159
7.2.1 Analysis Based Upon Evaluation Goal 1 159
7.2.2 System P1 Generated Opinion Polarity over three scale 163
7.2.3 Recall 163
7.2.4 Precision 163
7.2.5 F-score 163

7.2.6 GS1 163
7.2.7 GS2 163
7.2.8 Discussion 166
7.2.9 Analysis Based Upon Evaluation Goal 2 170
7.2.10 Discussion 172
7.2.11 Analysis Based Upon Evaluation Goal 3 173
7.2.12 Analysis Based Upon Evaluation Goal 4 174
7.3 Limitations 176
8 Chapter 8 – Conclusion and Future Directions 179
8.1 Closely Related Recent Research 185
8.1.1 Similarities between Approaches Used 186
8.1.2 Differences between Approaches Used 187
8.2 Limitations of Current Research 189
8.3 Future Directions 190
8.3.1 Level of Granularity 191
8.3.2 Utilise the Diversity of User Generated Data 191
8.3.3 Utilisation and Evaluation 192
9 References 193
Appendix A 1
Existing Corpora 1
MPQA Opinion Corpus 1
JDPA Sentiment Corpus 3
SentiWordNet 4
WordNet 5
A brief over view of Existing resources and corpora 5
Appendix B 9
List of part-of-speech (POS) tags used in the Penn Treebank Project taken from Bies et al. (1995) and Proofread_Bot (n.d) 9
Clause Level 9
Phrase Level 9
Word level 10
Appendix C 12
50 Sentences Dataset 12

List of Figures

FIGURE 1-1: METHODOLOGY AND THESIS STRUCTURE 10

FIGURE 2-1: DATA MINING HIERARCHY 22

FIGURE 3-1: THE GRAPHICAL REPRESENTATION OF THE OPINION ANALYSIS PROCESS 65

FIGURE 3-2: AN EXAMPLE FOR SUBSTITUTION OF WORDS AND THEIR GRAMMATICAL CATEGORIES, BIRD ET AL. (2009) 67

FIGURE 3-3: HIERARCHY OF THE RELATIONS IN STANFORD PARSER, MARNEFFE ET AL. (2006)72

FIGURE 3-4: EXAMPLE SENTENCE 75

FIGURE 3-5: DEPENDENCY TREE BREAK UP OF A SENTENCE WITH NEGATION AND CONJUNCTION86

FIGURE 4-1: THE SYSTEM FRAMEWORK BY JIN AND HO (2009) 107

FIGURE 4-2: THE PROPOSED UNIFIED FRAMEWORK BY LLORET ET AL. (2012) 109

FIGURE 4-3: CEC MODEL PRESENTED BY CONSOLI ET AL. (2008) 111

FIGURE 5-1: BLOCK DIAGRAM FOR FRAMEWORK 115

FIGURE 5-2: PRE-PROCESSING PIPELINE 118

FIGURE 5-3: CORPUS DESIGN 125

FIGURE 6-1: COMPARISON OF OPINION ANALYSIS IN P1 AND P2 139

FIGURE 6-2: OUTLINE OF PROOF OF CONCEPT PROTOTYPE P1 & PROTOTYPE P2 142

FIGURE 6-3: PROCESS FOR PHASE 1 OF EVALUATION PROCESS 146

FIGURE 6-4: PROCESS FOR PHASE 2 OF EVALUATION PROCESS 148

FIGURE 6-5: PROCESS FOR PHASE 3 OF EVALUATION PROCESS 149

FIGURE 6-6: PROCESS FOR PHASE 4 OF EVALUATION PROCESS 150

List of Tables

TABLE 1-1: EXAMPLE SENTENCES 6

TABLE 3-1: EXAMPLE SENTENCES 52

TABLE 3-2: SYNTACTIC CLUES ESTABLISHED BY WILSON ET AL. (2004) 56

TABLE 3-3: TYPES OF CLAUSES IN ENGLISH SENTENCE 69

TABLE 3-4: IDENTIFIED GRAMMATICAL RELATIONS FOR NOUN PHASES 73

TABLE 3-5: CLAUSE LEVEL TAGS AND THEIR DESCRIPTION FOR PENN TREE BANK, BIES ET AL. (1995) 76

TABLE 3-6: RULES AND EXAMPLES FOR NOUN PHRASE CONTEXTUAL OPINION CALCULATION 79

TABLE 3-7: RULES AND EXAMPLES FOR VERB PHRASE CONTEXTUAL OPINION CALCULATION 80

TABLE 3-8: RULES AND EXAMPLES FOR OVER ALL PREDICATE CONTEXTUAL OPINION CALCULATION 81

TABLE 3-9: RULES AND EXAMPLES FOR SIMPLE CLAUSE CONTEXTUAL OPINION CALCULATION 82

TABLE 3-10: METHODS FOR COMPLEX CLAUSE CONTEXTUAL OPINION CALCULATION 83

TABLE 3-11: RULES SPECIFYING NEGATION 85

TABLE 3-12: GENERAL RULES FOR CONTEXTUAL OPINION CALCULATION 87

TABLE 3-13: EXAMPLES FOR NEGATION 87

TABLE 3-14: METHODS FOR NEGATIVE CONTEXTUAL OPINION CALCULATION 88

TABLE 3-15: EXAMPLES FOR POS AND OPINION SCORES 89

TABLE 4-1: DEFINITIONS OF ENTITY CATEGORIES AND EXAMPLES, JIN AND HO (2009) 108

TABLE 6-1: APPROACHES USED FOR EVALUATION OF OPINION BASED RESEARCH 135

TABLE 6-2: CONFUSION MATRIX 153

TABLE 7-1: SYNOPSIS OF EVALUATION PLAN 159

TABLE 7-2: CORRELATION FOR OPINION ORIENTATION: WHERE EXPERT 1, SYSOUT, AND EXPERT 2 ARE GS1, SYSTEM (P1) OUTPUT AND GS2 RESPECTIVELY 161

TABLE 7-3: RECALL, PRECISION AND F-SCORE FOR OPINION ORIENTATION OVER SEVEN SCALES162

TABLE 7-4: RECALL, PRECISION AND F-SCORE OVER THE OPINION POLARITY SCALE 163

TABLE 7-5: THE VALUES OF RECALL, PRECISION AND F-SCORE FOR MATCHES IN SUBJECTS, OBJECTS AND VERBS IN SENTENCES 164

TABLE 7-6: THE VALUES OF RECALL, PRECISION AND F-SCORE FOR MATCHES IN OPINION WORDS AND OPINION TOPICS 164

Table 7-7: The values of Recall, Precision and F-Score for matches in Subjects, Objects and Verbs in Sentences between both GSs 168

Table 7-8: The values of Recall, Precision and F-Score for matches in Opinion Words and Opinion Topics between both GSs 169

Table 7-9: Percentage results for Stage 2 172

Table 7-10: Search results 176

Table 1: A fragment of the MPQA subjectivity lexicon Potts (2011). 2

Acknowledgement

First and foremost I am thankful to Almighty **ALLAH**, the most gracious, merciful and beneficent who helps me in every moment of life. By whose grace I have been able to complete this research successfully.

Undertaking a PhD is truly a life changing and most challenging experience of my life. Towards the submission of my work I would like to take this opportunity to say that it would not have been possible without support and guidance of many people.

I would like to pay earnest and sincere gratitude to my supervisor Dr. Darren Mundy, for all the support and encouragement he has shown towards my work. Without his vision, advice and patience I would not have been able to reach to this stage of submission.

I would also like to thank Dr. Tanko Ishaya who was my supervisor during initial stage of my PhD, for long discussions. He was a great help and had provided guidance and suggestions for setting up goals of my PhD.

I am deeply grateful to Paul Warren and Michael Farrelly for annotating my evaluation dataset and helping me designing and setting Gold Standards for my research. Their guidance and understanding of language helped me a lot throughout evaluation process.

I would also like to thank my mates in WF1 for providing me friendship and help throughout the process. They were a source of joy, support and assistance whenever I needed it.

I am thankful to all my friends in England and back home for being moral support during all this time. They also have helped immensely in gathering data for my research.

Above all, a heartfelt thanks to my family, who have been my inspiration, encouragement, strength and driving force. Who unconditionally waited for me so long, without any complains. I owe everything to them. I wish I could show how much I love and appreciate them. I hope I can make them proud someday. They are reason of my being. No words are enough to show their contribution and my gratitude.

Abstract

Opinion analysis is an area of research which deals with the computational treatment of opinion statement and subjectivity in textual data. Opinion analysis has emerged over the past couple of decades as an active area of research, as it provides solutions to the issues raised by information overload. The problem of information overload has emerged with the advancements in communication technologies which gave rise to an exponential growth in user generated subjective data available online. Opinion analysis has a rich set of applications which are used to enable opportunities for organisations such as tracking user opinions about products, social issues in communities through to engagement in political participation etc.

The opinion analysis area shows hyperactivity in recent years and research at different levels of granularity has, and is being undertaken. However it is observed that there are limitations in the state-of-the-art, especially as dealing with the level of granularities on their own does not solve current research issues. Therefore a novel sentence level opinion analysis approach utilising clause and phrase level analysis is proposed. This approach uses linguistic and syntactic analysis of sentences to understand the interdependence of words within sentences, and further uses rule based analysis for phrase level analysis to calculate the opinion at each hierarchical structure of a sentence. The proposed opinion analysis approach requires lexical and contextual resources for implementation. In the context of this Thesis the approach is further presented as part of an extended unifying framework for opinion analysis resulting in the design and construction of a novel corpus. The above contributions to the field (approach, framework and corpus) are evaluated within the Thesis and are found to make improvements on existing limitations in the field, particularly with regards to opinion analysis automation. Further work is required in integrating a mechanism for greater word sense disambiguation and in lexical resource development.

List of Acronyms

Abbreviations	**Acronyms**
AI	Artificial Intelligence
AMALGAM	Automatic Mapping among Lexico-Grammatical Annotation Models
BoW	Bag of Words
CeC	Customer enterprise Customer
CORTEX	COndensés et Résumés de TEXte
EMOTIVE	Extracting the Meaning of Terse Information in a Geo-Visualisation of Emotion
GI	General Inquirer
GS	Gold Standard
HMM	Hidden Markov Model
HTML	HyperText Markup Language
ICE	International Corpus of English
IE	Information Extraction
IR	Information Retrieval
JDPA	The J.D. Power and Associates Corpus
LOB	Lancaster-Oslo/Bergen Corpus Tag-set
LSA	Latent Semantic Analysis
MPQA	Multi-Perspective Question Answering
NER	Name Entity Relationship
NLP	Natural Language Processing
OOV	out-of-vocabulary
P1	Proof of concept Prototype
P2	Experimental Prototype
POS	Parts of Speech
POW	Polytechnic of Wales Corpus
SD	Standard Deviation
SNS	Social Networking Sites
SR	System Results
WSD	Word Sense Disambiguation
WWW	World Wide Web
XML	Extensible Markup Language

Chapter 1 – Introduction

The recent development of Web 2.0 sites and applications has broadly influenced the Social Web (the Web of interaction and communication). This demonstrated a common willingness of people for sharing their lives, knowledge, experience and thoughts with the rest of the world, through micro-blogs, forums, wikis, reviews and Social Networking Sites (SNS) (Lloret et al., 2012). Through this tendency towards social collaboration and sharing including the provision of perspectives on a variety of items; a growth in SNS, newsgroups, blogs and forums is observed (Abdelrahman and Moustafa, 2010). Increasingly these networks are opened up to a wide variety of individuals and people have started making comments and opinions available to strangers (Pang and Lee, 2008; Liu, 2010).

The increase in the prominence, size, number and awareness of SNS, forums, micro-blogs and reviews has fashioned major changes in the ways people communicate and share their knowledge and emotions. This increase in subjective information on the Web has influenced social, political and economic behaviours worldwide (Lloret et al., 2012) and can have a very strong influence on an individual's decision making process (Zhu and Zhang, 2010; Töllinen et al., 2012). For example, Pang and Lee (2008) have reported that 73% - 87% of Internet users have accepted that opinions expressed on the Web strongly influence their buying decisions (Lloret et al., 2012).

Large organisations and companies are also interested in the opinions of individuals as the opinions help them in monitoring public perception about their products, services, policies, etc. This public perception helps companies in maintaining and strengthening their market competitiveness. Companies generally look to conduct market surveys, opinion polls or focused group discussions in order to collect public opinion about their products and/or services (Liu, 2010; Lloret et al., 2012). Political parties also conduct public polls and try to understand and capture trends in public opinion. People tend to seek and compare opinions of others in order to support their decisions. For example, before buying a new car many individuals tend to discuss it with their friends and family to seek for their advice and opinion. Similarly, people tend to read movie reviews in order to decide which movie they are going to watch. Many people are willing to contribute to online discussions, giving advice and expressing their opinions through participation in forums or product review sites. For example, these may include reviews about products they purchase or the movies they have seen.

Current information retrieval (IR) approaches (e.g. general search algorithms) used on the Web are unable to manage the large amount of user generated subjective information that is being produced. Social media networks producing large amounts of subjective comments about people or organisations generate problems for individuals and organisations in sifting the material to discover subjective comments. Given the current state of opinion analysis techniques there is a need to integrate solutions from the areas of Natural Language Processing (NLP) and IR to work together in order to improve the process. Improvements to the process could reduce the user processing time and the search space for decision making, given that when different areas of research combine the overall performance and capabilities of a system may be extended (Lloret et al., 2012).

Historically, similar challenges were apparent as the World Wide Web (WWW) began to grow in popularity during the early-mid 1990’s. The number of websites was growing and the Web faced a mushroom of growth. Manual directories like Yahoo were created in order to provide an adequate access method to global web information. However, such directories were expensive and had poor scalability. Therefore, researchers turned to the existing field of IR, (which at the time was mainly used by librarians and journalists) to develop computer algorithms which provided automated methods of populating such directories for everyday users of the Internet and Web.

Researchers brought theory and logic together, by combining different existing fields and research (IR and statistical techniques), and developed ways to retrieve and search documents online. Word occurrences can be readily indexed by computer systems and retrieval technology can be constructed on top of such indexes. However, such technology is not enough in order to overcome the recent issues regarding retrieval of subjective information on the Web, as the traditional search does not employ any techniques to solve deep problems in human language analysis and semantics. Therefore concepts from within other field’s like NLP, computational linguistics, IR and psychology are needed to be explored in order to develop a novel opinion analysis approach for subjective information in the written text.

Opinion analysis can be defined as the determination of sentiments, feelings, emotions and attitudes of a source with respect to some target topic within a written text (Bethard et al., 2005; Kim and Hovy, 2006a; Lu, 2010; Lloret et al., 2012). The term sentiment is defined as "a single component of emotion, denoting the subjective experience process" by Scherer (2005). Thet et al. (2010) has identified sentiments as opinions, attitudes, thoughts, judgments and emotions. Wilson et al. (2005) stated that opinions are private states of individuals, which cannot be open to objective observation(s) and verification(s). The terms 'opinion' and 'sentiment' are used interchangeably as synonyms (opinion analysis, sentiment analysis, opinion mining, sentiment classification and etc.) and are used recursively in definitions of opinion and sentiments. Different terms used for the opinion analysis process and their specific definitions and tasks associated to them are discussed in Section 2.5.

Many commercial applications, products and services based upon opinion analysis have been developed in recent years. These products and services differ from each other on a basis of their goals and capabilities. 'Tweetfeel.com' (http://www.tweetfeel.com) and 'socialmention.com' (http://www.socialmention.com) search and retrieve opinion from social media and summarise data returned as positive and negative. 'Tweetfeel.com' is a search tool only for Twitter and is observed to sometimes give inaccurate results and 'socialmention.com' is a commercial system, which is not fully tested. In addition, 'attensity.com' and 'lexalytics.com' offer more sophisticated services extracting opinions and their respective opinion topics (http://www.attensity.com/, http://www.lexalytics.com/). Both 'attensity.com' and 'lexalytics.com' are pioneers in the area of commercial systems in opinion analysis. 'Lexalytics.com' collaborates with big companies like Microsoft and has clients like Cisco, whereas 'attensity.com' is a customer management application working for companies like Airbus, and Lloyds Bank in a partnership with Yahoo.

More recent research has focused on the use of opinion analysis alongside a more detailed analysis of topics, for example; product based opinion analysis in order to determine the features and parts of products which need improvement along with the details of these improvements (Zhang and Liu, 2011). Stoyanov and Cardie (2006) believed that overall performance and capabilities of a system might increase by combining more than one approach together. Therefore, one of the main challenges in opinion analysis is to design and evaluate an approach, which can bring existing research (resources and techniques) from related fields together and improve the overall state-of-the-art.

1.1 Research Problem

The high level of semantic variability of natural language, in addition to differences in the personal understanding of human sentiments, attitudes and cultural backgrounds has made opinion analysis a complex task (Lloret et al., 2012). Written communication depends crucially on information about the perspectives and attitudes of the author and the perception of the reader, which can differ based on an individual's inference about the topic and information covered (Greene, 2007). The use of slangs, cultural contexts, sarcasm, ambiguous words and the inherently informal nature of web based social media messages can increase the challenge for automated opinion analysis (Branthwaite and Patterson, 2011; Töllinen et al., 2012). In addition, written text has its own limitations, for example, no additional cues (e.g. body language, voice tone, facial expressions, etc.) can be obtained, which makes the opinion analysis of written text more complicated. Sarcasm is another phenomenon which is very commonly used in social media, but is inherently difficult to analyse (Maynard and Greenwood, 2014), not just for automated analysis, but also for humans.

Researchers have been working in the area of opinion analysis of textual data since the 1990s (Abbasi et al., 2008). Opinion analysis has been undertaken at different levels of granularity of textual data (e.g., document level, sentence level, clause level, phrase level and word level). The following paragraphs outline the progression of research in this area leading towards the presentation of the principal research problem.

The research at document level includes the work of Pang et al. (2002), Turney (2002), Dave et al. (2003), and Pang and Lee (2005) among others. Opinion analysis at document level generally makes the assumption that the whole document has maintained the same opinion about any particular topic (Greene, 2007; Liu, 2010). This is why at document level most research has focused on the use of review based datasets. This particular focus is based on an assumption that constructed reviews are about one particular product or service, and that they hold one opinion holder’s opinion (the reviewer’s). However, opinion analysis even at document level can often involve a study of a collection of opinions about one or more object (s) their different features, sub-components and other aspects. Therefore, even in review based data a clear understanding of the object, its complementary objects, alternatives, features, and subcomponents is necessary to analyse, summarise, and present in the analysis.

Recent research has taken a more sophisticated approach to opinion analysis focusing primarily at sentence level instead of document level analysis (Hatzivassiloglou and Wiebe, 2000; Riloff and Wiebe, 2003; Kim and Hovy, 2004b; Wilson et al., 2004). Sentence level opinion analysis mainly uses syntactic and lexical techniques for opinion identification, and extraction in written text (Liu, 2010). Textual data is divided into sentences, words, idioms, phrases, Parts of Speech (POS) and their relationships with each other. Research in the area of opinion analysis at sentence level; takes the assumption that the whole sentence reflects only a single opinion at a time. However, this does not hold true as a sentence can have multiple clauses depicting different perspectives and meanings (Bloom, 2011).

Each individual sentence can contain independent or dependent clauses (further detailed in Section 3.5.1). Each clause at least consists of a subject and a predicate, where the subject is based on a noun phrase and the predicate is an arrangement of object, verb, or adverbial phrases and complement words. Clause level opinion analysis is a further refined and sophisticated mechanism for in-depth opinion analysis, to extract and examine opinions based upon specific opinion topics and/or aspects (Fiscus and Doddington, 2002; Thet et al., 2010). In addition, mechanisms are provided to help determine opinion orientation and opinion intensity at clause level. These mechanisms help in comparing the intensity of an opinion across different clauses within a sentence or between more than one sentences, based upon particular aspects and topics.

As an example, a dataset of restaurant reviews can be considered, where opinions expressed not only covers the overall opinion about a particular restaurant, but further detail opinion about different aspects like food quality, cleanliness, the environment, etc. Therefore taking an example statement from such a dataset; “I loved the environment but the food quality was really pathetic.”, opposing views are expressed about two aspects (environment and food quality) of the same restaurant. Such opposing and comparative opinions can easily be compared and analysed by dividing the sentence on the basis of clauses, identifying the opinion orientation and the opinion intensity based on the topics/aspects covered.

In addition to document level, sentence level and clause level opinion analysis methods researchers investigated phrase level (Tan et al., 2011) and word level (Neviarouskaya et al., 2011) opinion. However, sometimes phrases are not detailed enough to capture the opinion, and the opinion topic(s) within them (Wiebe and Riloff, 2005; Wilson et al., 2005) and analysis at the word level produced resources which are not able to provide an understanding of the interrelationship of words within a sentence. However, the research based upon word based opinion analysis has helped in the generation of lexical resources, which can be used across different levels of granularity. Words within a sentence can change each other’s meanings and the intensity of opinion associated with them (especially when used in different sequences). This highlights the need to understand the interconnectedness of words within a sentence and how they impact on each other.

For example, given the sentences in Table 1-1:

Table 1-1: Example Sentences

Sentence Number	Sentence
1	The movie is fairly good.
2	The movie is very good.
3	The movie is not good.
4	The movie is not outstanding.

The use of ‘fairly’ and ‘very’ in sentences 1 and 2 change the intensity of the sentences. Both these words are adverbs, which intensifies the meaning of the related words. In this case, the intensity of ‘very’ is higher than the intensity of ‘fairly’. In sentences, 3 and 4 ‘not’ may transform the polarity of the related word and can change the intensity. In the case of the above sentences ‘not’ limits the intensity of the sentences, in the case of sentence 3 it completely transforms the polarity. In the case of 4 there is an argument over whether the polarity is changed and to what extent.

There is a need for a fully automated in-depth opinion analysis approach, which can utilise dependency grammars (as used in NLP and described in Section 3.5.2) coupled with linguistic and syntactic analysis. The use of syntactic dependencies in the resolution of syntactic contexts has earlier been explored for English and Chinese languages (Lu, 2010; Wu et al., 2012). However, these works have only considered individual words, and there is scope to identify the noun or verb phrases for topic target identification and extraction (Bethard et al., 2005; Jijkoun et al., 2010; Lu, 2010).

There is also a need to structure opinions within the analysis process in a different way. The way proposed is an opinion structure of an <*opinionholder*|*opinion*|*opinion topic*> triplet, which is very similar to the parts of sentence structure <*subject*|*verb*|*object*> triplet used in NLP. This would help in handling opinions across different opinion topics especially in comparative sentences (Kessler et al., 2010; Somasundaran, 2010). However, there would be still a need to aggregate all the different opinions based upon opinion topics at the sentence level, given that a sentence can be thought to be the most basic, independent unit in written text. For example; Shaikh et al. (2008) have used a linguistic approach based upon verb frames to calculate opinion intensity scores at sentence level, but they have not calculated the opinion scores for multiple aspects within a sentence.

This Thesis proposes a novel sentence level opinion analysis approach, which utilises a unique algorithm providing a hierarchical breakdown of each individual sentence. The breakdown provides a level of automated granularity (through the design and generation of new corpus) not found before in existing research. This work provides analysis at word, phrase, and clause levels. In addition, an aggregation of the overall sentiment at sentence level is calculated based on the relevance to the overall topic of the document.

Existing frameworks for opinion analysis are reviewed (Consoli et al., 2008, Jin and Ho, 2009, Lloret et al., 2012). It is thought the novel approach proposed in this Thesis lead to a need to re-evaluate these frameworks to analyse whether the different stages in opinion analysis process can be brought together (i.e., IR, opinion analysis, opinion summarisation and opinion representation for later usability) into a unified process.

Therefore, a new unified framework for opinion analysis is presented as a further contribution of this Thesis.

1.2 Research Question

Since 2010, researchers have investigated the identification of phrases within opinion holding text as a mechanism to improve opinion analysis. However, opportunities remain to link current research in the area of NLP with opinion analysis. Given research identified in Chapter 2, there remains opportunities to explore how phrase level identification can be best utilised in the identification of opinion holding text. Therefore, the research questions posed within this Thesis are as follows:

- Are there improvements targeted at phrase level that can be made to existing state-of- the-art systems that can bring the process of automated opinion analysis closer to manual 'expert' performance levels?
- Does phrase level analysis provide opportunities for the identification of additional information that can be used to support opinion analysis?

1.3 Aim and Objectives

This Thesis aims to thoroughly investigate the current state-of-the-art in the area of opinion analysis towards proposing improvements based on the integration of approaches to phrase level analysis.

From the above aim, the objectives of the research can be broken down as:

O1: Identification and review of approaches for opinion analysis, opinion based corpus generation and the development of an understanding of already existent frameworks and corpora.

O2: Understanding of issues and limitations in existing tools and techniques for opinion analysis and existing corpora.

O3: Proposal of an effective opinion analysis approach based on an understanding of the issues and limitations identified.

O4: Integration of the proposed opinion analysis approach into a framework for the generation of a fully automated corpus.

O5: Design of an evaluation plan and the development of a software prototype model, which demonstrates the effectiveness of the opinion analysis approach, generate an automatic semantically intelligent corpus for opinion and topic analysis and test the reusability of the corpus.

O6: Design and development of a test bed to gather data according to a specific evaluation plan.

O7: Evaluation of the effectiveness of the proposed approach through the proof of concept prototype and data gathered from the generated test bed.

O8: Assessment of the reusability and retrieval effectiveness of the corpus.

O9: Critical evaluation of the achievements and contributions of the research against the initial research objectives, including the suggestion of further work.

1.4 Research Methodology and Thesis Structure

In order to attain the aim and objectives defined above, a clear research methodology is important. The research methodology is a methodical and structured approach which is chosen and adopted to solve the defined research problem (Kothari, 2008). The research methodology includes the selected research methods, questions asked, data collection practises utilised, and techniques used for analysis.

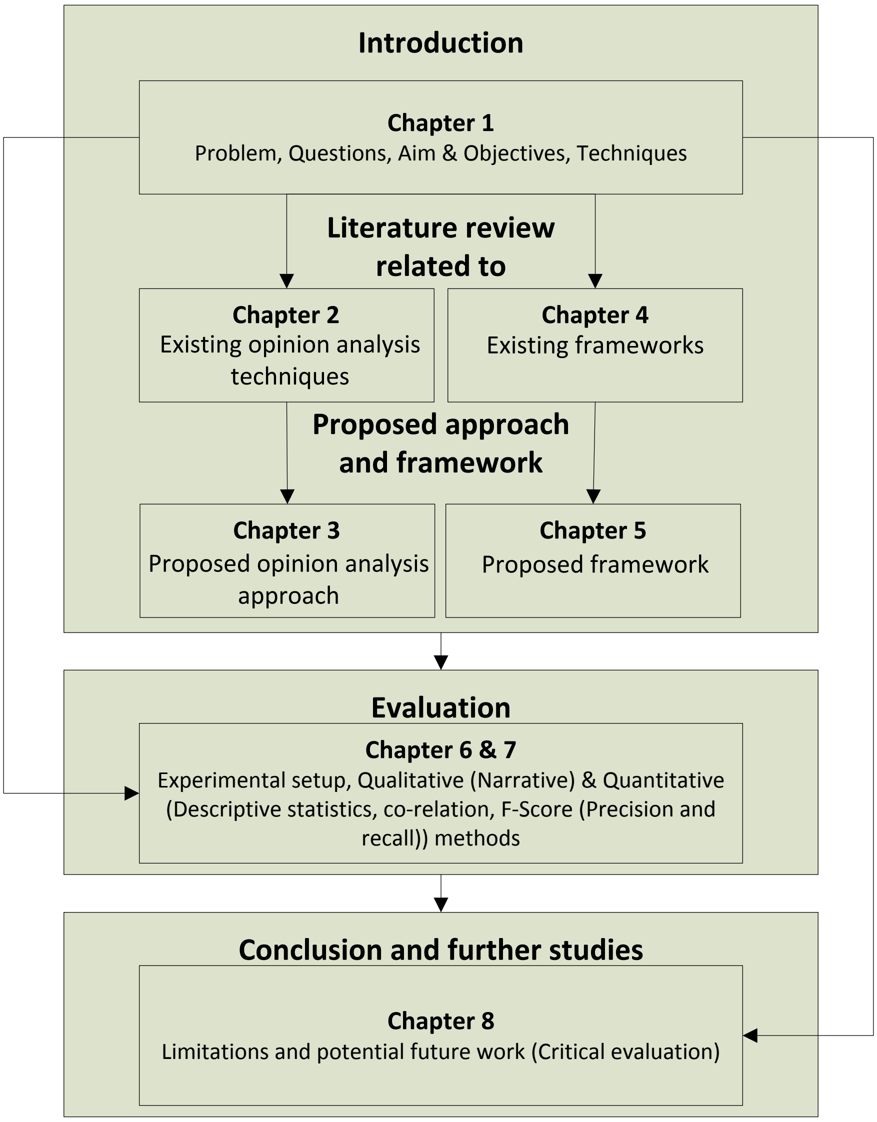

Introduction
Chapter 1
Problem, Questions, Aim & Objectives, Techniques
Literature review related to
Chapter 2
Existing opinion analysis techniques
Chapter 4
Existing frameworks
Proposed approach and framework
Chapter 3
Proposed opinion analysis approach
Chapter 5
Proposed framework
Evaluation
Chapter 6 & 7
Experimental setup, Qualitative (Narrative) & Quantitative (Descriptive statistics, co-relation, F-Score (Precision and recall)) methods
Conclusion and further studies
Chapter 8
Limitations and potential future work (Critical evaluation)

Figure 1-1: Methodology and Thesis Structure

The current research uses a mixed methodology, incorporating qualitative and quantitative research methods in the design and evaluation of the novel opinion analysis approach presented in this Thesis. The research follows a process where: 1) - A problem is identified and the related fields are reviewed through literature; 2) - This review identifies the limitations and gaps in the current state-of-the-art; 3) - The analysis of these gaps and limitations gives a basis for the design and proposal of an opinion analysis approach and framework; 4) - An evaluation mechanism is designed and the proposed approach and framework is evaluated; and 5) - Analysis of the results form a basis for recommendation of future work. The research methodology and structure of this Thesis is presented in Figure 1-1.

An introduction to the Thesis, the formulation of the problem, research questions and aim and objectives are provided based upon the literature (Chapter 1). The extended review of the literature is undertaken in order to identify and analyse the existing opinion analysis techniques and frameworks, their characteristics and limitations (Chapter 2) as well as some of the existing frameworks and resources employed for opinion analysis (Chapter 4).

Approaches to analyse the structure of sentences based on differing levels of granularity are investigated to gain an understanding of the limitations of current techniques used. The critique of the current state-of-the-art helps to identify the scope and limitations for the proposal of a novel approach for opinion analysis (Chapter 2). A short in-depth analysis of identified and selected research is carried out in order to establish requirements for the novel opinion analysis approach (Chapter 3 and Chapter 4). A design approach building on an understanding of current techniques in NLP, computational linguistics (literature review) and the set of requirements identified during in-depth analysis of selected research (Chapter 2), is used in the presentation of a novel opinion analysis approach (Chapter 3). Once existing frameworks are reviewed, there is a need to understand how they can be extended to provide better support for the presented approach (Chapter 5).

In order to evaluate the proposed theoretical designs an empirical research approach focusing on practical experimentation is undertaken. An experimental prototype is developed based upon the proposed solution(s). An evaluation plan is designed in order to verify and validate the solution(s) against the initial research questions, aim, and objectives of the research. The design of the evaluation process involves a series of steps including: determination of appropriate evaluation goals; generation of test datasets; construction of gold standards (GS); and the selection of evaluation metrics.

The selection of datasets to be annotated by human ‘expert’ users is undertaken based upon the perceived relevance of sentences to system specifications. In addition to the expert users, a sample group of other users are used to benchmark the systems performance. Both qualitative (narrative and comparative analysis) and quantitative measures (descriptive statistics, co-relation, f-score (recall and precision)) are chosen for evaluation (Chapter 6). These measures are analysed and investigated during the evaluation process (Chapter 7). The results of the evaluation in comparison with the research questions (Chapter 1) provides a basis for future work (Chapter 8).

As progress is made towards answering the research questions within the Thesis, time is taken to establish criteria to enable a comparative analysis between the approach proposed and existing state-of-the-art approaches. These criteria are provided as a mixture of qualitative and quantitative metrics enabling such analysis to occur. In the final chapter of the thesis, reference is made to the success or otherwise of the proposed approach having already evaluated these criteria.

1.5 Research Contribution

This Thesis through the analysis, design, creation and evaluation of a series of mechanisms to improve the opinion analysis process, provides contributions both theoretically and practically to the field of opinion analysis.

In theoretical terms, the work within the Thesis contributes the following:

- A novel opinion analysis approach based on clause level analysis

    The structure of an opinion in textual data often includes the opinion expressed and the opinion topic. It is observed from sentence structures that a clause is the smallest unit, which can hold an independent opinion. From these observations a novel clause level opinion analysis approach incorporating phrase level analysis for the accurate identification of opinion topics is proposed. This novel approach is evaluated through utilisation of existing research from the fields of NLP, computational linguistics and IR.

- The design of a unifying framework integrating the novel approach

    In order to integrate the novel opinion analysis approach into existing opinion analysis processes a framework is presented which unifies processes in existing state-of-the-art opinion analysis frameworks.

- A novel corpus design

  A novel automated corpus design is presented which enables greater reusability, and extensibility from current corpus designs.

This Thesis provides 2 peer reviewed academic contributions to the field of opinion analysis these are:

1. Asmi, A. and Ishaya, T. (2012a) A Framework for Automated Corpus Generation for Semantic Sentiment Analysis in: The World Congress on Engineering 2012. London, U.K.
2. Asmi, A. and Ishaya, T. (2012b) Negation Identification and Calculation in Sentiment Analysis. In: The Second International Conference on Advances in Information Mining and Management. Venice, Italy.

Further publications are planned from the research in the Thesis including:

- A journal paper discussing the overall evaluation of the algorithm highlighting the strengths and weakness of the approach.
- A conference and extended journal paper on the use of the algorithm in a practical social media based context, analysing the output of political party communication on their forums.

In practical terms the work within the Thesis contributes the following:

- A prototype system is designed and implemented based upon the proposed opinion analysis approach and the designed framework. This implementation is required for evaluation purposes.
- A corpus is generated in alignment with the corpus design proposed.

Chapter 2 – Background and Literature Review

The Web is a collection of billions of inter-connected and inter-related documents, designed and authored by millions of people (Xu et al., 2010; Zhang and Liu, 2011), which is experiencing a very rapid and exponential growth. This growth has attracted web users to use the Web as a medium for storing, distributing, broadcasting, and retrieving information especially to express and share opinions with the global community. Such opinion provision is enabled via various social tools such as user blogs, web forums, bulletin boards, and Social Networking Sites (SNS) e.g. Facebook, Twitter etc. As discussed in Chapter 1, the Social Web has influenced and changed the behaviour of individuals and can help to influence individual and business decision making processes. However, in order to better enable decision making, there is a need to continually improve solutions for the discovery and retrieval of relevant and accurate information (Kosala and Blockeel, 2000a) from the Web.

The first two Sections (2.1 and 2.2) of this chapter discuss the problem of information overload and outline how IR and Information Extraction (IE) methodologies can be utilised to bring solutions to this issue. Sections 2.3 and 2.4 outline the concepts of web mining and text mining, which broadly describe techniques for enabling the discovery and extraction of relevant information. The above four sections as a whole provide the background for the area of research of direct relevance to this Thesis that of opinion analysis.

Section 2.5 provides an introduction to opinion analysis as an area of text analysis, and includes a differentiation of terms used synonymously in the literature to describe the area. Opinion analysis involves a number of research challenges (opinion structure, opinion extraction, determination of polarity and the degree of polarity), these challenges are detailed, and an introduction is given to the state-of-the-art with respect to these in Section 2.7. Based upon the research challenges identified in the field of opinion analysis, it is observed that the understanding of the structure of an opinion and the extraction of an opinion from written text are the most crucial tasks. Opinion extraction can occur at multiple levels of granularity. These levels are further explained in Sections 2.8 and 2.9 with details provided for existing techniques which can be utilised to enable opinion extraction including an indication of the limitations of the existing state-of-the-art. Section 2.10 proposes a solution for sentence level opinion analysis using the identification of the linguistic hierarchy of clauses and phases and highlights the novelty of the proposed approach. This solution is then further explored in the Chapter 3 as the core contribution from this Thesis.

2.1 Information Overload

Information overload means the amount of information around us is growing beyond a reasonable threshold, and this leads to a need to put more effort into the processing and understanding of the information, otherwise, it might lead to poor decision making (Jacoby et al., 1974; Chen et al., 2009b). Improvements and changes made to communication technologies over the past decade have made it difficult to cope with the amount of content available to the user. The growth in the information is exponential and twofold in itself: too much information, and too many types of information (Gantz et al., 2009). With respect to decision making this can lead to problems such as users overlooking the information they are most interested in, not having the information they are most interested in presented to them, or the information may be presented in multiple types which makes it difficult to select the most appropriate resource. The challenge of information overload only continues to grow as further information is added to the Web.

There is a need to understand and prioritise information (Webster, 2010). Chen et al. (2009) have defined the online user as a human information processing system with limited processing capacity (Chen et al., 2009b). In addition, the online user is also limited in processing speed, and in various forms by their ability. However, technology at present is limited with respect to performing the types of complex decision making often made by the online user. Therefore, there is a need to continually improve methods used to extract relevant information for the human user in order to support the more complex decision making process.

2.2 Information Retrieval (IR)

The growth in the amount of information and document sources on the Web has amplified the need for effective ways to automate IR approaches (Lloret et al., 2012).

“IR is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)” (Manning et al., 2009). According to Manning, et al. (2009), the word ‘search’ tends to be used as a synonym of IR in modern times (Manning et al., 2009). However, search is not the same as IR; generally the process of retrieval can only be carried out once a search has been performed.

In order to search and filter objects, the earliest approach of retrieving relevant and required items is to create some data about the data, metadata. For example, storing shelf locations for books and presenting this information in a library catalogue or other index (Shalizi, 2009). Indexing in documents is similar as all documents can be pre-processed in order to extract all terms for indexing. Each term in the document can be collected with a pointer to where in the document it appears, this helps in searching across multiple documents to find the documents containing the required term(s) (Lloret et al., 2012). Concepts like metadata and indexes (or catalogues) are necessary for IR. Metadata provides data about the stored data itself and requires a knowledge representation method to save the data (an index or catalogue) along with the data to enable efficient data retrieval.

In a Web based context, web crawlers are used for indexing to efficiently capture web pages along with the link structures that interconnect them (Manning et al., 2009). IR needs to filter the information according to a specific query or the context of a search. Sheth and Maes (1993) explained information filtering as a process involving retrieval, routing, categorization, and extraction. Information filters are mediators between sources of information and their end-users.

Sheth and Maes (1993) have differentiated IR and information filtering. In their differentiation, IR is about the extraction of information from a large repository (Langville and Meyer, 2006), whereas, information filtering involves obtaining more relevant information from a stream of less relevant information (Webster, 2010). Currently, there are two approaches to information filtering:

- Searching with keywords to match metadata attached to information items;
- Browsing through the metadata extracted from information items (Webster, 2010).

In other words, information filtering is about the extraction of relevant or interesting information. The field of data mining can provide help in filtering information to enable more relevant information to be found and used across multiple different contexts. Research in the area of web mining (as a type of data mining) can be used in simplifying this process with web based data.

2.3 Web Mining

Web mining research comes at the crossroad of different research areas such as IR, IE, artificial intelligence (AI), NLP, databases, and machine learning (Kosala and Blockeel, 2000a). Given these roots in multiple research areas, approaches to web mining can be complex and multi-dimensional. Web mining itself emerged from the field of data mining, and therefore, can be further divided into subcategories based on the type of data being analysed. Initially, the focus of data mining was almost extensively on structured data, which could be found in databases, or organisational documentation (Gopal et al., 2011). Structured data whilst providing some initial research challenges (in the construction of filtering techniques and algorithms) can now be thought to be quite simple to manage, manipulate, and summarise in comparison to the types of unstructured data that can be found on the Web (Gopal et al., 2011). Web data can include: textual material, images, audio/video files, web server files, weblogs etc. (Gopal et al., 2011). The different types of web data can pose different research challenges with respect to mining information therefore, they can be divided into different fields as shown in Figure 2-1. For example, images in a web based context require image processing algorithms to help to analyse particular information of interest (shape recognition, facial recognition, digital signatures, watermarks, etc.). The analysis of web based textual data is of most relevance to this Thesis. In particular, the growing collection of user generated text in places such as forums, blogs, user review websites, etc. The focus of analysis on the data (primary and secondary) is to enable automated solution(s) to provide information about opinions within a specific text, which is a close match to that which could be perceived by a trained human user.

Primary data is composed of content (informational websites, forums, social networking websites, etc.) which is broadly based on text or multimedia data (Gopal et al., 2011). In fact, textual data is the most natural and most widely used method for storing and communicating information, 85-90% of information in organizations available on the Web is in text format (Kosala and Blockeel, 2000b; Mcknight, 2005) [NOTE: It is likely that these percentages have reduced over the past several years given the widespread propensity of video and graphical data as shown by Gupta and Lehal (2009), who placed this figure at 80%]. The other category of data is secondary data which is system generated data such as server logs, proxy server logs, cookies, session data, mouse movements, clicks, etc. Secondary data can help organisations to understand user journeys' on websites, network traffic generated from web navigation and other aspects of the Web experience (Stumme et al., 2006). In general, primary data can be thought to be less formalised, more chaotic but more readable, than the system generated data which is in general formally structured and machine oriented.

Textual data on the Web varies in form and structure. It can be more structured web pages or databases, or less structured user composed content such as emails, forum threads, social networking sites, chats or product/movie reviews etc. All this data is generally referred to as user generated content (Pang and Lee, 2008; Westerski, 2008). Whilst in many cases there are human moderators and administrators managing the forums, chat rooms, social networking groups, etc., most user generated content remains un-moderated. Users across the various places where user generated content is produced can be from anywhere, access anytime and use any language to post/respond. Therefore, in general user generated data does not follow any rules of structures including in some cases traditional language structures (Michael, 2004; Dey and Haque, 2008). For example, Twitter users post tweets on Twitter in any format and any language. In addition, tweets may be made up from forms of shortened dialogue e.g. Gr8 (great), F9 (fine) etc. Tweet content can define the language in which it is composed, however, there is no constraint about not using multiple languages in a single tweet. The complexity of user generated textual content and its inter-relatedness with research in NLP has presented text mining as an independent area of research.

2.4 Text Mining

Text mining is defined as "the non-trivial extraction of implicit, previously unknown, and potentially useful information from (large amount of) textual data" (Waegel, 2006). Data mining is about analysing information from pre-categorised fields, whereas text mining involves seeking information from disorganized, unformatted, and often fragmented collections of text (unstructured textual data means no HTML (HyperText Markup Language) or XML (Extensible Markup Language)) (Gupta and Lehal, 2009). Textual analysis mainly involves tasks around the identification and analysis of information from textual data through different analysis and mining techniques supported through a range of algorithms. Five core approaches to textual analysis are identified; four of these are described below, with the fifth approach, opinion analysis overviewed in Section 2.5 and Section 2.7.

The first core approach to textual analysis is text summarisation. The main aim of text summarisation is to reduce the length and details of the document while retaining the important points and overall meaning (Gupta and Lehal, 2009). Generally analysts try to achieve summarisation of text by reading and developing a complete understanding of the text before producing a concise representation. However, this process can be difficult to automate due to a limited understanding of the written text and semantics by machines. Therefore, different strategies to achieve this can be adopted. The identification and extraction of the most significant words is one of them. The significance of words is weighed by the frequency of words within textual data (Luhn, 1958; Gupta and Lehal, 2009). The markers of headings and subtopics can also help in the identification of significant words. More recent text summarisation tools extract the most significant sentences within documents. To ascertain which sentences are deemed to be most significant algorithms may focus on significant phrases such as ‘in conclusion’ or ‘the contribution’. Summarisation involves a reduction in the dimensionality of the document. This might be achieved by eliminating common words, extracting keywords, stemming, etc. A simple example could be de-constructing textual data into lists of keywords or indexes of contained values. A more complex example would be creating short textual descriptions of large blocks of textual data e.g. of a document, as a commercial solution Microsoft Word’s ‘AutoSummerise’ function is an example for this (Witten et al., 1999; Lihui and Lian, 2005).

Text categorisation is another core approach to textual analysis. Text categorisation identifies the main subjects which a document is addressing. In text categorisation a pre-defined hierarchical structure (of topics) is required (generally a form of ontology is used) in order to classify textual data (in the form of documents) into a taxonomical representation (Cohen, 1995; Hong and Weiss, 1999; Luo et al., 2011; Ur-Rahman and

Harding, 2011; Li et al., 2012). During computer based text categorisation a document is treated as a “Bag of Words” (BoW). BoW technique is further detailed in Section 2.9.1. The categorisation only counts words. Once the word count is complete a process of identification of the documents main topics take place based on the popularity of key words, discounting stop words e.g. ‘and’, ‘the’, ‘of’, ‘is’, etc. (Gupta and Lehal, 2009). Categorisation techniques mainly rely on a pre-defined taxonomy of topics and their relationships by identifying synonyms (words with similar meanings), related meaning, and the broad/narrow meaning of terms (Gupta and Lehal, 2009). For example, a document about a ‘dog’ may be categorised in a ‘dogs’ category, this could be structurally contained below documents with a category of ‘pet’.

The third core approach is concept categorisation: an extension to text categorisation. In this approach, textual data is analysed prior to categorisation, and concepts within the data are identified to enable a more accurate categorisation to take place (Rauber and Merkl, 1999; Ur-Rahman and Harding, 2011). Concept categorisation not only identifies the concepts within the written text, but also tries to relate documents to each other by identifying the linkages and relations between the concepts. Concept categorisation techniques are mainly used in the bio-medical field, where research has taken place to identify links between diseases and their treatments. So taking the text categorisation example above, during analysis of the concepts within the dog document, we may discover that the document focuses on a ‘dog movie’, therefore the category chosen would be more specific, perhaps ‘movie review’.

The fourth approach is factual analysis. In this approach, key terms and facts are identified from within the text. A key term is a piece of textual data determined to be important (e.g. a specific word such as ‘dog’, ‘cat’). A fact is a piece of textual data which is objective (e.g. “The dog was brown.”). Analysis takes place focused on linkages and relatedness between the terms identified through glossaries or other structures (Perrin and Petry, 2003; Ur-Rahman and Harding, 2011). An example of such analysis is concordance. Concordance is a standard study where one can look up a word and find references to all the passages, pages and chapters in the target work where the word appears.

In all the above approaches, the written word is considered as one of the most important and primary units of our communication. Therefore, text summarisation and text categorisation have focused on using the frequency of words in the identification of the most significant words within textual data. Once the most significant words have been identified, documents can be classified in relation to them (Luhn, 1958). However, this is not a sophisticated approach to textual analysis, as words and phrases create a context for each other, therefore patterns of words, the position of any word within a sentence, linguistic features, and punctuation can provide more specific syntactic and semantic information. Therefore, later research has emphasised the identification of syntactical units and patterns of words (Baxendale, 1958).

The focus on textual data patterns provides a better platform from which to achieve the identification of concepts for concept categorisation. Edmundson (1969) used the structure of documents and concept key words for automatic screening of documents. He was first to use lists of stop words in order to identify the words which can be classed as non-informative words. His idea of using keywords: the words with core significance in terms of frequency, is still widely used even in area of opinion analysis (Das and Martins, 2007). Later in the 1990s, more advanced and structured techniques were introduced for textual analysis. These techniques mainly involve structuring the input text on the basis of linguistic and statistical features, deriving patterns from the data, and interpreting the output. Often linguistic, statistical, and machine learning techniques are used in the process of automatically recognising and learning complex patterns in textual data, placing these in representative models and structuring the content (Das and Martins, 2007; Wajeed and Adilakshmi, 2009). This helps in the summarisation and visualisation of large blocks of textual data. Efforts have been made in the syntactic and semantic analysis of textual data using concept identification and heuristic based systems (Saggion and Lapalme, 2002).

Classification and categorisation of text can depend upon the usage and application of the text. Textual data could contain a complex taxonomy of topics, or a specific limited set of topics, or even sometimes contain a binary classification. In addition to complexity regarding taxonomy, the understanding of textual data can be dependent on the perspectives and attitudes brought to the text by the author or reader. Perspective, opinion, attitude, feelings are complex phenomenon. The final approach of textual analysis described in this Thesis focuses on the need to determine and classify the sentiment/opinion within a text, opinion analysis.

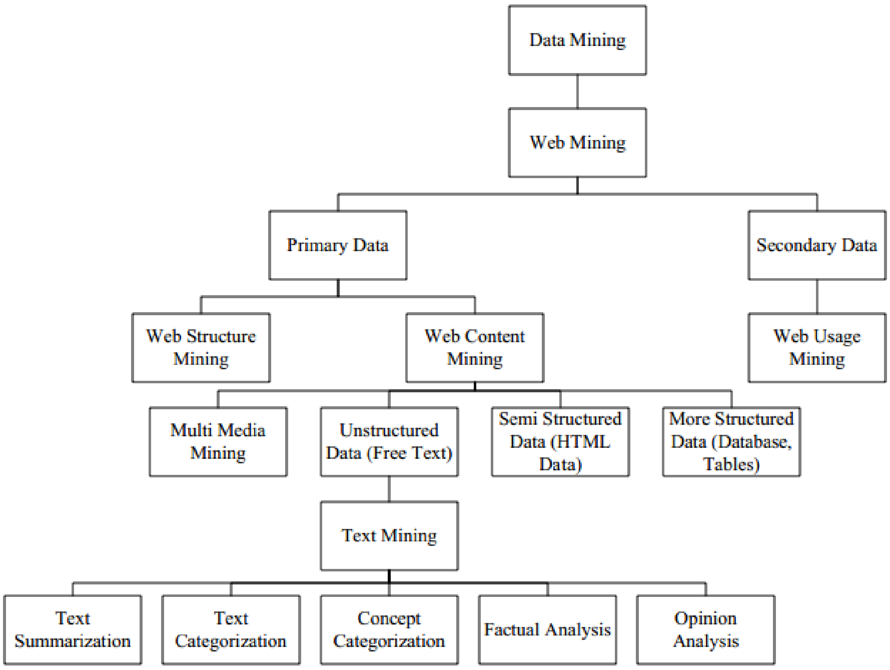


Figure 2-1: Data Mining Hierarchy

2.5 Opinion Analysis A Brief Introduction

Hearst (1992) and Wiebe (1994) were the first researchers to propose the idea of mining text, for direction based (positive or negative) data, for example; the mining of opinions, sentiments, affects or biases from the text (Abbasi et al., 2008). A multiplicity of names for research within the area of mining direction based information from textual data, have arisen over the past two decades (Pang and Lee, 2008; Somasundaran, 2010). Some of these names are listed below:

- “sentiment analysis” (Abbasi et al., 2008; Pang and Lee, 2008; Westerski, 2008; Liu, 2010; Hamouda et al., 2012)
- “opinion analysis” (Akkaya et al., 2011; Xu et al., 2011; Rill et al., 2012b)
- “opinion mining” (Esuli and Sebastiani, 2006a; Esuli, 2008; Bhuiyan et al., 2009; Binali et al., 2009; Liu, 2011)
- “sentiment extraction” (Bai et al., 2004; Goel and Hui, 2004)
- “sentiment classification” (Pang et al., 2002; Michael, 2004)
- “emotion detection”(Liscombe et al., 2005; Quan and Ren, 2010; Schuller et al., 2011)

- “affect analysis”(Abbasi, 2007; Neviarouskaya et al., 2007; Osherenko and Andr, 2007; Calvo and D'mello, 2010)
- “mood classification” (Mishne and Rijke, 2006b; Balog and Rijke, 2007)

The dictionary meaning of ‘sentiment’ and ‘opinion’ are fairly similar, and they are used interchangeably by many researchers who have identified them as near synonyms (Esuli, 2008). Although psychologists and social scientists study these matters and try to be precise in their definitions, Greene (2007) has identified that despite their careful treatment of terms such as ‘attitude’ they tend to use the word opinion . However, there are some distinct differences identified by Kim and Hovy (2004). An opinion suggests a perspective on a particular object/topic, whilst sentiment suggests aspects of point of view, emotions and desires (Kim and Hovy, 2004a; Lin et al., 2006; Greene, 2007).

Greene (2007) has described sentiment analysis as an investigation at document and sub-document level, whereas he describes opinion mining as particularly seeking and distinguishing between objective and subjective statements, at the level of clause, sentence or passage.

In addition to differences in word meaning there can also be differences in specific tasks associated with each of the given names. For example; according to Esuli (2008), ‘opinion classification’ generally focuses on classification of whole documents as positive or negative, whereas ‘affect analysis’ provides both the classification of expressed feelings (such as happiness, sadness, fear, excitement) and the polarity of the documents. The term ‘opinion analysis’ is chosen for this current research as it emphasises the investigation of polarity, intensity of polarity, and analysis of the properties associated with each opinion term. This gives an in-depth analysis of opinion instead of just classifications.

2.6 Fact/Opinion

Textual information can broadly be categorised into two main types: facts and opinions (Liu, 2010). Facts are objective expressions about entities, events and their properties. Opinions are subjective expressions describing the sentiments, appraisals, or feelings towards entities, events and their properties (Liu, 2010). For example, a fact could be “The temperature is 29°.”, an opinion could be “The air is quite warm.”, a mixture could be “The temperature is 29° and it feels quite warm.” Liu (2010)’s categorisation is very similar to how Hayek (1945) has defined and classified knowledge, as scientific and unscientific knowledge. Hayek defined ‘scientific knowledge’, as knowledge of facts, and defined ‘unscientific knowledge’, as “…the knowledge of the particular circumstances of time and place...special knowledge of circumstances of the fleeting moment, not known to others” . A very similar distinction is given by Polanyi (1966), he divided knowledge between ‘explicit’ and ‘tacit’ knowledge. Explicit knowledge is defined as knowledge that is or can be documented, easily communicated and interpreted. In contrast, tacit knowledge derives from experience and involvement in a specific context, and often only resides “in the heads” of individuals. Tacit knowledge includes individuals' beliefs, mental models, and viewpoints, and thus is inherently difficult to communicate (Polanyi, 1966). Opinion analysis is a field of research which attempts to construct systems that can determine the sentiment/opinion of an author from textual data.

In order for opinion analysis to be achieved manually, one has to find relevant resources, extract the related sentences, read the extracted sentences, summarise the extracted sentences, and organise the summarisation into a usable form (Bhuiyan et al., 2009). This process involves categorisation of textual data (at different levels of granularity e.g. word, sentence, document) according to a binary (positive/negative), ordinal (3 star, 4 star, etc.) or an interval based attribute (a degree of positivity e.g. a value in the range [-1, 1]) (Pang and Lee, 2008).

2.7 Opinion Analysis

Opinion analysis in written text is a field of research which attempts to construct systems that can determine the sentiment/opinion of an author from textual data. Research in the area of opinion analysis started with determining whether given textual data contained any form of expressed opinion (included some form of polarity) (Hatzivassiloglou and Mckeown, 1997; Hatzivassiloglou and Wiebe, 2000; Pang et al., 2002; Turney, 2002). Over the last ten years research challenges have shifted from the initial now less complex identification challenge, to more specific challenges related to the analysis of opinion related data. Challenges in this area include determination of the intensity of the opinion (degree of polarity) (Pang and Lee, 2008; Li and Wu, 2010), identification of the propagation of opinion related data over multiple sentences (Sarmento et al., 2009; Qiu et al., 2011), and comparative analysis of opinion oriented data (Jindal and Liu, 2006b; Liu, 2010; Wu et al., 2012). These and other challenges are further discussed below.

2.7.1 Identification and Extraction of Opinion Oriented Data

The main purpose of IE is to locate and extract the information of interest from unstructured text, based on a prescribed set of related concepts i.e. an extraction scenario defining why we are extracting information, and what is our target (Turmo et al., 2006). Traditionally information required for intelligent systems/knowledge based systems has been acquired manually with the help of domain experts.

Generally, the IE process involves a collection of the following steps: document pre-processing, syntactic parsing, semantic interpretation of parsed data, discourse analysis to provide semantic interpretations, and the organisation of output (generally in the form of a template) (Turmo et al., 2006; Bloom, 2011). Early automated IE systems used single words to represent a unit of textual data (in most cases a document). Instead of parsing the whole text, they just used to analyse the data and use a single word (generally a topic keyword) or in some cases a series of words to identify the whole document (Lehnert et al., 1991). In later systems, the analysis took place at the level of a sentence. However, the analysis focused primarily on indexing syntactic segments of noun phrases, verb phrases, subjects, objects, etc.; rather than parsing the whole document. Such parsing could have focused on constructing a parse tree from the complete text through the use of a syntactic grammar based approach (Lehnert et al., 1992).

In recent research, emphasis has been placed on automation of IE. In such systems, analysis is focused at multiple levels of granularity. Today’s approaches break down the extraction process and focus on textual data as a series of data elements. For each of these data elements an approach is selected to enable the identification of targeted information. Present research in the field is focused on pre-processing textual data to identify textual data patterns (e.g. a repeated word segment in a document or series of documents, a targeted group of words, for example “In conclusion”), recognition of core domain based features (e.g. in the analysis of restaurant review data targeting a feature such as food quality), resolution of co-references within the data (e.g. “Mary has a dog. She likes it so much.” – in this sentence ‘Mary’ and ‘She’ would need to be related and ‘Dog’ and ‘it’),relation prediction between extracted elements (e.g. a possession relationship in that Mary has a dog), and determining how to unify extracted information and binary relations into an organised representation (Bloom, 2011).

The identification and extraction of opinion(s) from textual data is a complicated task as while extracting the opinion from the data, the system needs to distinguish the role and relationship of entities contained within the text. For example, if extracting information about terrorism, there must be a distinction between the person who is the perpetrator, and the person who is the victim (Choi et al., 2005). Choi et al. (2006) presented a global inference approach in order to capture entities that express opinions and entities that denote sources of opinions . Their results improve when their system is incorporated with a semantic role labelling system. Another such approach was taken by Bathard et al. (2005), who tried to extract opinion propositions and opinion holders by using syntactic and lexical cues. Their results were preliminary, however, their focus on opinion clauses, and use of rich syntactic features, pointed to an important new direction in opinion detection. The clause level opinion analysis and syntax based opinion analysis techniques are further analysed in Section 2.8.3 and Section 2.9.3 respectively.

Although, many researchers have developed different opinion identification and extraction methods, the field of opinion analysis has not yet reached a level where decisions can be made about the identification of opinions on the basis of theories of cognition, affect and emotion without skilled human intervention. Theories in the above areas are being used in the formulation of annotation schemes for training datasets and in helping to define the dimensionality of labels for annotation and classification. Appraisal theory is one which is quite actively researched and worked in this context (Bloom, 2011; Balahur et al., 2012). Most of the existing work in opinion analysis has focused on three basic parts of appraisal expressions: attitude, evaluator and target. According to appraisal theory there are other parts of appraisal expressions that can be explored like subordinates, aspects and processes (Hunston and Sinclair, 2000). However the interest of this Thesis only lies with understanding and analysing the core parts of appraisal expression and their further in-depth analysis. The core parts of appraisal expression (i.e., attitude, evaluator and target) map to the structure of opinion (i.e. opinion, opinion holder and opinion topic). This structure of opinion (opinion,

opinion holder and opinion topic) properly maps onto the parts of a sentence (Subject, Object and Verb) which is employed for clause based opinion analysis as proposed in research presented in this Thesis.

2.7.2 Opinion Polarity and Degree of Polarity

The task of labelling any document/sentence/word as positive/negative either in two opposing classes or assigning a value on a continuum between two of these values for positive/negative is called polarity determination. Analysis of the opinion in text is similar to the textual classification process, where words are classified into one of two classes (positive/negative). This binary decision task is named as sentiment classification in the literature, Pang and Lee (2008) have explained sentiment classification as broadly referring to binary categorisation, multiclass categorisation, or ranking. Polarity determination involves locating a value for the opinion on a scale. Research in this area has generally concentrated on binary classification mechanisms for example: thumbs up/thumbs down for movie review data (Pang et al., 2002); likely to win verses unlikely to win from election data (Kim and Hovy, 2007); positive or negative sentiment from 1987 Wall Street Journal corpus (Hatzivassiloglou and Mckeown, 1997).

The main problem with binary classification systems is the lack of insight this gives into the degree of polarity or strength of opinion e.g. in simply classifying textual data as one extreme or the other. Further research has focused on classifying opinion polarity on an ordinal scale, for example; one to five stars for review based data (where one star means low quality of movie and five star means the movie is rated as a good quality movie) (Pang et al., 2002). Whilst this improves the granularity of the degree of polarity value, it can still be thought to provide a limited range of potentials (Esuli and Sebastiani, 2006b; Li and Wu, 2010).

Therefore, the most sophisticated way at present for determination of degree of polarity is through a real scale. A real scale being one where the maximum positive is represented, the maximum negative is represented and all the values in between, including the neutral space. This is often represented as [-1, 1] (Esuli, 2008). The process of determining a real value and the process of determining a neutral state is much more complex than binary identification or ordinal classification.

The determination of neutrality is always difficult as it may mean a number of things (Demars and Erwin, 2005) based upon different levels of granularity and the definition of opinion structure. At document level this could mean a document has a balanced number of positive/negative statements. At sentence level it may mean that there are equal numbers of positive/negative words. At a word level it may mean that a word has no significant polarity ('indifferent', 'neutral', 'impartial'). Cabral and Hortacsu (2004) have observed very interesting results about the perception of users regarding neutral comments on eBay. They have found that users consider neutral feedback as much closer to negative feedback than positive . Neutral opinions are also discussed in Section 2.8.2.

There are more refined and fine grained opinion structures than determining the polarity and polarity strengths, these fine grained structures can provide opinion and emotion based scores on a real number scale e.g., Affect database and SentiFul are designed based upon nine emotions ('anger', 'disgust', 'fear', 'guilt', 'interest', 'joy', 'sadness', 'shame', and 'surprise') and each emotion is assigned a numeric value on the scale of 0.0 to 1.0 (Neviarouskaya et al., 2007; Neviarouskaya et al., 2009). These numeric values across nine scales are represented into the vector space representing the emotion across each word. There is a need to further refine the structure of opinion and researchers are employing theories from psychology and cognitive sciences for this purpose (http://www.saaip.org/). A similar approach using a set of eight emotions, (i.e. six of the emotions defined by Ekman in 1985: Anger, Disgust, Fear, Happiness, Sadness and Surprise, plus Shame, and Confusion) has been proposed by Sykora et al. (2013a). They used an ontology engineering approach to generate resources.

Further related work exists in relation to the above where researchers try to find and capture reasons for positive or negative comments from the written text. For example in product review data this sort of research gives a better insight as to why any customer gave a positive or negative comment about the product and what they expected with respect to that particular aspect. Kim and Hovy (2006) developed a technique to extract opinion as well as the reasons behind the opinion expressed. They have used negative comments and their associated opinion topics to identify the features of products which might need improvements. In this way, they have identified on a basis of assumption that customers write negative comments about features and indirectly suggest improvements for them (Kim and Hovy, 2006b). Another example is the work by Niu et al. (2005) where they have determined the outcomes (improvement of health or death) from medical texts, on the basis of degree of positivity.

Aspect is also a key term to consider. Aspect means the identification of a specific set of features about textual data in specific contexts. People express their opinion about different entities, each entity can have multiple aspects (features/attributes), and people can express their opinion about each of these aspects instead of expressing opinions about an entity as a whole. For example, in restaurant review data, an aspect may be the quality of the food or the cleanliness of the environment. These aspects are difficult to determine from less structured text, often therefore textual structures are put in place to better enable this analysis (e.g. a questionnaire format or some specific tagged elements). Snyder and Barzilay (2007) in their research, analysed restaurant review data with a focus on better understanding the aspects within the data. They used an ordinal ranking process (1-5) to provide a value for the degree of polarity across three aspects: food quality, service and ambiance. Similarly, Mishne and Rijke (2006) have used multiple aspects to predict the moods of blog authors. They have compared the mood of written blog posts with a specified “current mood”, which is selected from a list by the post author when composing the post. The use of aspect based opinion analysis can be helpful and with a clear instruction and list of relative aspects the quality of annotation for machine learning and/or evaluation of opinion analysis systems can be improved.

2.7.3 Structure of Opinion in Textual Data

There are different theories of affect and emotion which can be used in opinion analysis especially in order to understand an opinion and its composition. There are overlaps in understanding what an opinion comprises of, and how it can be differentiated from sentiments as described earlier in this chapter. However there is a need for examination of emotion theories in relation to computing practices for effective opinion analysis (Calvo and D'mello, 2010). Some of the theories of affect and emotion are stated below:

- Scherer's (1984) typology of affective states, presents a series of situations which have the potential to raise particular emotions and emotional responses. The typology also identified and analysed differences between emotion, mood and attitude.

- Ekman (1985) from an analysis of facial expressions provides a different perspective in his work which proposes that there are six basic emotions: surprise; happiness; anger; fear; disgust; and sadness.
- Scherer (1986) proposed that emotional states could be predicted through changes in acoustic parameters, and vocal cues. A large amount of research and models in the subject domain of psychology and cognition are based upon gestures, body language and vocal cues. Written text has limited information as it misses any such cues (gestures, body language and vocal cues).
- Pultchik (2002) developed a wheel of emotions which provides a visual representation (a 3D visualisation) of how emotions relate to one another.

Affect and/or emotion detection is very challenging task as emotions are conceptual constructs which cannot directly be measured and are expressed and experienced with great variations (Calvo and D'mello, 2010). The first researchers who integrated emotion with written text were anthropologists and social psychologists, in their research to find the similarities among communication of people from different cultures (Lutz and White, 1986; Calvo and D'mello, 2010). Theories such as those above (and many others) provide ways in which to classify different aspects, shades of emotion, and/or affect. However, in their most basic form many of these theories can be mapped directly onto a simple classification system, like the one used in the Affect database (Neviarouskaya et al., 2007). The Affect database has a basis of nine emotions. For each word in a defined list of opinion oriented words an associated numerical weight is assigned for each of the emotional states, this gives a multi-dimensional vector based value e.g.; for the word regret ['guilt:0.2', 'sadness:0.1', all other emotions 0] (Neviarouskaya et al., 2007). In an extension of the work, SentiFul (Neviarouskaya et al., 2009), each of the nine emotions are also mapped onto positive and negative values.

All of the above emotion and affect theories have used different classifications, different cues for identification, and different mechanisms for the extraction of opinion, emotion or affect. Many of those cues cannot be identified in written text, for example, there no identification of voice quality, use of facial gestures, etc. in written text. Therefore, the presence of these theories has brought forward limitations in the analysis of opinions within written text.

In order to understand the structure of opinion in the written text, it is necessary to understand a few of other basic concepts.

Opinion is basically the expression of positive/negative view, feeling, attitude and/or emotion about an object by an opinion holder, which may or may not be true.

Object is any product, service, individual, event, topic, feature or part of feature and/or sub- component about which an opinion is expressed. This makes the object a target of opinion.

Opinion is expressed by a subject, person, or even organisation; it is called the opinion of an **opinion holder**.

The structure of opinion stated above is more straightforward and can easily be found in statements which are directly expressing opinion. The direct opinion is simple to understand and identification of object, feature of object, opinion orientation, opinion holder and polarity and strength of opinion is easy and direct (Liu, 2010). For example, "I like Vanilla ice cream." Ice cream – object, Vanilla – feature (Flavour of ice cream), like – positive opinion, I – opinion holder and strength of opinion is mild.

However the expression of an opinion in written text can be not so direct and explicit, for example, comparative opinion. A comparative opinion is expressed by giving the relationship between two or sometimes more than two objects and/or features of objects. These relationships can be similarities, differences or preferences, etc. Mostly comparative or superlative forms of adjectives or adverbs are used to express comparative opinion (Jindal and Liu, 2006c).

These comparative sentences have many different types however they could be gradable in terms of positive and negative opinions. For example: "Samsung S3 has big display screen, but iPhone 5 has relatively smaller." and "Coke tastes different than Pepsi". In this example the bigger screen display of Samsung S3 does not classify as positive or negative feature, it can be one's personal choice. The example of gradable comparison is; "The picture quality of Nikon is better than that of Sony." An example of equality is; "The picture quality of Nikon is as good as Sony."

Opinion identification and extraction of comparative sentences can be more complex as the task is divided into firstly, the identification of comparison structures and secondly, the identification of opinion (Liu, 2010). The opinion structure here can get really complex as there can be more than one opinion, multiple opinion holders, and objects and a multiplicity of relationships between them which can be important to identify (Jindal and Liu, 2006b).

2.8 Level of Granularity

The level of granularity relates to decisions regarding the unit of analysis for opinion analysis. Decisions regarding this can vary from document, paragraph, sentence, and phrase/word level. Determining the level of granularity has an impact on the decision about which resources (corpora, dictionaries, lists, etc.) should be used in opinion analysis. Different levels of granularities for opinion analysis are further discussed below.

2.8.1 Document Level Opinion Analysis

Document level opinion analysis aims to discover the orientation of opinions expressed in a whole document (Pang et al., 2002; Turney, 2002; Dave et al., 2003; Riloff and Wiebe, 2003; Pang and Lee, 2005; Ku and Chen, 2007). It seeks to determine if the document indicates an opinion in favour of the topic of discussion, or in opposition (Greene, 2007). It is based on the assumption that the document has opinions from a single opinion holder, and about a single object. Therefore, this technique is usually used for the analysis of review based data, as reviews generally hold opinions about a single object (e.g. product/movie), and are often only based on the opinions and experiences of a single reviewer. Document level analysis of opinion is a less effective opinion analysis approach when in-depth analysis of text is required (Thet et al., 2010).

Opinion analysis at document level takes place in a similar way to classic topic based text classification, where topic related words from a list of pre-definite topics e.g.; sports, science, movies, etc. are important. Whereas, in opinion analysis all opinion oriented words should be identified, i.e. the words which indicate positive or negative opinions, for example, ‘excellent’, ‘bad’, ‘amazing’, and ‘pathetic’, etc.

Machine learning approaches, for example; naïve Bayesian and support vector machine are very readily applied in opinion analysis at document level (Pang et al., 2002; Kennedy and Inkpen, 2006; Hamouda et al., 2012), as they introduce the capability in the system to adapt to the changing input, give general machine understandable rules, as well as a possibility to compare and measure the degree of similarity of the input with these rules (Boiy and Moens, 2009). In supervised machine learning the manually annotated training data set is given in order to train the system and the system can continue to learn and update the rules based on the input and training dataset (Pang et al., 2002). The main task of opinion analysis using machine learning techniques is to engineer a suitable set of features, which can be learned through rules (Liu, 2010). A few examples of features used in opinion analysis are listed below and are explained further in Section 2.9.

Terms and their frequencies (BoW)

BoW is also mentioned in Section 2.4 and detailed in Section 2.9.1, is a classification using the representation of text as individual words or word counts as frequencies (tf-idf, weighting scheme from IR (Hamouda et al., 2012)),or presence or the absence of certain opinion based words (Wilson et al., 2005). The usage of terms and their frequencies are very commonly used in traditional topic based text classification, and have worked quite well in document level opinion analysis as well.

Parts of Speech (POS) tags

The use of the information related to POS of the opinion words is also very common. It helped in automated analysis of the polarity of opinion based words and generation of lexical resources (Hatzivassiloglou and Mckeown, 1997). Early researchers considered adjectives as important indicators of subjective and opinion based information. More latterly nouns, verbs and adverbs have also been used to produce machine learning algorithms (Riloff et al., 2005; Chesley et al., 2006; Benamara et al., 2007; Subrahmanian and Reforgiato, 2008) to varying success.

Opinion words and phrases

Apart from opinion based words, phrases and idioms are also instrumental in opinion analysis; for example, “This has cost me an arm and a leg.” Here a simple word count or other word based analysis can miss a lot of information and may change the perception of a sentence. Therefore phrase based resources are developed and employed in opinion analysis (Rill et al., 2012b).

Syntactic dependency

Word dependency based features are also used by many researchers for machine learning algorithms using dependency relations. There are many words which might not have an opinion polarity and strength associated with them, however, when these words are used in combination with other opinion based words they can change the opinion polarity and strength of those words. Example words include: ‘very’; ‘hardly’; etc. These words are known as valance shifters or modifiers (Kennedy and Inkpen, 2006; Li and Wu, 2010). Information about modifiers and valance shifters allows for the identification of the words that are modified by these.

Document level opinion analysis is less effective for in-depth opinion analysis even for review based data, as different types of reviews, such as critic reviews, blog posts, message posts on discussion boards, and social networking site posts can have different characteristics. Document level sentiment analysis using a BoW approach may be suitable for some genres with long texts, but it is not ideal for other genres having rather short texts (Na et al., 2010; Thet et al., 2010). A new genre of data, web based textual data (tweets and chats): mostly shorter in length generally less than 140 characters in length, has emerged very widely. Therefore there is a substantial focus on understanding this web based user generated data (Bollen et al., 2011). Researchers have created more sophisticated methods like exploring documents at further refined levels of granularity e.g. sentence level and beyond.

2.8.2 Sentence Level Opinion Analysis

Sentence level opinion analysis, recognises the presence, polarity and intensity of positive/negative sentences within a document (Hu and Liu, 2004; Ding et al., 2008; Ganapathibhotla and Liu, 2008; Shaikh et al., 2008). Sentence level opinion analysis mainly uses syntactic and lexical techniques for opinion identification, and extraction in written text (Liu, 2010). Textual data is divided into sentences, words, idioms and phrases, POS and their relationships with each other. However, researchers in the opinion analysis field at sentence level, take the assumption that the whole sentence reflects only a single opinion from a single opinion holder at a time. However, this does not hold true as each sentence can have multiple clauses depicting different meanings (Bloom, 2011).

The opinion analysis at sentence level can be divided into two sub tasks.

1. Determination of subjectivity of the sentence
2. Determination of opinion polarity (positive/negative) of the sentence, if the sentence was subjective.

The decision whether a document/sentence/word contains opinionated information or not is defined as subjectivity. Opinion analysis starts with the assumption that textual data contains some opinion, and analysis is undertaken to understand the polarity and degree of polarity. Therefore, subjectivity detection is a very important step in opinion analysis especially at sentence level. The detection of subjectivity in a sentence provides evidence that the sentence is an opinion, and not a fact. After determining the existence of an opinion the polarity and strength of the opinion can be calculated (Wilson, 2008; Gyamfi et al., 2009). According to Hatzivassiloglou and Wiebe (2000) the presence of adjectives provides good support for the determination that a sentence is subjective.

Wiebe and Riloff (2005) have used screening for objective sentences in textual data, in order to narrow down the amount of text to be analysed for automated opinion analysis. However, in screening for and/or ignoring objective sentences individuals may be missing important opinion related content. For example; Wilson et al. (2004) argue from their clause based opinion analysis that the absence of opinion might mean the presence of a neutral opinion e.g.; a balanced, a mediocre, or a so-so perspective. This increases the level of difficulty of subjectivity/objectivity classification, as this raises the need for further classification of the non-positive/negative (neutral) sentences. Mihalcea et al. (2007), summarise that, “the problem of distinguishing subjective versus objective instances has often proved to be more difficult than subsequent polarity classification, so improvements in subjectivity classification promise to positively impact sentiment classification” . The overall opinion orientation even after adding the subjectivity analysis cannot guarantee that the author had maintained the same perspective throughout the document. There might be different opinion orientations about different features of products, movies, political stances, etc.

Hu and Liu (2004) have performed opinion analysis and provided a visualisation of this material for customer review data. They have extracted features of the products and presented the opinions associated to them as positive/negative. Their technique only classes positive/negative opinions, if the opinion is explicit. No pronoun resolution or intensity calculation of opinion is employed. Yi et al. (2003) have proposed a method to extract opinion from online review and news articles data. They have extracted opinions related to specific subjects. Their identification of opinion topic is based upon extraction of noun phrases. Their technique of manually designed pattern matching has not gained much popularity. However, their idea for extracting opinion based upon topic-feature level analysis has gained much popularity (Hu and Liu, 2004; Popescu and Etzioni, 2005).

Miyoshi and Niagami (2007) used a linguistic approach for the opinion analysis of customer review data. They have extracted and analysed adjective-noun pairs from a specific set of sentences. They have taken extra care for contextual valance shifters in order to understand the change in semantic orientation triggered by them. Similarly, Shaikh et al. (2008) have used a linguistic tool SenseNet in order to extract the verb frames of sentences to calculate contextual valance for whole sentence. However, they have not calculated scores for different aspects within a sentence.

Ding et al (2008) have also used a lexicon based approach for binary opinion classification, using product review data (using features and aspects) at sentence level. They have used a linguistic parser to associate POS to the sentences. However, they have not used grammatical dependencies at all. Grammatical dependencies are discussed in detail in Section 3.5.1. They have used occurrences of words as a key to opinion analysis where the occurrence of a positive word means +1, and a negative word means -1. The overall opinion is classified as positive if there are more positive words in the sentence and vice versa. The absence of grammatical dependencies in their analysis and ordinal level classification are a few of the limitations of their work which are further discussed in Section 2.10.

Sentence level opinion analysis is not suitable for compound sentences. As Wilson et al. (2004) have pointed out a single sentence can contain multiple opinions, but also can have multiple subjective and factual clauses. Therefore only classifying sentences into subjective/objective sentences is not enough. An opinion can be inferred from many objective sentences and such opinions are called implicit opinions. For example; “The earphone broke in just two days.” Although the sentence is an objective sentence which is indicating a fact, it is implicitly communicating a negative opinion about the earphone. Thus, in order to analyse the opinion there is a need to analyse both (subjective/objective) types of sentences and process different opinions on multiple aspects expressed within a sentence separately in each clause. This requires a further finer level of granularity in text: clause level.

2.8.3 Clause Level Opinion Analysis

Clause level opinion analysis is a more complex and refined level of opinion analysis, as it first divides a sentence into different (dependent and independent) clauses. These clauses are then processed in order to assign the opinion scores based upon the opinion topic (Fiscus and Doddington, 2002; Wilson, 2008; Thet et al., 2010). Clause level opinion analysis helps in the further analysis and comparison of the strength of opinion in different clauses and sentences based on respective aspects and topics.

Wilson et al. (2004) in their research identified opinions and calculated opinion intensities at clause level. They identified and gathered a wide range of clause level features based on syntactic and lexical cues. These cues are generated through dependency parse trees. Their research divides sentences into nested clauses, this is further explored in the proposed system in this Thesis in Section 3.3.1. They analysed the grammatical relationships of words, in order to classify the intensity of the opinion (as neutral, low, medium or high) for each clause. They also presented the clauses and sentences with respect to feature vectors, where grammatical relations are used as features. Further three different machine learning techniques (BoosTexter, Ripper, and SVM) were used in their research in order to determine the feature vectors for the classification of opinion intensity values. Their classification of opinion at clause level using an ordinal scale, and the association of a maximum intensity level to each opinion without any consideration of opinion topics are a few of the limitations of their work.

Similarly, Thet et al. (2010) have used clause level opinion analysis based on a linguistic approach to find grammatical relationships. These relationships are used with a rules based approach in order to associate sentiment scores at clause level. These real value scores (between +1 to -1) are associated with pre-defined aspects. They calculated sentence level opinion scores by averaging the opinion scores for all aspects within a sentence. However their research has the following limitations: an absence of any analysis for the aggregation of opinion scores; non-inclusion of noun phrases for the opinion topic/aspect, and overall topic of textual data; and a lack of analysis between the opinion topic and the overall topic. These limitations present an opportunity for further exploration of this area. Lu (2010) also has proposed along with others to extend the targets to verb phrases and embedded clauses in addition to noun phrases (Bethard et al., 2005).

2.8.4 Phrase and Word Level Opinion Analysis

Phrase and word level opinion analysis assigns opinion polarity and opinion scores to words and phrases. These opinion values either just indicate a positive, neutral or negative polarity with discrete values (+1, 0 and -1) or take continuous values between +1 and -1 providing a finer resolution in the measure of their opinion polarities (Rill et al., 2012b).

In the majority of opinion analysis techniques and approaches words are employed to understand and analyse opinion expressed in textual data. The words which are used to express positive opinions are known as positive words whereas the words which are employed to express negative opinion are known as negative words. ‘Beautiful’, ‘gorgeous’, ‘elegant’, ‘excellent’ and ‘amazing’ are examples of positive words, and ‘bad’, ‘poor’, ‘terrible’, ‘horrible’ and ‘ugly’ are examples for negative words. However, many times single words do not hold the complete meanings and therefore phrases and idioms are used in order to improve the opinion analysis (Wilson et al., 2005; Tan et al., 2011c). Wilson et al. (2009) observed that the series and combinations of words (phrases, intensifiers and modifiers) can change the intensity of the polarity of the opinion expressed. Most of the word and phrase based research in opinion analysis has contributed to the development of lexical resources. The utility of phrase level linguistic analysis is also witnessed in literature, but it mainly relies on the extensive manual annotation provided through training datasets (Wu et al., 2009; Toprak et al., 2010) and scoring techniques (Agarwal et al., 2009; Rill et al., 2012b). A number of phrase level analysis techniques use heuristic rules and methods of pattern extraction in written text (Choi et al., 2005), based upon POS (Tan et al., 2011c) or dependency parsers (Tan et al., 2011b).

Wilson et al. (2009) used a machine learning approach by recognising the appropriate features to automatically determine the contextual polarity of a phrase in opinion analysis. Moilanen and Pulman (2007) and Shaikh et al. (2008) used constituents within phrases to determine the opinion polarity.

Shaikh et al. (2008) used semantic relations within a sentence using verb frames. They combined verb based relationship with contextual valence of the words and generated the rules to calculate the opinion score for the overall sentence. They utilised a number of internet based resources, cognition and common sense knowledge to manually score verbs and adjectives. Moilanen and Pulman (2007) proposed a sentiment composition model based on the concept that the overall polarity of the sentence is a function of the polarities of its parts.

However, none of the above opinion analysis approaches has really investigated multi-word phrases using typed dependencies to provide a fine grained analysis of the relations between words. In addition, none of the above approaches have subsequently calculated the opinion polarity based upon rules defined from typed dependencies (Tan et al., 2011a).

Sometimes even phrases are not detailed enough to capture the opinion, and opinion topic within them (Wiebe and Riloff, 2005; Wilson et al., 2005), therefore analysis at the word level utilising manual or automated approaches has a lot of disadvantages. These disadvantages and limitations revolve around the completeness, validity, consistency, and domain dependence of the resource. The details of some of the existing lexical resources and their limitations are discussed further in Chapter 3 and Appendix A.

2.9 Existing Techniques for Opinion Analysis

There is an extensive body of work that addresses different techniques and approaches adopted for the analysis of opinion in written text based upon different types of data and requirements of analysis. The following sections provide a focus on key techniques which are of most relevance to this Thesis. These approaches are also presented in the form of a hierarchy in Figure 2-2.

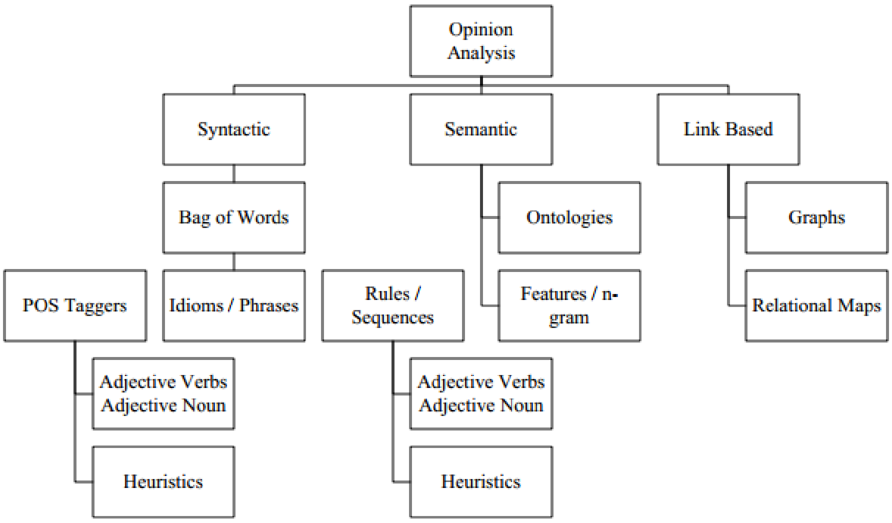


Figure 2-2: Opinion Analysis Hierarchy

2.9.1 Bag of Words (BoW)

In order to understand the meaning of communicated text, words are always given importance. The selection of appropriate words is necessary to get an individual's thoughts communicated. Therefore, the frequency of words used in a textual data segment has always been of much importance for textual analysis as discussed earlier.

Similarly for opinion analysis, BoW is the most commonly used technique for opinion extraction (Turney, 2002; Pang et al., 2002; Kennedy and Inkgen 2006). BoW is a representation where each word in a document is represented by a separate variable numeric value (weight) (Grobelnik and Mladenic, 2004). The BoW technique enables analysis of the vocabulary and choice of words as a way of communicating opinion and meaning in written text.

The concept of analysing the choice of words and the vocabulary used is called compositional semantics (Choi and Cardie, 2008). Compositional semantics takes the meaning(s) of a sentence to be dependent upon the meaning(s) of the parts of the sentence and the way that those parts interact with each other (Choi and Cardie, 2008). Within this technique the opinion bearing words, polar words, or opinion oriented words are deemed to be of most importance in order to find and track opinions within a document or a sentence (Pang et al., 2002b). The analysis of opinion within textual data segments is unlike verbal communication, where people convey their opinions and emotions through many modalities: linguistic content, speech, vocal variations, pauses, facial gestures and body language (Quan and Ren, 2010).

Human language is ambiguous, therefore, many words can be interpreted in multiple ways depending on the context in which they occur (Navigli, 2009). According to LingPipe, in their tutorial on word sense, words are fluid, living things that change meanings through metaphor, extension, adaptation, and just plain randomness (Lingpipe, n.d.). For example; given two sentences **(a)**"The use of a double bass gives them an original sound.", and **(b)**"They like grilled bass."; both contain the word 'bass', however, 'bass' has different meanings in each. Therefore by just reading or listening a word, sometimes even human beings do not get a very clear picture about its meaning, without clear knowledge of context and topic. An automated system when processing unstructured textual information and transforming it into data structures, finds it hard to differentiate the underlying meanings of each word used. This means one has to identify the exact sense of the word before attaching the correct opinion orientation and intensity score to the word. This leads to the issue of Word Sense Disambiguation (WSD). The problem of WSD does not directly fall into the scope of current research, however if employed coupled with this research can improve the results significantly as current research relies heavily on lexical based analysis, which mainly counts on the identification of correct opinion associated within the resource for the word encountered in textual data.

Opinion bearing words are captured from textual data on the basis of POS. The POS tagging can also be considered as a basic way towards the solution of WSD (Wilks and Stevenson, 1998). It is thought that adjectives and adverbs in a sentence mostly convey the opinions (Pang et al., 2002; Agarwal and Bhattacharyya, 2005; Chesley et al., 2006). Therefore lists of positive/negative words (Li and Wu, 2010), adjectives/adverbs (Hatzivassiloglou and Mckeown, 1997), intensifiers/diminishers (Subrahmanian and Reforgiato, 2008) have been quite extensively used in the process of text based opinion analysis.

In the literature a large amount of research is based upon list based approaches, analysing textual data through lists of positive and negative words. These are then compared with gathered word(s) from the sentence (adjectives, adverbs and verbs) and an associated polarity is assigned to the text (Li and Wu, 2010). If a word is found in the positive list, it means that it has a positive polarity in the sentence or vice versa. List based techniques have their limitations, one of the basic limitations is all the words in a list may be attached with the same score. Whereas, it is not true, all positive words like, good and adorable do not have the same strength of opinion attached with them. Therefore, words with different strengths and behaviours are classed in different lists, giving researchers a way of using more than two lists for opinion analysis (Li and Wu, 2010). Another limitation faced while dealing with lists is that lists can never be complete enough to capture all the words encountered during analysis. Therefore incomplete lexical resources can result in an inappropriate analysis. As the words which are not found in lists during analysis are classed as neutral, whether they are neutral or not. This may lead to inaccurate opinion analysis. In addition to list based approaches there are also techniques which discover interaction between words within textual data (Kessler and Nicolov, 2009). Thet et al (2010) have criticised the usage of the BoW approach for short texts, however, they thought it to be suitable for long texts. Turney (2002) have predicted document level opinion orientation using phrase based resources, where phrases were extracted based on predefined POS patterns (Rill et al., 2012b). Their pre-defined patterns were mainly based upon selected POS, i.e., adjectives, adverbs, verbs etc. If their technique encountered any pattern other than their predefined patterns, the system is not able to interpret and analyse the opinion. User generated textual data available online may not be constructed in a formal pattern, the structure of sentences do not always follow structural rules of language therefore there are more chances that the sentences encountered during opinion analysis fall outside of pre-definition leading to erroneous opinion analysis. Different lexical resources are developed using words as a basic unit, and some of them are discussed and evaluated in Chapter 3.

2.9.2 Rules Based

In extension to a basic textual analysis technique such as BoW (or lists) is the addition of a rule set. Rules such as negation rules (Wilson et al., 2005; Kennedy and Inkpen, 2006; Das and Chen, 2007; Li and Wu, 2010) or conjunction rules (and, but, however, punctuations, question marks or exclamation marks) in the sentences (Na et al., 2004; Wilson et al., 2004; Wilson et al., 2005; Airoldi et al., 2006; Kennedy and Inkpen, 2006) are defined separately as their usage can change the meaning of opinion oriented words (Li and Wu, 2010b). Similarly, phrases and opinion oriented idioms are also very important as a source of identifying opinion within textual data (Turney and Littma, 2003; Yi et al., 2003). For example; “He is a good and intelligent person.” and “He is not good but intelligent.”, in these two sentences ‘and’ and ‘but’ are changing or enhancing the opinion. Therefore, these kinds of rules have to be predefined for the analysis of written text within any language. Sarcasm is a phenomenon which is mostly ignored in opinion analysis, however a rules based approach can be used to detect and analyse sarcasm for opinion analysis on Twitter. Though this approach heavily relies on pre-assigned hashtags (Maynard and Greenwood, 2014).

2.9.3 Syntax

Many researchers have used syntactic dependencies as a way of understanding opinions in documents (Dave et al., 2003; Michael, 2004; Higashinaka et al., 2006; Liu, 2010). Dependency structures represent all the relationships in a sentence, uniformly as typed dependency relations between pairs of words (a governor and a dependent) using a syntactic parse tree (Lu, 2010). The tree structure is also used quite extensively to understand relationships between words in sentences, for example; the use of a Sentiment Progression Graph (Fei et al., 2006; Kessler and Nicolov, 2009; Wu et al., 2012). While parsing textual data syntactic dependencies can help in the modelling of valence shifters such as intensifiers and diminishes (Kennedy and Inkpen, 2006). The information from a dependency parser can also help in the resolution for the scope of relations between words within sentences (Kennedy and Inkpen, 2006) also discussed in Section 2.9.5. For example, relations of negation and conjunction etc. as shown in Figure 3.5.

2.9.4 Machine Learning

Machine learning techniques in the area of opinion analysis for textual data are very popular, as they introduce the capability in the system to adapt to changing input, give general machine understandable rules, as well as a possibility to compare and measure the degree of similarity of the input with these rules (Boiy and Moens, 2009). In supervised machine learning, manually annotated training datasets are given in order to train the system and the system can continue to learn and update the rules based on the dataset (Pang et al., 2002).

Manual annotation of a dataset can be a very slow, time consuming and biased process which is discussed in detail in Section 4.2 and Section 6.4.2. One of the main limitations in machine learning techniques is that each wrongly annotated instance may generate an incorrect rule and if a similar instance reoccurs, it can propagate and may have a knock on effect to other values as the dataset is generally used for training purposes.

2.9.5 Structural Opinion Extraction

The above sections have focused on techniques which emphasise analysis of the syntactic structure of textual data. A deeper approach that has been undertaken is the analysis of the semantic understanding of the relationships within textual data. This approach involves determination of the parts of opinion, such as what the opinion is about (the *target*) and who is expressing it (the *source*). Kim and Hovy (2004) have proposed that a quadruple [Topic, Holder, Claim, Sentiment] value should be assigned to an opinion in its extraction. However, they have not taken this to a further level of sub-topics of the topics. In 2006, they extended this work through the use of FrameNet data, and using verb or adjectives as opinions (Kim and Hovy, 2006a). Kim and Hovy (2006) based this extension on the manually annotated semantic roles of FrameNet data. However, they identify that other POS such as adverbs and nouns can also affect the performance of the system. Bethard et al. (2005) have tried to capture opinion holders by identifying propositional opinions. They used manually annotated training and evaluation corpus initially to enable this. However, their results show that when this initial analysis is coupled with the usage of another externally generated opinion oriented words list it showed improved results. Similarly, they found that the use of semantic-role-detection processes can also be very useful in improving the result set. Efforts have also been made to identify opinion holders and targets by using syntactic dependency parsers (Lu, 2010). Lu (2010)'s work was based on the use of verbs and can further be extended to include verb, noun phrases and clauses. With the multi-lingual nature of user generated content on the Web, the field of opinion analysis has also seen advances in the area of multi-lingual opinion analysis, with work ongoing in this area (Guo et al., 2010; Xu et al., 2011).

Work related to the field of topic specific opinion analysis is based on syntactic contexts, how opinion words syntactically link to targets of opinion (Jijkoun et al., 2010). This can be achieved by manually annotating the contextual polarity of a text or series of texts and generating a corpus (Wilson et al., 2005), using a clustering technique to detect topics of the stories (Fiscus and Doddington, 2002). There can however be problems with the approach when dealing with multi-topic stories, and stories with no clear topics (Fiscus and Doddington, 2002). There is another way of resolving syntactic contexts through the help of syntactic dependencies (Jijkoun et al., 2010). Though Jijkoun et al (2010)'s work has only considered individual words and there is still scope to identify the noun or verb phrases for topic target identification and extraction.

In addition, Jijkoun et al. (2010) have generated a topic specific lexicon resulting in a list of triples (clue word, syntactic context and target) as lexicon. This is quite similar to the works in domain and target specific opinion mining. Similar to the approach used by Kim and Hovy (2004), Godbole et al. (2007) have generated a lexical resource by manually selecting seed words in the field of health and business for their work in domain specific subjectivity and polarity calculation for specific topics (Kim and Hovy, 2004a). There are limitations in these works as Jijkoun et al. (2010) have only used individual words in their research instead of having opinion targets to include multi-word phrases (Noun Phrases (NP) and Verb Phrases (VP)), and, Kim and Hovy (2004) have only presented an idea to use a quadruple [Topic, Holder, Claim, Sentiment], which they further extended in their work in 2006.

Wilson et al. (2004), Ding et al. (2008), and Thet et al. (2010) have also presented work in the past which has similarities to the proposed work detailed in Section 2.10. However, there exist differences and improvements in the proposed work which are presented below. The work of Wilson et al (2004) and Thet et al. (2010) are presented in detail in Section 2.9.3. However, the differences with the proposed work are further presented in the following section.

This section is an overview of the proposed technique for opinion analysis, presented in order to understand the gaps in the present state-of-the-art and to suggest how these gaps can be filled through the proposed technique of opinion analysis.

2.10 Limitations in State-of-the-Art

The above sections have detailed literature related to understanding the state-of-the-art in opinion analysis. This section brings forward identified limitations of the state-of-the-art in order to provide a basis for the proposition of an improved opinion analysis approach in Section 2.11 (and further detailed in Chapter 3).

L1 - The manual annotation of datasets for use as training data for opinion analysis techniques is time intensive and can result in inconsistent annotation and bias (Greene, 2007; Tadano et al., 2009; Lu et al., 2011). In addition, there is a large body of work focused on identification of idioms, phrases and other sophisticated linguistic features, but the majority of this work is conducted manually (Rill et al., 2012a; Rill et al., 2012b).

L2 Existing techniques for opinion analysis based upon clause level and phrase level analysis require improvements in order to develop the opinion structure as opinion, opinion holder and opinion topic. (Toprak et al., 2010).

L3 - Kim and Hovy's (2004, 2006) research about the structure of extracted opinion presented an improvement on previous state-of-the-art systems. However, research related to finer granularity levels still requires further exploration, particularly in relation to complex and comparative sentences (Wu et al., 2012).

L4 - Aspect based opinion analysis has a pre-requirement of a hierarchy of 'objects'. At present hierarchies need to be pre-defined, which makes things difficult for a general purpose opinion analysis system (Snyder and Barzilay, 2007; Thet et al., 2010).

L5 - 'Objects' are generally expressed through phrases (noun phrases). State-of-the-art techniques depend on word based approaches to determine the interaction or relationships within sentences. There is a need to incorporate phrase identification within clause level analysis (Thet et al., 2010; Toprak et al., 2010).

L6 - Sophisticated linguistic features to identify the local (phrases) and global relations between words within sentences may bring improvements to traditional opinion analysis techniques utilising typed dependency structures (Marneffe et al., 2006; Wu et al., 2009; Rill et al., 2012b).

L7 - POS is presently widely used in opinion analysis. However, POS does not communicate the structure of sentences within datasets (Penn_Treebank, 1992). Understanding sentence structure may lead to a greater semantic understanding of opinion based data (Ku et al., 2008; Athar, 2011). The sentence structure (Subject-Verb-Object) should be integrated with the hierarchical structure of a sentence to better enable clause level opinion analysis (Thet et al., 2010). In relation to present approaches using phrase level analysis there are restrictions in how POS are used and only a limited subset of POS are employed during analysis.

L8 - Present scoring methods used in opinion analysis approaches mainly rely on discrete values as determined most often on likert scales. Improvements in relation to other state-of-the-art limitations have an impact on scoring mechanisms. A refined opinion scoring technique based on a real score (between -1 and +1) is required. Thet et al. (2010) and Rill et al. (2012b) have taken steps towards this, however, in the author's opinion such approaches need to be adapted to accommodate the local structures and the global inter-dependence of words (negation, intensifiers etc.) in order to calculate an overall opinion score.

L9 - There are substantial problems as noted in Section 2.8 with regards to the aggregation of scores across different levels of granularity. For example, aggregation may result in determination of neutrality, where neutrality does not necessarily hold true for elements of the sentence. Therefore further research is required into understanding approaches to aggregation and calculation of overall opinion scores associated with any topic within a document repository (Rill et al., 2012b).

2.11 Proposed Technique

Limitations of the state-of-the-art would suggest that original approaches to opinion analysis based on an understanding of data items at clause level, coupled with phrase identification may result in an improved understanding of opinionated data. Development in the approach may also impact on mechanisms for scoring data items and could result in finer grains of granularity relating to opinion aggregation. This provides the foundation for the development of a proposed technique for opinion analysis. This technique is briefly outlined below, then outlined in greater detail within Chapter 3.

The proposed opinion analysis approach will be a clause level approach using linguistic and syntactic analysis, based upon dependency grammars, for automatic resolution of dependencies in different parts of sentences. For each sentence, the approach will require the generation and use of a dependency tree and constituent tree, in order to identify different parts of sentences based upon the local structures (phrases) of words. The top level clause in the constituent tree will represent the entire sentence. Sentences will be divided into nested clauses based on their syntactic dependencies, and phrase level dependency trees will be generated. These phrase level dependency trees will have sub-trees which help in dividing clauses into phrases (noun phrase, verb phrase, adverbial phrase, etc.). This approach will help in identifying the clauses in sentences which are based upon one aspect (topic). For example; given the sentence, “I loved the restaurant’s environment but hated the food.”, this can be divided into two clauses. “I loved the restaurant’s environment.” and “I hated the food.” The first sentence demonstrates a positive opinion about the environment of the restaurant and the second sentence is expressing a negative opinion about the food.

Analysis of noun phrases will be undertaken in order to provide a better understanding of the opinion topic and opinion holder, as syntactic dependencies are only based upon single words not phrases. For example; “My mother found Sony very expensive.” is a sentence and a syntactic parser will only identify mother and Sony as Subjects of the sentence. However, mother is part of a noun phrase (my mother), which needs to be further analysed especially if there is a comparison with another sentence, “His mother thought Sony is cheaper". Therefore phrase level syntactic dependencies will be analysed to provide a semantic understanding of relations and dependencies.

In the proposed approach, opinion scores for the strength of opinion will be calculated over a range of real numbers (+1 maximum score for most positive and -1 for most negative). As a distinction from present state-of-the-art approaches (Li and Wu, 2010; Thet et al., 2010) this proposed scoring and aggregation technique will take into consideration the opinion topics and overall topic of a document to aggregate intelligently.

An opinion score will be attached to each opinion oriented word and these scores will be aggregated and calculated for all of the phrases and for clauses in a sentence. The sentence structure will be utilised and aggregation will be performed based upon a topic analysis. If the clause level opinion topic and the document topic have any relation, than the opinion will be aggregated and calculated to generate an overall opinion of the document, according to the relation between document topic and opinion topic.

The reliability and validity of the proposed approach will be evaluated by developing a prototype system; however, the development of a prototype will require the use of existing lexical resources and the pre-extraction of opinion related data. In addition, a series of algorithms to provide opinion analysis and to provide a scoring mechanism will be required to be designed by the author of this thesis. Therefore, some of the existing lexical resources and existing frameworks are reviewed and explored in Chapter 4, in order to understand the specific features of these resources.

2.12 Summary

This chapter began with exploring the roots of opinion analysis in fields like web mining and text analysis. This then moved into the examination of opinion analysis as an independent area of research and explained the basic components of opinion analysis and structure of opinion in written text. Different existing levels of granularity to analyse opinions and techniques have been reviewed and the motives and reasons behind the development of different techniques are analysed. Limitations of current state-of-the-art are highlighted and based on these limitations an opinion analysis technique is suggested. This approach will be further explored in Chapter 3 and its novelty will be discussed.

Chapter 3 – Proposed Opinion Analysis Approach

In Chapter 2 the state-of-the-art for opinion analysis is critically analysed and the chapter concluded with an identification of the limitations of the existing approaches, along with the proposal of a novel new approach for opinion analysis.

This chapter initially developed an understanding of the basic elements underlying the proposed opinion analysis approach. This approach aligns with a number of the previous state-of the-art mechanisms identified in the previous chapter, and the most relevant of these are further explored through a more detailed analysis in this chapter. Time is then taken to detail aspects of the proposed new opinion analysis approach including detail of key decision points in the process regarding particular challenges in construction e.g. the decisions made about the level of granularity, parsing and scoring techniques. Finally, the chapter completes with identification of issues impacting on the operation of the approach which fall outside the scope of the current research, including those identified as limitations of the proposed approach. It is noted that with improvement to these challenges the results of the proposed opinion analysis approach may improve.

3.1 Opinion Analysis Technique

Traditionally, words used for textual information in any language are considered to be highly influential in understanding and inferring meaning from the text. Therefore, lexicon based methods for opinion analysis (simply using opinion bearing words) in order to identify the opinion orientation and strength of opinion have been commonly used (Ding et al., 2008; Li and Wu, 2010; Thet et al., 2010). However, there are a wide variety of words and phrases. People also use words and phrases in a multitude of different combinations within a sentence, in order to express their opinion or point of view. Words used in a sentence can change the semantic orientation of other terms they are used with. Therefore, sentences are treated as expressions and lexico-syntactic patterns can also be used for the identification and extraction of opinion in textual data (Riloff and Wiebe, 2003; Wilson et al., 2005; Subrahmanian and Reforgiato, 2008).

The current research is based upon an understanding that a sentence is a sequence of words presented in a particular order, in order to convey an idea, event or description. The order of words in a sentence is very important to communicate the proper meaning and perception. There are defined and set rules specifying the order of words based upon English Language sentence grammar. The boundaries of a sentence are defined by: they begin with a capital letter and terminate with a punctuation mark (period, question mark or exclamation mark). However this definition of a sentence may cause difficulty in user generated content as the majority of Web users do not have English as their first language and they mostly do not follow the formal structure of sentences, missing for example punctuation marks. These missed punctuation marks may result in the coupling of two or more sentences together into a single sentence. This scenario is further discussed in Section 7.2.2 with the help of an example sentence: *"I am worried my 7yr old boy is bully at and they told me if he doens ' t stop it they are goin to expell him from , please help me how do i stop him from being bully ?"*.

The opinion analysis approach briefly described in Section 2.11, is based upon some basic understandings. Some of the understandings are as follows:

1. A sentence is a combination of words, which can be used in different combinations and the sequence can change the overall meaning of the sentence.

Table 3-1: Example Sentences

Sentence Number	Sentence
1	The picture quality of this camera is good.
2	The picture quality of this camera is very good.
3	The picture quality of this camera is not good at all.
4	The picture quality of this camera is not so good.
5	The picture quality of this camera is good, but the battery life is too short.

The first four sentences 1,2,3,4presented in Table 3-1 include the opinion bearing word 'good' used with different combinations of words. Sentence 1 and 2 are positive sentences about the picture quality of a camera, with, sentence 2 being more positive than sentence 1. Similarly, sentences 3 and 4 provide a negative orientation, with sentence 3 being more negative than sentence 4.

2. The structure of a sentence in terms of identification of clauses and phrases is very important for resolving the parts of sentences (Subject, Verb, and Object).

In sentence 5 presented in Table3-1, two opposing opinions are expressed with opinion bearing words, ‘good’ and ‘short’; and both these opinions are about different features of the same camera. “The picture quality of this camera is good.” “The battery life of this camera is too short.”

There is a need to identify the clauses and their respective Subjects and Objects in order to understand the opinion topics and opinion holders of both opinion bearing words, ‘good’ and ‘short’. Therefore, the basic unit of analysis for the proposed opinion analysis technique is considered as a clause.

3. The semantic orientation of opinion bearing words is to be identified by using some lexical resource.
4. Different language constructs, semantic and syntactic dependencies change the semantic orientation of opinion bearing words.

For example:

- In sentence 2 the adjective ‘good’ is modified by the adverb ‘very’, and ‘good’ is intensified;

- In sentence 3 “is not good at all” is a verb phrase, which contains “at all” as an adverbial phrase and ‘good’ as an adjectival phrase. This provides a negation relation between the words ‘not’ and ‘good’ where the adjective ‘good’ is modified by the adverbial phrase “at all”.

Therefore, dependency grammar and syntactic structures can be used for the resolution of semantic and sentiment orientations; and phrase level analysis can also be employed in order to understand the higher level of syntactic dependencies.

It is to be understood that ‘short’ in itself is not a negative word. It is just an adjective. However, when ‘short’ is attached as a property of the ‘battery life’ of a camera, it becomes a word with negative orientation. Therefore the topic associated with the word can help in understanding its sense.

5. The assignment and aggregation of opinion strength at a sentence, paragraph and document level is based upon the opinion topic as identified in the respective clause.

The proposed opinion analysis technique as previewed in Chapter 2, is a clause level opinion analysis approach, based upon a nested sentence structure. The sentence is broken down into a nested structure of clauses and phrases. The basic POS are tagged and tokenised. These tagged tokens are grouped in the form of phrases and clauses into a hierarchy. This gives the sentence as a tree of parsed words named as a ‘constituent tree’ by Wilson et al. (2004).

A lexical resource can be used to associate opinion orientation and strength to each word encountered during the parsing of the sentence. This assigns an initial opinion value to each word. Non-opinion bearing words are assigned with ‘neutral’ for their orientation and strength.

There is a need to resolve the relations between the words within a sentence and the scope of the words in terms of modifiers (for the modification of semantic and sentiment orientation and strength). The linguistic and syntactic analysis is carried out based upon the dependency grammar, and dependency parser. The dependency parser identifies the grammatical relation between two words, and also identifies the governor and dependent words. This helps in resolving the scope of words.

The grammatical relations along with the constituent tree provide a set of rules and conditions to calculate the overall opinion orientation and strength at phrase/clause level. This is done by first resolving the scores for individual words (adjectives, adverbs, verbs, etc.), then resolving all of them at the level of phrases (adverbial phrase, adjectival phase, etc.). Finally a score is calculated for the overall opinion at clause level. Opinion analysis at clause level consists of the identification of opinion as (Opinion Words, Opinion Topic, Opinion Orientation, and Opinion Strength) as detailed in Section 3.4.2.

The opinion topic and opinion holders are mostly contained in noun phrases and the aggregation of opinion based on the topic can be completed by further analysis of noun phrases. For example; sentence 5 “The picture quality of this camera.”, has multiple noun phrases “the picture quality”, “this camera”, and “the battery life”. Where “the picture quality” is Subject for the first clause, and “the battery life” is the Subject for the second clause. As both the clauses are joined with the conjunction word ‘but’ this specifies an opposing opinion orientation for both subjects, with “this camera” being the Object for both clauses. This property of conjunction ‘but’ is further discussed in Section 3.5.4. If there is an additional opinion about “the picture quality” of the same object it is aggregated accordingly (depending upon orientation and strength). For example in sentence 5 the opinion about the object “this camera” is retrieved, the opinion orientation and strength across both the clauses is aggregated as they are opinions expressed about the same object of “this camera”.

3.2 Methodology

During the preliminary examination of opinion analysis as an area of research in Chapter 2, three existing research projects are identified which are most related to the opinion analysis approach proposed. These projects are further analysed in Section 3.3, in order to understand the principle areas of interest. A comparison of each of these research approaches will be drawn with the proposed approach and the novelty of the proposal will be established through this analysis.

Following this, an in-depth analysis will be undertaken in order to propose effective techniques and resources at each stage of the opinion analysis process. Further analysis of the most closely related research will help to identify additional limitations to those presented in relation to the area of opinion analysis stated in Chapter 2.

3.3 Analysis of Identified Research

3.3.1 Wilson et al. (2004)

Wilson et al. (2004) claim that their research is the first that automatically classified the strength of opinions in clauses. They used clause level analysis as they argued that there can be more than one opinion in a sentence and that a single sentence can has both opinion as well as factual information. They identified a variety of subjectivity clues in the language based on syntactic clues using a dependency parse tree (POS tags and grammatical relations) and previously established clues supplied by the research community (Hatzivassiloglou and Mckeown, 1997; Baker et al., 1998; Dave et al., 2003; Riloff and Wiebe, 2003). They used a lexicalised English parser for this purpose (Collins, 1997). In addition, they identified clauses based upon non-leaf verbs using the Collins parser. Through this work, they established five syntactic clues for each word in a dependency parse tree as presented in Table 3-2.

Table 3-2: Syntactic Clues established by Wilson et al. (2004)

Clue	Description
root(w, t**)**	Word w with POS tag t is the root of a dependency tree (i.e., the main verb of the sentence).
leaf(w, t**)**	Word w with POS tag t is a leaf in a dependency tree (i.e., it has no modifiers).
node(w, t**)**	Word w with POS tag t.
bilex(w, t, r,wc, tc**)**	Word w with POS tag t is modified by word wc with POS tag tc, and the grammatical relationship between them is r.
allkids(w, t, r1,w1, t1, . . . , rn,wn, tn**)**	Word w with POS tag t has n children. Each child word wi has POS tag ti and modifies w with grammatical relationship ri, where 1 _ i _ n.

They used a manually annotated MPQA corpus and clearly accepted the low inter-annotator agreement for this corpus (see Section6.1.4 for an explanation). In their work they claim that no specific efforts were made to align the strength scales of different annotators.

The MPQA corpus itself provided the following attributes:

- Who is expressing opinion?
- What is target of opinion?
- Type of attitude
- Subjectivity strength {neutral, low, medium, high}
- Neutral means absence of opinionated data (subjectivity)

A wide range of experiments were constructed to establish the clues, both from the literature as well as the identification of new syntactic clues. These helped to develop strength in the classification process of the system based upon the techniques of boosting, rule learning and support vector regression. They evaluated and validated their system using a baseline mean squared error and accuracy for all algorithms.

Similar to the proposed opinion analysis approach, Wilson et al. (2004) presented a technique based on clause level opinion analysis and used syntactic analysis as well as other rule based techniques. However, they did not recognise the nested structure of the sentence beyond clauses. They have not identified and considered phrases.

The presence of subjectivity and strength of opinion was presented on an ordinal scale and there was no provision for the calculation of the strength of opinion using any of the modifiers (negation, diminishes or intensifiers)/valance shifters. Modifiers were only considered if they had already been annotated in the MPQA corpus, and any un-annotated combination of words would guessed using previously established subjectivity clues.

No mechanism was defined in order to aggregate the opinion scores to the levels of paragraphs or to the document level.

3.3.2 Ding et al. (2008)

Ding et al. (2008) provided a sentence level opinion mining approach for online product review data. Their approach was based upon the lexicon-based approach proposed by

Hu and Liu (2004). The opinion lexicon used as a resource was obtained through a bootstrapping process using WordNet (Fellbaum, 1997).

The semantic orientation (positive, negative, neutral) of product features was assigned. In addition they identified the requirement of finding semantic orientation for implicit and explicit opinions as expressed in textual data, based upon the context and domain. For example, “This camera is too large.”, identifies the size of a camera as being very large as a negative opinion. ‘Large’ is just a feature indicator and something being large is not necessarily a negative opinion. It is context dependent and there is a need to employ an expert in the context to identify the semantic orientation of each word in any domain.

Their approach can help a system to handle words which were context dependent. They presented a special function to aggregate multiple conflicting opinion words within a sentence. A distance based scoring method was presented for the aggregation of word orientation. Negation, conjunction and synonym/antonym rules were handled for the scoring technique.

The usage of a lexical based approach, initial assignment of scores to all opinion bearing words, and the proper handling for negation and conjunction rules were features which made this technique close to the proposed approach of opinion analysis. However, there were a few limitations of their research which made it different from the proposed approach. Ding et al. (2008) very clearly stated that the technique could not deal with comparative sentences. It could only work if the domain/context based knowledge was known. It had not considered syntactic information. Context dependency was also not handled in this research, which is really necessary to handle free web based textual data, as web based user generated text may not follow the formal language or grammatical structures.

3.3.3 Thet et al. (2010)

Thet et al. (2010) presented an aspect based, clause level opinion analysis technique. They used movie review data on discussion boards and found the opinion orientation and strength of opinion at clause level. The technique used for opinion analysis is highly dependent upon the dependency structure at clause level, i.e., splitting sentences at clause level and then identifying the clause level dependency structures. They used a linguistic approach using grammatical relations in order to understand the dependencies. The orientation and intensity of opinion were found based upon the pre-defined aspects in movie review data.

Their approach is similar to that proposed in the current research in many ways. Thet et al. (2010) also presented clause level opinion analysis technique based upon linguistic features. They assigned opinion scores based upon the intensity of opinion across each clause. However, Thet et al. (2010) had not used a nested sentence structure (clauses and phrases) in order to assign scores and had not used phrase level analysis in the aggregation of opinion scores. The current research employs other semantic and conjunction rules for the calculation of overall opinion across the sentence, whereas Thet et al. (2010) only relied on grammatical dependencies. Limitations exist in the approach for example: there was no consideration of conjunction rules; the pre-defined nature of aspects used during opinion analysis limits the scope for re-use; the lack of phrase level analysis at topic level reduced summarisation; and the approach taken in the aggregation of scores limits flexibility.

3.4 Development of a Novel Opinion Analysis Approach

The sections below provide detail of an approach for opinion analysis proposed as an advancement of the state-of-the-art by this Thesis. The proposed approach is focused on improving opinion analysis in the area of review based data. Improving the analysis of review based data provides an area which can produce significant organisational value, especially when applied in the context of social media networks, blogs, or forums. Manual retrieval and tracking of this review based data without the help of any IR and analysis tool is very difficult. The retrieval and analysis of this data is equally important for organisations to gain an understanding of their present customer perception; and for prospective customers, who are comparing a range of products before purchasing. Companies may benefit through keeping track of public opinion in relation to their products and in improving the products according to the customer opinion. Prospective customers may benefit from an analysis of the reviews of people who are already using the products, helping to inform them of the positives/negatives across a range of similar items.

3.4.1 Research Design

This study combines both the syntactic and semantic methods for automatic opinion based analysis. The dependency parsers, constituent parsers, and lexical resources are used together with some syntactic and semantic rules in order to identify the opinion orientation and the intensity of textual data.

A dataset of 600 manually gathered sentences selected from existing research and datasets is used (Bethard et al., 2005; Jindal and Liu, 2006a; Kim and Hovy, 2006a; Kessler et al., 2010; Framenet, n.d.-a). Manual selection of the dataset is undertaken as the dataset is used to evaluate different aspects which the opinion analysis claims to cover i.e., negation, conjunction, comparative and complex sentences. Most of the sentences are collected from product review data in the areas of Cars and Cameras. A topic is also assigned with each sentence, as topic related data is needed for the testing of IR at a later point.

3.4.2 Structure of Opinion

Before providing details about the proposed opinion analysis approach it is useful to outline the approach taken in relation to understanding the structure of opinion oriented data. There are multiple approaches to understanding opinion oriented data in textual data segments (as outlined in Section 2.7.3). In this section the particular approach that forms the basis for more detailed opinion analysis de-construction is defined.

In general an opinion can be voiced about anything; a product, a feature of the product, an individual, a habit of the individual, an organisation, an event, a topic, etc. Therefore, a general name ‘object’ is given to the entity about which the opinion is expressed. The ‘object’ can has a set of components (parts) and it can also has a set of features (attributes/properties). This allows the object to be disintegrated into a hierarchical structure on a basis of part-of or feature-of relations. For example; “a digital camera” is an object having a ‘lens’, ‘battery’, etc. as its parts and ‘size’, ‘weight’, and ‘picture quality’ as its features.

More formally:

Object: an object ‘o’ is an entity which can be any product, person, organisation or event, etc. It can be represented as o (P, F) where P is a hierarchy of parts or sub-parts, and F is a set of attributes/properties. Each part or sub-part can has its own set of sub-parts and features therefore P (P, F).

As object is defined and represented as a hierarchical structure (tree). The root of the tree is the object itself, and a non-root node is a part or sub-part of the object. Each node has its own set of features or properties. An opinion can be expressed about any node of the tree.

Opinions can be voiced by any person, or on behalf of any person, or by an organisation. They are determined to be the source of the opinion. Generally the reviewers and authors of posts and blogs are the opinion holders. In news articles generally opinion holders are more explicitly stated and mentioned.

Opinion holder: the holder of an opinion can be a person or an organisation that has expressed the opinion.

Opinion: an opinion is a point of view, emotion or attitude (positive/negative) about an object from an opinion holder.

Opinion Orientation: the orientation of an opinion about an object is an indication that the opinion holder holds a positive, negative or neutral opinion (point of view) about the object.

Opinion Intensity: the intensity of the opinion expressed means how strongly the opinion is expressed. The opinion can be weak, mild or strong; it can be expressed as a score value from 0 to 1 where 0 is weakest and 1 is strongest.

Therefore combining orientation and intensity of opinion together gives (-1, 0, +1) values to scoring. Where –ve and +ve show orientation and real numbers show strength attached to the respective orientation.

3.4.3 Proposed Opinion Analysis Approach

The main idea behind the technique proposed for the opinion analysis is that all the words within a written sentence are knitted together in order to communicate the point of view of the author/opinion holder. Therefore, they must be used individually as well as together to understand the opinion expressed through them. The rules to bring them together in a particular pattern are generated through languages, grammars, syntactic and semantic rules.

The limitations in the state-of-the-art are identified in Section 2.10, later in Section 3.1 some basic understandings primarily based upon the inter relation of words in terms of local structures (phrases) and global structures (dependencies) within a sentence and semantic understanding of opinion in terms of opinion orientation and strength are presented. All these limitations and basic understandings provide the background knowledge for the current proposed opinion analysis approach. This current section gives a detailed understanding of the proposed approach which is followed by a number of sections providing the details of different levels of proposed approach.

The local structures and their dependencies give insight into the semantic roles of different parts of sentences. These semantic roles clarify the opinion holders and opinion topics for any opinion expressed in textual data. The semantic roles further help for the analysis at the topic level. This topic level analysis gives a basis for aggregation and summarisation of opinion orientation and opinion scores at higher level of granularities.

The analysis (presented in the literature review and Section 3.5.1) has argued the clause level of granularity to be the most appropriate. In relation to this, linguistic and semantic rules can be used to work together for opinion analysis of textual data at clause level.

The data is pre-processed using some established NLP tools and techniques, which help in obtaining POS, the constituent tree, and dependency tree across each sentence. At the same time, all the opinion based words within the textual data are pre-assigned with the opinion orientations and their intensities by using a suitable lexical resource as shown in Figure 3-1. During the analysis of lexical resources in Chapter 4, it is observed that no single lexical resource is complete. Therefore, improvement in the lexical resource domain improve the performance of the overall opinion analysis process, but the improvement of the lexical resources do not fit within the scope of the current research.

During the preliminary stage of pre-processing dependency trees provide bigrams. Bigrams are part of the output of the dependency parsers (which gives grammatical relations and two words from the sentence, such that first word is determined to be a governor and the second word is determined to be dependent; bigrams are made up of both the related words). The extracted bigrams are analysed based upon POS and grammatical relations for generation of phrases.

Section 3.5.2 provides details of how different grammatical relations, their bigrams and respective POS correspond to different phrases. The patterns of the POS tags associated with each grammatical relation and an analysis of other corresponding relations provide local structures (phrases) and dependencies. This whole process help to generate a revised dependency tree based upon the phrases instead of words as generated through a refined analysis of the typed dependencies and constituent tree together.

Using the nested structure of a sentence based upon the revised parsed tree, with all the POS (words) already assigned with an opinion orientation and opinion intensities in pre-processing, the scores are calculated using the formula and rules explained further in this chapter in Section 3.5.4. An opinion score is attached to each opinion oriented word and later the scores are aggregated and calculated for all the phrases. These phrase level scores are aggregated for clauses. The clause level scores are attached to Opinion Topics associated with each clause. The clause level opinion topic and overall document topic are analysed. If the clause level opinion topic and document topic have any relation, then the opinion is aggregated and calculated to generate an overall opinion of the document, according to the relations.

The extraction of opinion words, their respective opinion holders and opinion topics are identified by resolving the grammatical relations used in the revised phrase level dependency tree identified above. The scopes for conjunction and negation is also calculated with the help of rules generated on a basis of grammatical dependencies, identified through Literature Review in Chapter 2.

This automated process of opinion analysis generates opinion based words, their respective opinion holders and opinion topics, the orientation and intensities of opinion across each clause; this clause level opinion orientation and intensity is aggregated based upon further analysis of conjunction rules and topic level analysis.

Functional relations of words are detected using typed dependency parsing, with refined analysis of grammar and semantics of textual data.

3.4.3.1 Algorithm

```
For each sentence
    Generate dependency structure
    Generate constituent tree
    Identify POS tags
    Regenerate phrase level dependency structure using POS patterns and grammatical dependencies
    Break each sentence into clauses
    For each clause
        Calculate polarity of each phrase
        Negation identification
        Calculate clause level polarity
        Identify subjects, objects, verbs, adverbs and adjectives
        Calculate clause value [Subject (noun + adjective) + (verb + adverb + object / complement)]
    End For
    If Subjects and Objects (Noun Phrases) are same or belong to same feature
    Add polarity of each clause to find out the polarity of each sentence using additional conjunction rules
    End If
    Generate Tuples using identified Parts of Sentence and revised dependency structure
    Save processed information into Corpus
End For
```

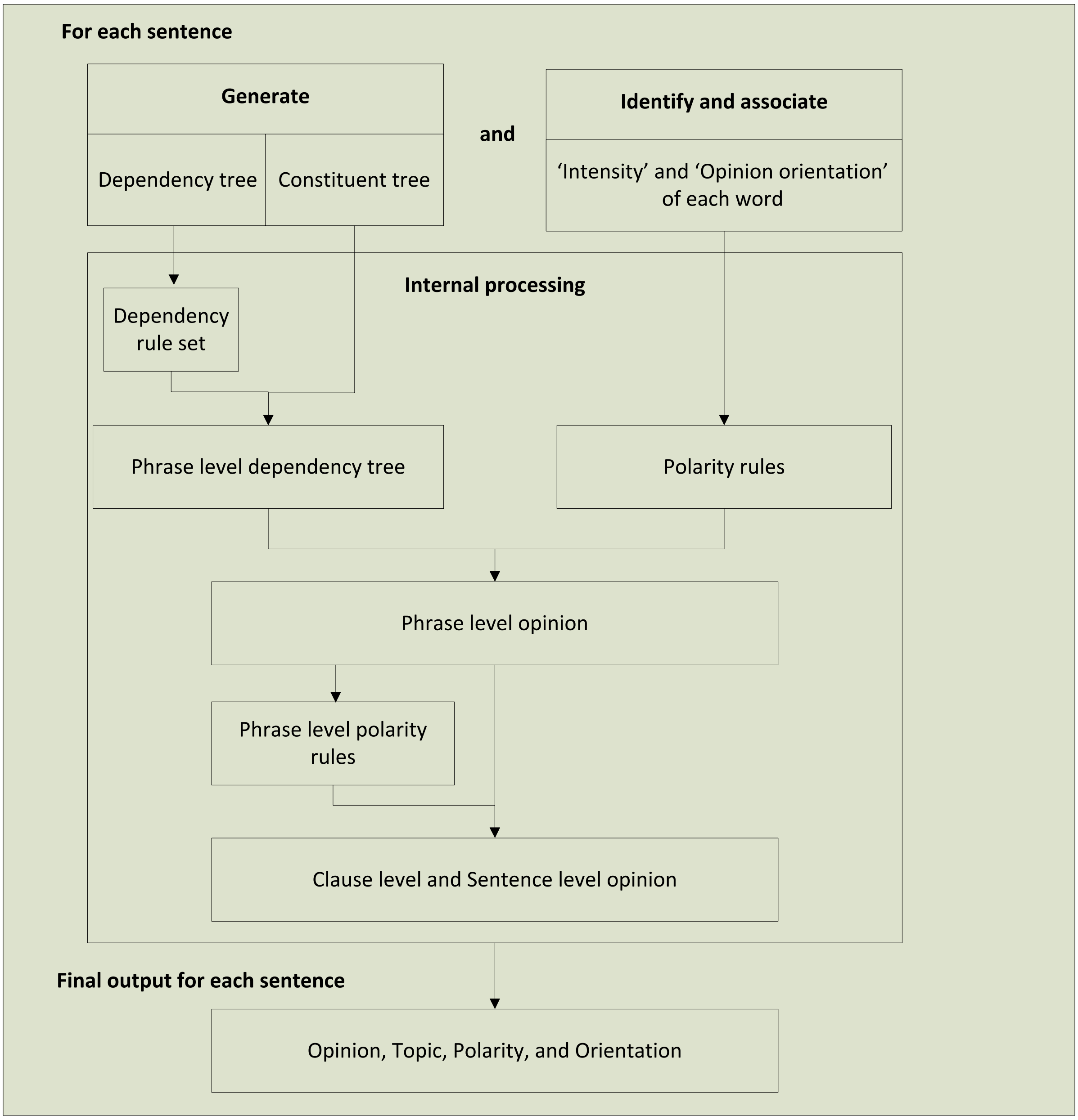

For each sentence
Generate
Dependency tree
Constituent tree
and
Identify and associate
'Intensity' and 'Opinion orientation' of each word
Internal processing
Dependency rule set
Phrase level dependency tree
Polarity rules
Phrase level opinion
Phrase level polarity rules
Clause level and Sentence level opinion
Final output for each sentence
Opinion, Topic, Polarity, and Orientation

Figure 3-1: The graphical representation of the opinion analysis process

3.5 In Depth Analysis of Different Components and Decisions Made For Opinion Analysis

3.5.1 Granularity

In order to provide an understanding of how the proposed opinion analysis approach operates, there is a need to provide information regarding how individual documents are to be analysed and de-constructed. In this section a detailed description is provided for how the level of granularity is approached for each sentence. To enable this, there is a need to understand the basic sentence structure in the English language along with the type of dataset in order to make any decision about the level of granularity for opinion analysis.

Linguists have problems in proposing an agreed definition of ‘sentence’ as structurally sentences can be analysed in more than one way depending upon many factors (Harley, 2007; Bird et al., 2009). For example: the synthetic model: morpheme ➔ word ➔ phrase ➔clause ➔ sentence; and the analytic model: sentence ➔ clause ➔ phrase ➔ word ➔ morpheme (Moore, 2002), are two of many ways of analysing sentences. For this Thesis a sentence is taken as a set of words grouped together in order to convey an idea, description or an event in the form of a constituent structure. A constituent structure is based upon the observation that words combine with other words to form units (Bird et al., 2009). The main reason for using a constituent structure of sentence is substitutability i.e., the sequences of words in a well formed sentence can be replaced by a short sequence without ruining the sentence. The example is shown in Figure 3-2.

<table>
<tr><td>**Det**
The</td><td>**Adj**
little</td><td>**N**
bear</td><td>**V**
saw</td><td>**Det**
the</td><td>**Adj**
fine</td><td>**Adj**
fat</td><td>**N**
trout</td><td>**P**
in</td><td>**Det**
the</td><td>**N**
brook</td></tr>
<tr><td>**Det**
The</td><td colspan="2">**Nom**
bear</td><td>**V**
saw</td><td>**Det**
the</td><td colspan="3">**N**
trout</td><td>**P**
in</td><td colspan="2">**NP**
it</td></tr>
<tr><td colspan="3">**NP**
He</td><td>**V**
saw</td><td colspan="4">**NP**
it</td><td colspan="3">**PP**
there</td></tr>
<tr><td colspan="3">**NP**
He</td><td colspan="5">**VP**
ran</td><td colspan="3">**PP**
there</td></tr>
<tr><td colspan="3">**NP**
He</td><td colspan="8">**VP**
ran</td></tr>
</table>

Figure 3-2: An example for substitution of words and their grammatical categories, Bird et al. (2009)

The research presented in this Thesis is based on the English Language and uses user generated data. The user generated online data does not follow the strict rules of language and takes a flexible structure. Therefore substitutability is an important aspect and may help in the analysis. Further, substitutability enables greater adaptability regarding the Subject-Predicate structure of English sentences, this is further detailed in the current section while explaining clauses. The structure of a sentence in different languages may differ. As the sentence structure of English (Subject-Verb-Object) and German (Object-Verb-Subject) languages are very different. The word order of any sentence is important and it can be defined with the help of a few basic rules of grammar. The boundaries of a sentence are documented as; it begins with a capital letter and terminates with a terminal punctuation mark (period, question mark, exclamation mark). Words are organised into subgroups of words within a sentence. These subgroups have their own structures and linguists represent them as syntactic structures. The simplest structure in the English language is Subject-Verb or Subject-Verb-Object. simplified as Subject-Predicate.

There are three types of sentence in English; a simple sentence, compound sentence and complex sentence. Each sentence is constructed from independent or dependent clauses. There are rules in formal English to identify and differentiate dependent and independent clauses; for example, the second clause in a sentence is independent if there is a comma used before the coordinating conjunction. “Jim studied in the Sweet Shop for his chemistry quiz, but it was hard to concentrate because of the noise.”

Each clause consists of a subject and a predicate, where the subject is based on a noun phrase and a predicate is an arrangement of object, verb, complement and adverbial phrase. A simple sentence contains only one independent clause and a compound sentence contains two or more independent clauses, whereas, a complex sentence has a minimum of one independent and one dependent clause. There are seven different types of clauses defined as; subject-verb (S-V), subject-verb-object (S-V-O), subject-verb-complement (S-V-C), subject-verb-adverbial phrase (S-V-A), subject-verb-object-object (S-V-O-O), subject-verb-object-complement (S-V-O-C) and subject-verb-object-adverbial phrase (S-V-O-A). Table3-3 contains examples of each type of clause.

Clauses are further divided into grammatical structures based on the basic POS, these structures are called phrases. Phrases are sets of words which broadly belong to the same category. For example, a Noun Phrase (NP) is made up of an initial determiner (DT), then an adjective (ADJ), then a noun (N), a complete set of tags used in the Penn Treebank Project is presented in Appendix B.

A clause is a group of related words containing at least a subject and a verb. A clause can easily be distinguished from a phrase, which is a group of related words not containing a subject-verb relation. For example, “in the morning”, or “running down the street”.

The general structure of a complete English Sentence based on POS can be

```
(Sentence
   (Noun Phrase (Pronoun, Noun))
   (Adverbial Phrase (Adverb))
   (Verb Phrase (Verb)
     (Sentence
       (Verb Phrase (Verb)
         (Noun Phrase (Noun))
        )
     )
   )
 )
```

The subcategory Sentence depicts the presence of a clause.

Table 3-3: Types of clauses in English sentence

Clause Type	Sentence
S-V	John sleeps. John (S) sleeps (V).
S-V-O	John has car. John (S) has (V) car (O).
S-V-C	Car is red. Car (S) is (V) red (C).
S-V-A	The teacher is over there. The teacher (S) is (V) over there (A).
S-V-O-O	He gave her a car. He (S) gave (V) her a car (O, O).
S-V-O-C	She thought the car rather expensive. She (S) thought (V) the car (O) rather expensive (C).
S-V-O-A	He parked his car in the garage. He (S) parked (V) his car (O) in the garage (A).

Levels of granularities for opinion analysis are defined on a basis of hierarchy of a document i.e., document, paragraph, sentence, clause, phrase, words. Opinion analysis at all these different levels of granularities and their limitations have been analysed in detail in Section 2. 8. The decision about the level of granularity to analyse for current research is taken on the basis of the structure of the opinion to be extracted (defined in Section 3.4.3), and the features of the dataset used. As the proposed opinion analysis approach mainly targets extraction of an opinion tuple where an opinion topic relation is important, it is identified this relation is similar to verb-object relation in English language. The most refined level where this verb-object relation can be extracted is clause level.

The dataset used for this research is defined in Section 3.4.2. The dataset is comprised mostly of product review data. This genre of data is generally very rich in terms of the hierarchical structure of the products (object of opinion). One product can has multiple features and can be reviewed by multiple reviewers. The products and their features are generally not described through single words. They are generally phrases. The proposed opinion analysis is a lexicon based technique. However, it cannot solely rely on the opinion orientation of words as a basic unit (BoW approach) as this may lead to confusion at later stages of analysis, as it could result in incomplete information. Therefore opinion analysis of this form of data should incorporate phrases into it.

As explained earlier, phrases differ from clauses based upon their structure. Phrases barely communicate complete meaning and opinion as they don’t have any subject within them. Whereas, the structure of opinion as detailed earlier in this chapter requires an opinion and opinion topic relation. It is something similar to the verb-object structure of a sentence.

Therefore while analysing comparative, complex and compound sentences, it is identified that there are generally more than one opinion expressed about more than one object. For example; in the sentence “I highly recommend the Canon SD500 to anybody looking for a compact camera that can take good pictures.”, ‘the Canon SD500’and ‘pictures’ are objects whereas, ‘recommended’ and ‘good’ are opinion bearing words. However ‘the Canon SD500’ is connected to ‘recommended’ by the author, but it cannot be concluded that the ‘pictures’ taken are also ‘recommended’. The correct relations cannot be generated by a Cartesian product of opinion bearing words and objects.

The clause is a unit in a sentence which has its own subject and object, so using parts of sentences the relations can be identified. Therefore clause level opinion analysis utilising the hierarchical structure of a sentence (phrases and words) can be used in conjunction to deliver a more refined level of granularity.

3.5.2 Parsing

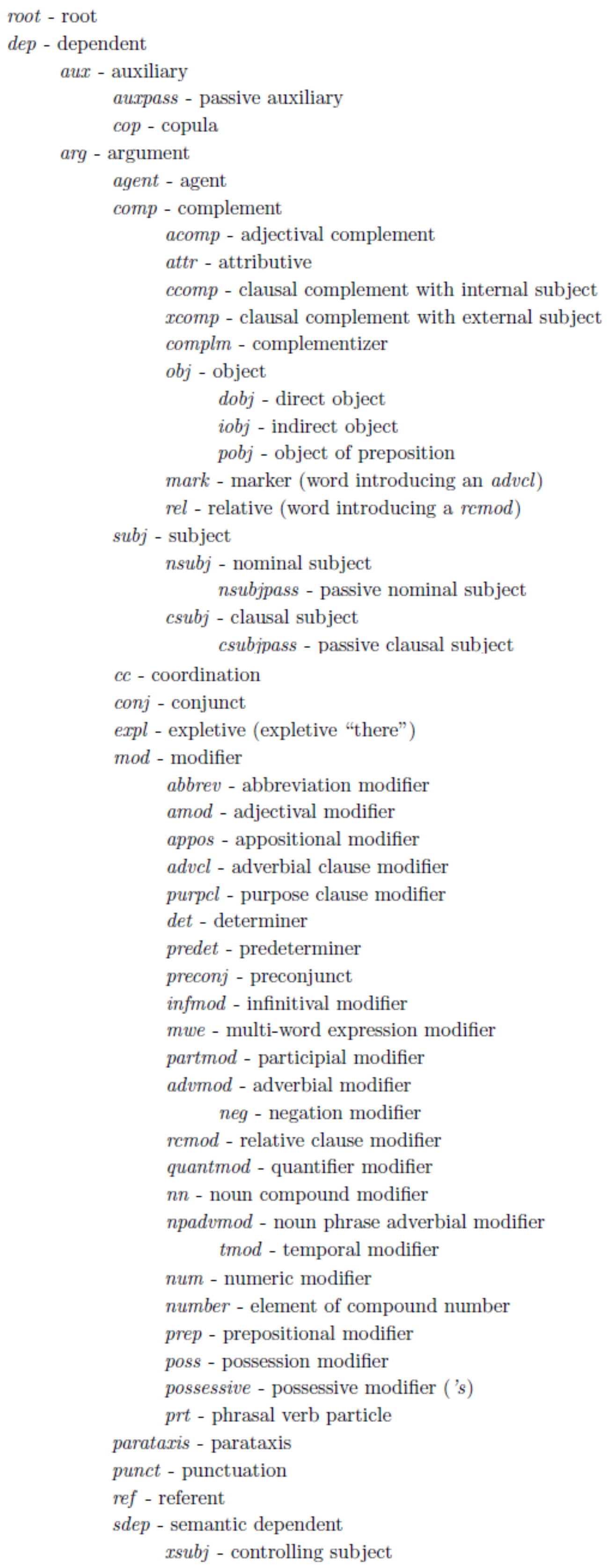
Any textual analysis application depends upon the basic process of parsing a sentence. The first question that arises is what is parsing? Parsing is a process of reading and breaking a string of elements, into its components, based on a defined set of rules. There are different ways of parsing for sentences in the English language. These approaches include mechanisms for defining the structure of a sentence as being flat or tree like (Schlenker, 2006). Parsing depends upon different requirements of analysis, as a parsed sentence can be parsed and re-examined for different contexts.

Dependency parsing is conducted using syntactic elements; relations are built between binary relations between single words. However, it doesn’t provide information about the local structure and syntactic categories (phrase level POS) of words. Thus a dependency tree can be used to provide the relations and connections between distinct words. Dependency parsing is very useful for the extraction of relations where words are separated by more than one word. The combination of dependency parsing and an analysis of syntactic categories can be used to provide a dependency tree giving local structures. In addition, this combination can also give a good insight into local/global dependencies and relations.

Previous work (Marneffe et al., 2006; Wu et al., 2009) on relation extraction has usually used a head word to represent the whole phrase and to extract features from a word level dependency tree. Unfortunately, this approach is inadequate in providing an extended phrase level analysis e.g. the product topic analysis (Yi et al., 2003; Stoyanov and Cardie, 2008; Mukherjee and Bhattacharyya, 2012).

In order to solve this issue there is a need to *introduce* the concept of a phrase level dependency tree (Marneffe et al., 2006, Wu et al., 2009, Tan et al., 2011). Many tools are available in the context of NLP which can provide POS tagging, the development of a constituent tree, and dependency parsing for sentences based upon words. However, there are no tools available which can provide dependency parsing of phrases. Therefore some rules can be introduced and utilised along with the dependency parsing, POS tagging and constituent tree solution provided by the Stanford Parser and Penn TreeBank. The Stanford Parser and Penn TreeBank are discussed in detail in Section 3.5.3.

The proposed opinion analysis approach utilises the rules and patterns applied through dependency and constituent trees in order to generate phrase level dependency parsing. There are 56 grammatical relations defined by the Stanford Parser (Www.Stanford.Edu, n.d), which have been analysed for their definitions and their respective POS. These relations follow an internal hierarchy. For example; auxiliary, argument and modifier relations can be interpreted for dependent relations. However, an argument relation can be further divided into complement and subject relations and so on. The hierarchy of the relations identified by Stanford Parser is shown in Figure 3-3. In extension to this, the most common grammatical relations for each of the phrases can be identified. For example, the identified grammatical relations for Noun Phases are given in Table 3-4



- *root* - root
- *dep* - dependent
    - *aux* - auxiliary
        - *auxpass* - passive auxiliary
        - *cop* - copula
    - *arg* - argument
        - *agent* - agent
        - *comp* - complement
            - *acomp* - adjectival complement
            - *attr* - attributive
            - *ccomp* - clausal complement with internal subject
            - *xcomp* - clausal complement with external subject
            - *complm* - complementizer
            - *obj* - object
                - *dobj* - direct object
                - *iobj* - indirect object
                - *pobj* - object of preposition
            - *mark* - marker (word introducing an *advcl*)
            - *rel* - relative (word introducing a *rcmod*)
        - *subj* - subject
            - *nsubj* - nominal subject
                - *nsubjpass* - passive nominal subject
            - *csubj* - clausal subject
                - *csubjpass* - passive clausal subject
        - *cc* - coordination
        - *conj* - conjunct
        - *expl* - expletive (expletive “there”)
        - *mod* - modifier
            - *abbrev* - abbreviation modifier
            - *amod* - adjectival modifier
            - *appos* - appositional modifier
            - *advcl* - adverbial clause modifier
            - *purpcl* - purpose clause modifier
            - *det* - determiner
            - *predet* - predeterminer
            - *preconj* - preconjunct
            - *infmod* - infinitival modifier
            - *mwe* - multi-word expression modifier
            - *partmod* - participial modifier
            - *advmod* - adverbial modifier
                - *neg* - negation modifier
            - *rcmod* - relative clause modifier
            - *quantmod* - quantifier modifier
            - *nn* - noun compound modifier
            - *npadvmod* - noun phrase adverbial modifier
                - *tmod* - temporal modifier
            - *num* - numeric modifier
            - *number* - element of compound number
            - *prep* - prepositional modifier
            - *poss* - possession modifier
            - *possessive* - possessive modifier (*’s*)
            - *prt* - phrasal verb particle
        - *parataxis* - parataxis
        - *punct* - punctuation
        - *ref* - referent
        - *sdep* - semantic dependent
            - *xsubj* - controlling subject

Figure 3-3: Hierarchy of the relations in Stanford Parser, Marneffe et al. (2006)

Table 3-4: Identified grammatical relations for Noun Phases

Relation	Description	POS
Amod	Adjective modifier	(noun, adjective)
Rcmod	Relative clause modifier	(noun, verb)
Det	Determiner	(noun, determiner)
Partmod	Participial modifier	(noun phrase/verb/phrase/clause, verb)
Infmod	Infinitive modifier	(noun, noun/adjective)
Prep	Proposition modifier	(word, preposition)
Aposs	Apposition modifier	(noun, noun/adjective)
Nn	Noun compound	(noun, noun)
Num	Numeric modifier	(noun, number)
Number	Element of compound number	(currency/unit, number)
Abbrev	Abbreviation	(Noun, abbreviation)
Cc	coordinating conjunction	(noun, conjunction)

There are many relations which are used to explain the internal relations of the noun phrase (NP), however the most common are det, nn, aposs, and cc. In addition, there is a need to identify the head of each constituent of the sentence. This head is not the syntactic head but preferably a semantic head e.g. subject, object of clause/sentence. The head words for noun phrases are generally from the category of ‘subj’ and ‘obj’ in the hierarchy of Stanford Parser relations as shown in Figure3-3.

Similarly any Verb Phrase (VP), Adjectival Phrase (ADJP) and Adverbial Phrase (ADVP) also need to be identified. The nested structure of a sentence is presented in Section 3.5.1 and similarly the phrases are presented in a constituent tree. The phrases identified in the above process are cross checked with the phrases presented in the constituent tree. A head is identified for all of these phrases. The head is identified based upon the higher level relations (root, subject and objects).

For Example:

Sentence:

The new phone book and tour guide have not impressed me.

Noun Phrase (The new phone book and tour guide)

Verb Phrase (have not impressed (noun Phrase (me)))

Typed dependencies

Determiner (book, the)
Adjectival modifier (book, new)
Compound noun (book, phone)
Subject (impressed, book)
Compound noun (guide, tour)
Conjunction_and (book, guide)
Subject (impressed, guide)
Auxiliary (impressed, have)
Negation (impressed, not)
root (ROOT, impressed)
Direct object (impressed, me)

The Subject of the sentence is 'book', which is a 'new book', this is further modified as a 'new phone book'. There is a ***Conjunction_and*** relation between 'book' and 'guide', which means if 'book' is a ***Subject*** in the sentence, then 'guide' must also share the same relation. Therefore, 'guide' is also a ***Subject***, which is modified by another noun as 'tour guide'. The complete noun phrase (NP) is 'the new phone book and tour guide'. ***Root*** is 'impressed', however the verb phrase (VP) is 'have not impressed' based upon ***Auxiliary*** and ***Negation*** relation. The ***Object*** of the sentence is NP 'me'.

The example sentence is presented as shown in Figure 3-4.

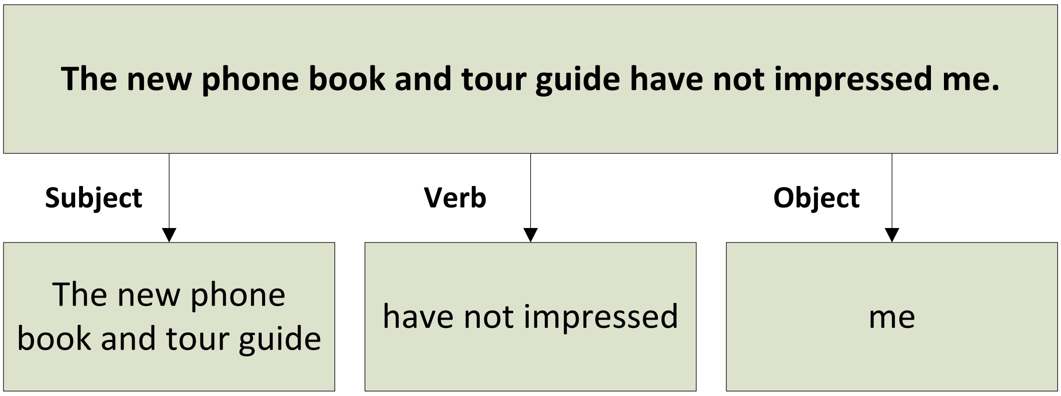

The new phone book and tour guide have not impressed me.
Subject
Verb
Object
The new phone book and tour guide
have not impressed
me

Figure 3-4: Example Sentence

3.5.3 How to Resolve Clauses

As discussed earlier in Section 3.5.1 sentences in English (compound and complex) have more than one clause in them. In more structured English there are some rules which can be encoded in order to automatically identify the clauses. However, there are some issues, one of the issues is that these rules are not few in number and are not stated anywhere as a ruleset, but with the complexity of language the number of rules vary. Another issue is that these rules are structured only for highly formal language, whereas online users of web based forums do not always follow general language structures and rules. In online user generated data, generally two or more sentences can be combined together without any punctuation marks.

Therefore there is a need to identify and resolve clauses based upon the constituent tree and dependency tree generated during the pre-processing phase of sentence parsing. Wilson et al (2004) have used Collin's parser and have used non-leaf verbs in the parse tree as a basis for the identification of clauses.

The Penn TreeBank used in Stanford Parser for the constituent tree uses a number of tags for the identification of clauses as presented in the Table 3-5 and a complete set of tags is presented in Appendix B.

Table 3-5: Clause level Tags and their description for Penn Tree Bank, Bies et al. (1995)

Tag	Description
S	Simple declarative clause, i.e. one that is not introduced by a (possible empty) subordinating conjunction or a wh-word and that does not exhibit subject-verb inversion.
SBAR	Clause introduced by a (possibly empty) subordinating conjunction.
SBARQ	Direct question introduced by a wh-word or a wh-phrase. Indirect questions and relative clauses should be bracketed as SBAR, not SBARQ.
SINV	Inverted declarative sentence, i.e. one in which the subject follows the tensed verb or modal.
SQ	Inverted yes/no question, or main clause of a wh-question, following the wh-phrase in SBARQ.

Based upon the tags in the constituent tree clauses are identified and are parsed based upon the algorithm presented in Section 3.4.3. Opinion words, their respective opinion holders and opinion topics are identified using the set of relations and rules identified by the analysis of the relations.

For example, in the sentence “He is in great problem as he lost his award.”

The sentence has two clause “He is in great problem” and “he lost his award”.

Whereas in another sentence “It was first time he passed, otherwise he rarely succeed in exams.” The sentence has three clauses “it was first time”, “he passed”, and “he rarely succeeds in exams”.

There are some conjunctions and/or prepositions joining these clauses and with the help of conjunction rules their analysis is conducted.

3.5.4 The Scoring Technique

Assigning opinion based scores in different opinion analysis techniques remains a delicate task. Decisions made regarding the generation of opinion scores in other research are always based on a number of elements, including the opinion structure, an understanding of opinion, and the level of refinement required for the analysis. Different scoring methods at various levels of granularity and lexical resources are discussed in Chapter 2 and Chapter 3.

The proposed opinion analysis approach uses a real number based approach for opinion analysis where numbers between the range of -1 (most Negative) and +1 (most Positive) are assigned during opinion analysis. Refined calculation rules are used for more accuracy in the calculation of opinion scores by using, grammatical dependencies, phrase level structures, phrase level dependencies, POS, and word based opinion scores from existing lexical resources.

The description of the proposed opinion analysis approach and sentence parsing as discussed earlier in Section 3.4.3, has introduced the basis for opinion based scoring.

In the pre-processing stage the opinion scores for all the opinion bearing words are assigned. The semantic and syntactic relations between words are captured through dependency parsing and constituent trees. These relations generate phrases; for example, verb phrases, adjectival phrases etc. each type of phrases has its own set of POS and structure. For each phrase, the opinion scores are calculated, using Equation 3.1

$$|\text{Word1}| + (1-|\text{Word1}|) * |\text{Word2}| \qquad \text{Equation 3.1}$$

As these phrases are resolved within each clause so the type of clause (as discussed in Section 3.5.1) is also considered in the calculation of opinion scores. Clauses have a subject-predicate structure, where a predicate can have further sub-parts as object, verb, complement and adverb. The subject/object in any clause is presented as a noun phrase, a verb is presented as a verb phrase, adverb as an adverbial phrase, and a complement as an adjectival phrase.

Clauses can have a more complex structure and one, or more than one, clause can be used to construct a sentence. Therefore opinion analysis is performed by the calculation of opinion scores and a determination of the opinion orientation for all the independent phrases. In addition, further calculations are performed related to these phrases at clause and sentence level and these are discussed in the sections ahead.

3.5.4.1 Subject/object

The subject and object in a clause are presented as a noun phrase. Generally, in a noun phrase the nouns and adjectives are opinion based content, with the rest of the noun phrase being treated as discussed in Section 3.5.2 (shown in Table 3-4) as generally neutral, i.e., does not hold opinion; for example, ‘the’ (determiner), ‘of’ (preposition) etc. When an adjective and noun come together, and they have same opinion orientation, they have a tendency to intensify each other. In order to calculate the score of this kind of noun phrase, the following Equation 3.2 is used and the rules for the combination of positive/negative values of noun and adjective are shown in Table 3-6.

$$|\text{Noun}| + (1 - |\text{Noun}|) * |\text{Adjective}| \qquad \text{Equation 3.2}$$

A positive adjective can intensify a positive noun, and the resulting score should be an intensified (higher than both noun and adjective scores) positive score. Similarly, a positive adjective with a negative noun again intensifies the negative opinion of noun. The negative adjective with a positive noun diminishes the positive opinion of the noun and gives a negative orientation to the resulting score. Whereas, the negative adjective with a negative noun intensifies the negative opinion of the noun. Examples for this are shown in Table 3-6.

Table 3-6: Rules and examples for noun phrase contextual opinion calculation

Adjective	Noun	Output	Examples
+ Great=0.75	+ Surprise=0.125	+ 0.78125	Great surprise
+ Big=0.125	- Disaster=-0.5	- 0.5625	Big disaster
- Lousy=-0.625	+ Star=0.125	- 0.67187	Lousy star
- Worst=-0.875	- Disaster=-0.5	- 0.9375	Worst disaster

3.5.4.2 Predicate

As already discussed the predicate of a sentence may contain a verb phrase, an object (noun phrase), adverbial phrase, and an adjectival phrase/complement. Mostly negation is presented in verb and adverbial phrases. The scope of the negation is identified by dependency parsing. The dependency parser gives grammatical relations between two words from within a sentence. Therefore where negation is identified, then a relationship is identified between the negating word and the word which is directly negated. However, this word based scope does not capture the complete information.

In order to capture the complete scope, further analysis is required. This analysis process captures the direct scope term, identified through the negation relation in the dependency grammar, and further analyse, to check if the negated term holds another modification relation (adjectival modifier, adverbial modifier etc.) with any other term in the sentence/clause. If the term is modified by any other term, the whole phrase is extracted using grammatical relations and a constituent tree is built. The complete scope of negation is identified through this process, with the orientation and score calculated accordingly. Therefore the negation rules are explicitly handled while calculating the opinion scores for phrases. These rules are specified in Section 3.5.5.

In addition, there is a need to accommodate conjunction rules as more than one phrase might be needed to calculate the scores of the predicate, this is explained in Section 3.5.5.

3.5.4.3 Verb phrase

A verb phrase may contain a verb and an adverb. In a verb phrase, a negating adverb can be introduced, as negation terms (not, never etc.) are generally adverbs (NOTE: In circumstances where these are not adverbs these are detected as modifiers). The negating adverb can change the meaning of the original opinion of a dependent verb/adjective. A method of negation calculation (see Section 3.5.5) is used for the identified negating adverb. Equations 3.3 and 3.4 are used in the score calculation of a simple verb phrase or a verb phrase which contains a negating adverb respectively. Whereas, Table 3-7 presents the rules for assigning the opinion orientation, examples of opinion scores and the orientation calculation.

$$|\text{Verb}\ | + (1 - |\text{Verb}|) * |\text{Adverb}| \qquad \text{Equation 3.3}$$

$$\text{Negation } (+/-)\ (1 - |\text{Scope of negation (Verb/Adverb)}|) \qquad \text{Equation 3.4}$$

Table 3-7: Rules and examples for verb phrase contextual opinion calculation

Adverb	Verb	Output	Examples
+ Greet=0.125	+ Cheerfully=0.125	+ 0.2343	Greet cheerfully
+ Proudly=0.125	- Gossip=-0.5	- 0.5625	Gossip proudly
- Wrongly=-0.5	+ Release=0.25	- 0.625	Release wrongly
- Badly=-0.75	- Burned=-0.5	- 0.875	Badly burned

3.5.4.4 *Verb object/complement*

The predicate of a sentence may be comprised of a combination of verb phrases with an object (noun phrase) or a complement (adjectival phrase). There is a need to identify the rules to bring these different phrases together in order to calculate the opinion orientation and scores. Therefore Equations 3.5 and 3.6 are presented. The rules governing the overall opinion orientation and examples of these are presented in Table 3-8.

$$|\text{Object/Compliment}| + (1 - |\text{Object/Compliment}|) * |\text{Verb Phrase}| \qquad \text{Equation 3.5}$$

$$\text{Negation } (+/-)\ (1 - |\text{Object/Compliment}|) \qquad \text{Equation 3.6}$$

Table 3-8: Rules and examples for over all predicate contextual opinion calculation

Verb Phrase	Object/ Complement	Output	Examples
+ Give=0.375	+ Blessing=0.375	+ 0.6093	Give blessings
+ Add=0.25	- Trouble=-0.375	- -0.5312	Adding troubles
- Missed=-0.5	+ Award=0.5	- -0.75	Missed award
- Suffer=-0.5	- Failure=-0.25	- -0.625	Suffers failure

3.5.4.5 *Clause*

A clause includes a subject and a predicate. The equation used to calculate the opinion score at clause level is presented in Equation3.7. The opinion orientation rules and examples are presented in Table 3-9.

$$|\text{Predicate}| + (1 - |\text{Predicate}|) * |\text{Subject}| \qquad \text{Equation 3.7}$$

For example; the scores of subject ‘superstar’ is +0.125 (calculated using Equation 3.1) and predicate ‘greeted cheerfully’ is 0.2343 (calculated using Equation 3.5), so the score of the clause ‘superstar performed poorly’ can be computed using the Equation 3.7.

Table 3-9: Rules and examples for simple clause contextual opinion calculation

Subject	Predicate	Output	Examples
+ Superstar=0.125	+ Greeted cheerfully=0.2343	+ 0.33	Superstar greeted cheerfully.
+ Superstar=0.125	- Badly burned=-0.875	- 0.8906	Superstar was badly burned.
- bad=-0.75	+ plausible=-0.5	- 0.875	It was bad casting for a plausible script.
- bad=-0.75	- spoil=--0.125	- 0.78125	The bad casting spoiled the script

3.5.4.6 Complex clauses

A complex sentence contains multiple clauses (complex clause). To calculate the score of the complex clause, the score of the each clause is calculated and then all clauses are added using Equation 3.8 and the simple opinion orientation rules as shown in Table 3-10.

$$|\text{Clause 2}| + (1 - |\text{Clause 2}|) * |\text{Clause 1}| \qquad \text{Equation 3.8}$$

$$\text{Negation } (+/-)\ (1 - |\text{Clause}|) \qquad \text{Equation 3.9}$$

For example, in the sentence “This camera takes amazing images and its size cannot be beaten.”, ‘this camera’, ‘amazing images’ and ‘its size’ are noun phrases, and ‘takes’ and ‘cannot be beaten’ are verb phrases. There are two clauses in this sentence, ‘this camera takes amazing images’ and ‘its size cannot be beaten’.

The sentence hierarchy below shows both these clauses separated by a tag ‘(S’**.** There is a conjunction ‘(and)’ between both these clauses.

```
(S
  (NP This camera)
  (VP takes
        (NP amazing images)))
         (and)
(S
(NP its size)
(VP cannot be beaten))
```

All phrases are resolved for their scores and orientations and then their scores and orientations are aggregated using clause based methods. Finally, an aggregation of all clauses is calculated based upon the orientation rules.

Table 3-10: Methods for complex clause contextual opinion calculation

Clause A	Clause B	Output	Examples
+	+	+	This camera takes amazing images and its size cannot be beaten.
+	-	-	I will recommend to cancel the trip
-	+	-	It is hard to see this match
-	-	-	It is hard to find problem

Negation is a complex phenomenon, which can be used in more than one way, i.e., denials or rejections. The example of denial is "The food quality of this restaurant is not very good." and an example for rejection is "Given the poor reputation of the restaurant, I expected to be disappointed from food.". Researchers have undertaken different ways of handling negation for opinion analysis, for example: some use lists of words with different levels of denials i.e., lacked, rarely, hardly etc.; (Choi and Cardie, 2008; Li and Wu, 2010). Wilson et al. (2005) uses words and syntactic patterns to determine the negation features in written text ;and Moilanen and Pulman (2007) used morphological negation features i.e., prefixes (dis, non) and suffixes (less) to determine the effects of negation on polarity calculation in written text. In the current research negation is only considered in the case of denials i.e., explicit use of no, not, never, doesn't etc. for negation cues. The proposed opinion analysis technique is mainly dependent on linguistic features, typed dependencies and lexical resources. The lexical resources take care of all diminishers and intensifiers in terms of assigning the opinion based scores to the words. The scope of explicit negation terms are identified and scores are calculated using rules and equations specified in Equations 3.6, 3.9 and 3.11.

Negation has a property to reverse the actual polarity of words. However, the application of this property does not always result in an exact reversal. For example, given the sentences: 1) “This story is good.”; and 2) “This story is not good.” Sentence 2 has negation but it is not exactly reversing the opinion expressed in 1. ‘not good’ does not suggest the story is completely the opposite of ‘good’ i.e. ‘bad’. Here negation is only diminishing the overall impact of opinion expressed in 1. Therefore, the dependency structure for negation has to be handled carefully both using POS as well as the dependency tree. Negation words are generally depicted as ‘adverbs’. There have to be rules specified based upon the scope of negation especially in complex sentences/clauses. Table 3-11 is exhibiting some rules for negation. These rules are defined on the basis of POS. Most negation words are classified as adverbs, suffix, prefix or verbs. However, the nouns are generally there to determine the meaning of another noun. The scope of negation will be identified by the dependency tree, which indicates how negation is interacting with other words in the sentence. This dependency will identify the scope of the negation - whether it is a single word or a phrase / clause within a sentence. In the case of a clause or phrase, the noun phrase/ clause is first calculated for the sentiment polarity before the verb phrase or clause sentiment polarity is calculated. The negation is handled in each phrase accordingly. The intensity of polarity will not exceed (+/-) 1, where + is for positive and – is for negative polarity. The intensity of a sentence is calculated as in Equation 3.1.

The positive/negative value of words in the Equation 3.1 is extracted from the SentiWordNet in order to calculate the polarity of a sentence. The extracted value from the SentiWordNet is reversed during this process if negation is ‘True’ as presented in Table 3-11

.

Table 3-11: Rules Specifying Negation

First Word /Phrase /Clause	Second Word /Phrase /Clause	Negation	Result
Positive	Positive	TRUE	Negative
Positive	Positive	FALSE	Positive
Positive	Negative	TRUE	Positive
Positive	Negative	FALSE	Negative
Negative	Positive	TRUE	Positive
Negative	Positive	FALSE	Negative
Negative	Negative	TRUE	Negative
Negative	Negative	FALSE	Positive

For example: a sentence "They have not succeeded, and will never succeed, in breaking the will of this valiant people.", can be deconstructed as shown below:

```
(S
  (NP They)
  (VP
  (VP (have not)
  (VP (V succeeded)))
(and)
  (VP (will)
  (ADVP (ADV never))
  (VP (succeed)))
  (PP (in)
    (S
    (VP(breaking)
    (NP
    (NP(the will))
    (PP (of)
    (NP (this valiant people))))))))
)
```

The negation word ‘not’ is affecting the ‘succeeded’ (+) whereas ‘never’ is effecting ‘succeed’ (+) where ‘succeeded’ and ‘succeed’ are joined by ‘and’ (joins with the same polarity). Both successes are in ‘breaking’ (-) the will of people who are ‘valiant’ (+) people. As they have not succeeded in doing something Negative and the polarity of sentence is Positive as shown in Figure 3-5.

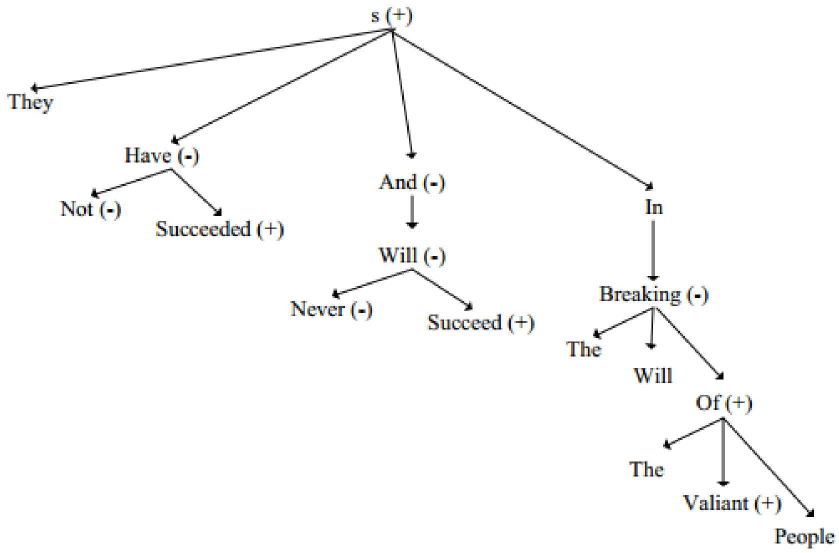


Figure 3-5: Dependency Tree Break up of a sentence with negation and conjunction

3.5.5 Generalized Scoring Method

It becomes very difficult to comprehensively cover all grammatical relations in each phrase and clause, as each phrase may contain multiple grammatical relations and the same relations may appear in other phrases.

There is a need to identify and specify default and generalized rules which can be applied to unspecified and unmatched phrases. The need for generalized rules is also strengthened as there can be many patterns and types of sentences in any language as the online user generated content can be produced by users from diverse backgrounds. Therefore, this user generated content may not follow formal rules of language.

Based on the previously given equations (Equation 3.1 to 3.8), the generalized Equation 3.10 is presented. If two terms (positive/negative) modify each other, then the score is calculated using Equation 3.10 and the opinion orientation is calculated using the rules specified in Table 3-12.

$$|A| + (1 - |A|) * |B| \qquad \text{Equation 3.10}$$

Table 3-12: General rules for contextual opinion calculation

Term A	Term B	Output
+	+	+
+	-	-
-	+	-
-	-	-

3.5.5.1 Negation Identification

Handling of negation requires more sophistication in opinion analysis. It is considered as one of the key processes in opinion analysis, both in the scoring and in the calculation of the orientation of the opinion. Some example sentences and their opinions are presented in Table 3-13.

Table 3-13: Examples for Negation

Sentence Number	Sentence	Opinion Score	Opinion Orientation
1	I am happy.	0.125	+
2	I am not happy.	0.875	-
3	I am very happy.	0.5625	+
4	I am not very happy.	0.4375	-

There is a need to understand the scope of negation. The negation ‘not’ is used in sentences 2 and 4. The scope of negation in both these sentences specifies different meanings. In sentence 2 negation is only affecting the word ‘happy’, whereas, in sentence 4 negation is affecting the phrase ‘very happy’. The effect of the negation word ‘not’ on the whole adjectival phrase ‘very happy’, is to change the opinion orientation of the complete phrase limiting the negation. This means that sentence 2 has a greater negative score than sentence 4 e.g. ‘not very happy’ is not as negative as ‘not happy’.

In sentence 4 the negation relation is identified for the word ‘happy’ and ‘happy’ has an adverbial modifier ‘very’. The word ‘very’ is not further modified by any other word in the sentence therefore it completes the phrase ‘very happy’. Therefore, in sentence 4, the scope of negation is the complete phrase ‘very happy’.

Negation can be calculated using the Equation 3.11 and Table 3-12 specifies the rules for orientation for scope of negation as specified in Table 3-14.

$$|N| * (1 - |T|) \qquad \text{Equation 3.11}$$

Table 3-14: Methods for negative contextual opinion calculation

Negation	Term	Output	Examples
-	+	-	Not fine
-	-	+	Not terrible

3.5.5.2 Conjunction rules

Clauses can sometimes be conjoined with the help of conjunction words. These conjunction words can affect the opinion orientation of the proceeding or preceding clauses. For example, the conjunctions ‘and’ and ‘which’. Both these conjoining words ‘and’ and ‘which’ mean that the conjoined clauses are having a similar opinion therefore they are intensifying each other’s opinions. The normal rules for orientation as specified in Table 3-10 are followed.

The conjunctions ‘but’ and ‘however’ mean the opinion of the preceding clause is to be weakened and minimized by the opinion of the proceeding clause. For example, “This camera is compact but is really heavy”. In this case the opinion of first clause is reversed by using the negation Equation 3.11.

The conjoining word ‘or’ is used when the opinion of both conjoined clauses are opposing and normal opinion scoring rules as specified in Table 3-11 are used.

3.5.6 Sentence Level Opinion Score Calculation

As explained in Section 3.4.3 the sentence is first parsed using a dependency parser, Penn Treebank parser and a lexical resource.

For example, the sentence *"I highly recommend the Canon SD500 to anybody looking for a compact* camera *that can take good pictures.",* is parsed using all three resources. Table 3-15 gives POS, opinion scores and opinion orientation of opinionated terms in the sentence.

Table 3-15: Examples for POS and opinion scores

Word	POS	Opinion Scores
Highly	adverb	0.375
recommend	Verb	0.375
looking	Verb	0.125
compact	adjective	0.375
Take	Verb	0.125
Good	Adjective	0.5

The sentence is parsed as

```
(S
   (NP I)
   (ADVP highly) 0.375
      (VP recommend) 0.375
     (NP the Canon SD500)
 (PP to)
    (NP anybody)
    (VP looking) - 0.125
        (PP (IN for)
      (NP a compact camera)
  (WHNP that)
      (VP can take) 0.125
      (NP good pictures) 0.5
   )
```

The above demonstrates that the sentence is divided into three clauses through this process: "I highly recommend the Canon SD500""anybody looking for a compact camera", and "that can take good pictures".

In the first clause “I highly recommend the Canon SD500”. ‘I’ is the ***subject*** and ‘the Canon SD500’ is the ***object.*** ‘Highly’ is an ***adverbial modifier*** for ‘recommend’. “Highly recommend” as a phrase gives a score of +0.609 from the Equation 3.3. And as ‘I’ and ‘the Canon SD500’ are neutral, the overall score of the first clause is +0.609.

For the second clause “anybody looking for a compact camera” has ‘Camera’ as the ***Object*** however; it is a noun phrase “compact camera”. ‘Camera’ is neutral so “compact camera” is +0.375. According to Equation 3.2. ‘Looking’ is a ***verb*** having an opinion score of 0.125. The predicate has an overall opinion score of 0.531 according to Equation 3.4.

In the third clause “that can take good pictures”, ‘that’ is ***Subject*** and ‘pictures’ is ***Object***. ‘Pictures’ is ***modified*** with ‘good’. So according to Equation 3.2 the noun phrase “good pictures” has opinion +0.5 and the overall predicate “take good pictures” has +0.5625.

As all three clauses are about the features of the same ***object*** ‘camera’ and are from the same ***opinion holder*** ‘I’ therefore, the overall opinion score and orientation for the whole sentence can be calculated using Equation 3.7. The calculation of opinion for the overall sentence starts with calculation of the opinion scores for the bottom most clauses, i.e. calculation from right to left in the sentence of the opinion score. So the overall opinion scores for the second and third clauses are 0.531 and 0.5625 respectively. According to Equation 3.7 the score is +0.7948. This 0.7948 is aggregated with +0.609. So the aggregate score for the whole sentence is +0.919. This score means it is a highly positive opinion that is being expressed in the sentence.

In all the above steps, a generalized scoring method is used to calculate the score and finally an aggregated value is declared as the polarity of the given sentence. If the objects (Opinion Topics) of all the clauses are not the same, there is a need for further topic based analysis, in order to aggregate the scores. This might require additional resources (topic based, topic hierarchical) i.e., ontologies explaining products and their respective features and attributes. Currently only WordNet based basic relations are used for this Thesis.

3.5.7 Other Issues

While performing opinion analysis especially the division of sentences at clause level, it is observed that in most sentences a proper noun (opinion holder/opinion topic subject/object) is only mentioned once. Mostly, nouns are represented by using related pronouns. As sentences are divided into their hierarchical structure (clauses) and the opinion in each clause captures the opinion holder, and opinion topic for each clause. It becomes difficult to find the exact opinion holder and the opinion topic because of the use of pronouns instead of proper nouns. This can be illustrated through analysis of the following two sentences:

1) Nokia-N8 is a good mobile.
2) It has a large screen.

In sentence 1, ‘Nokia-N8’ is the ***subject,*** and the ***adjectival phrase*** “a good mobile” is holding an opinion about the ***subject*** (Nokia-N8). Whereas, in sentence 2, ‘it’ is the ***subject*** and ‘screen’ is the ***object***. An ***adjectival modifier*** ‘large’ is modifying the ***object*** (screen), which makes a ***noun phrase*** “a large screen”. Here, the resolution of ‘it’ (***subject***) is necessary in order to aggregate the opinion and capture a correct understanding of the opinion expressed.

good mobile ➔ Nokia-N8

larger screen ➔ It

It ➔ ???

This is a significant limitation of the state-of-the-art and steps towards resolving this could involve the use of noun co-referencing for given inputs before the execution of any syntactic or semantic parsing. Noun co-referencing has the potential to significantly improve opinion analysis in the sentences above, however, this does not fall under the scope of the current Thesis.

Another important step performed during the pre-processing stage is the assignment of opinion scores to all the opinion based words in the textual data. This process of assignment of opinion orientation and scores to all the words is carried out with the help of a lexical resource. The approach proposed in this Chapter uses SentiWordNet 3.0 as the lexical resource for opinion based word determination, which is a word based resource. While using this resource it is found that there can be more than one sense associated with each word, this causes problems.

Words which have different meanings and are spelt the same way are very common in English writing. These kinds of words cause problems in understanding the meaning from the text. For example; the word 'bank' has a number of meanings including financial institution, and step, or edge as in snow bank, or river bank. Therefore there can be a difficulty in the resolution of Word Senses. This is considered a very common problem in NLP. The process of resolving this issue is called Word Sense Disambiguation (WSD). Consideration of this problem would help to improve results from the proposed approach. However, it is also a separate area of research in NLP and does not fit into the scope of the current Thesis.

Similarly as mentioned in Section 3.4.3, and Section 3.5.6 the availability of a resource (ontology) related to the dataset, i.e., the field of research, providing the hierarchical structure possible list of aspects, Opinion Topics and their features and attributes, for opinion analysis provides the potential to be used coupled with the proposed approach in order to provide improved results.

3.5.8 Limitations and Strengths of Proposed Opinion Analysis Approach

The proposed novel approach for opinion analysis is a clause level approach, utilizing syntactic and linguistic analysis for the resolution of the scope of different words within a sentence, and for understanding interaction between the words, i.e., how words modify the meanings of other words in textual data. This approach analyses the structural hierarchy and flow of a sentence in order to resolve the clauses and phrases. The scope of negation is handled in a novel way based upon the phrases and/or parts of phrases affected by negation. The opinion based scores are assigned to the clauses based upon the real numbers between -1 and +1. Later these clause level scores are aggregated based upon the topic analysis, opinion orientation, and opinion scores are calculated for sentences.

The main strength of this approach lies in the use of the sentence structure and hierarchy for opinion analysis. This approach gives a novel way to resolve the scope of negation, and a different way to calculate the orientation and scores for opinion based clauses.

The main limitations of the proposed approach for opinion analysis is that it mainly relies on existing resources, like Penn TreeBank, Stanford Parser, SentiWordNet. Penn TreeBank and Stanford Parser, which are well established resources and are the most widely used for POS tagging and for finding relationships between words. However, in the domain of lexical resources for opinion analysis there is still a need to develop and extend these resources. SentiWordNet 3.0. is one of most widely used word based resources for opinion analysis, but there are issues relating to opinion scores assigned to some of the words. Problems like WSD are yet to be sorted for the assignment of opinion scores to opinion based words.

There is a need to refine and develop heuristic rules utilizing the results of dependency grammars in order to resolve the hierarchical structure of sentences, and to refine phrase level dependency parsing. Refinement of the resources and closely related issues like noun co-referencing can significantly improve the performance of the proposed approach.

Another limitation of the opinion analysis approach is the assumption that the sentences are relevant and cleaned. There is a need for a preprocessing step in the proposed opinion analysis approach, which can help in cleaning stop words from the sentences and perform stemming for the words before matching them in a given lexical resource. Stop words are words, which are very often used in the written text and are important in order to understand the structure of sentence. However they don’t have any orientation and do not add much meaning to the opinion analysis of a sentence. Stemming is a process to remove the suffixes from the words and bring them to their basic form. For example, ‘helping’ is a word which might not be part of the lexical resource, however after stemming it is ‘help’ which is part of the lexical resource to be looked up for its semantic orientation. Similarly the preprocessing step may also contain a ‘spell check’. Such a preprocessing step is necessary for web based textual data as web based textual data can be written by people who have English their second language, and it may not follow formal language rules or structures.

All of the above mentioned issues and limitations can be improved through future work and can be explored further to bring about improvements in the opinion analysis approach.

3.6 Summary

The above chapter has presented the detailed description of a novel opinion analysis approach. The chapter began with an account of the basic understandings considered for the proposal of the approach. The next section presented the methodology followed during the analysis in order to develop the novel approach. This methodology mainly relied on an analysis of existing techniques, therefore the next section presented an in depth analysis of existing closely related research, which had been identified during the literature review in Chapter 2.

The in depth analysis is followed by the proposal of the novel opinion analysis approach. This details all the components of the approach. It details the reasons behind the selection of the level of granularity, and the parsing and scoring techniques used. In the final sections some other issues are discussed, which are not directly part of this research but, their improvement can continue to improve the proposed approach. The chapter concludes with identification of the limitations and strengths of the proposed approach for opinion analysis.

The limitations and issues explained in the last two sections make it clear that opinion analysis is not an independent and stand-alone task. Opinion analysis makes use of research from other domains and the outputs of the research in opinion analysis can help other research areas to improve. For example the opinion analysis in the proposed approach starts after a few pre-processing steps like opinion extraction from web data, noun co-referencing, etc. and the resulting tuples from opinion analysis can be saved in a corpus and reused later as a training and testing corpus for other opinion analysis research. All these steps are presented in the form of a framework in Chapter 5.

Chapter 4 – Background for Existing Resources and Frameworks

As established in Chapter 1, there is a need to bring research from different related fields together in order to find some solution for the problem of information overload encountered in relation to user generated opinion based textual data. Therefore a novel opinion analysis approach is proposed in Chapter 3, where techniques from NLP and computational linguistics are brought together. The current chapter presents and highlights some of the resources developed by a number of the researchers in the fields of NLP and computational linguistics which can help in the complete opinion analysis process. It further explicates and reviews the major lexical resources which can support in the development of the proposed opinion analysis approach. The current chapter establishes an argument for the development of a new unified opinion analysis framework based on the need to inter-link and bring together related fields within web text analysis

As mentioned by Thelwall et al. (2012) the research in opinion analysis and opinion identification in written text is divided into three categories based upon the techniques used: linguistic analysis, lexical resource based/machine learning, and polarity estimation from term co-referencing.

It is established in Chapter 3 that written text is not merely a set of words, but there are a formal set of rules in language and grammar which govern the position of words within a sentence, and establishes the relations between words. These relations between words help in the development and interpretation of meanings for sentences and the understanding of different parts of sentences (Subjects, Objects and Verbs). The proposed opinion analysis technique emphasises the analysis of phrases, and clauses in order to analyse opinion orientation at clause/sentence level. Therefore one of the preliminary steps for analysis of textual data is parsing. Parsing is discussed in Section 3.5.2 in detail. In the current section some of the established resources from the fields of NLP and computational linguistics for parsing textual data in the English language and establishing the relationships between words are overviewed.

4.1 Existing Resources (NLP) (Based on Linguistic Analysis)

Traditionally, the basic level of parsing for a sentence is to assign a lexico-grammatical annotation to a sentence such that each word is assigned with a tag representing a grammatical structure. The main problem in this parsing process arises with the fact that NLP researchers are unable to agree on one standard way to parse a sentence. This is not the only problem as they are also unable to agree on a single parsing scheme and tagset. Some examples of the tagsets are Brown Corpus tagset, Penn TreeBank tagset, and International Corpus of English (ICE) etc. (Greene and Rubin, 1971; Penn_Treebank, 1992; The_Ice_Project, n.d). The reason behind provision of these different syntactic annotation schemes is that each of these schemes provide different levels of refinement for grammatical classification. There has been research conducted in order to provide a standard tagset for annotation, by providing a mapping scheme between different syntactic annotation schemes (tagsets). Automatic Mapping Among Lexico-Grammatical Annotation Models (AMALGAM) (Leeds_University, n.d) is one example of such research. However, as mentioned earlier all different annotation schemes not only comprise different representations for different grammatical units, but also use different mechanisms to segment text; for example: compound words or idiomatic phrases are tagged differently in different annotation schemes, some of them are given a single tag, whereas, some others strip off the affixes and assign them a separate tag. For example: “n’t” used in don’t, can’t, etc., is used differently in different taggers, in BROWN tagger there is no tag to resolve “n’t”, whereas Penn TreeBank defines it as ‘RB’ (Adverb), and Lancaster-Oslo/Bergen Corpus Tag-set (LOB) and Polytechnic of Wales Corpus (POW) define it as ‘XNOT’. There is no single standard scheme agreed by NLP researchers and even AMALGAM doesn’t provide a standard, it only proposed a method to map different annotation schemes. Provision of one standard scheme appears to be unfair to some existent schemes as they are forced to compromise on their refined levels, especially for dependency grammars which have no grammatical classes.

This difference in tagging schemes gets increasingly complex with different sources of data and with the availability of different social media sites. Twitter provides different formats for communication from that of Facebook. The restriction in the number of characters to communicate adds difficulty to it. Therefore, Twitter uses different special characters in order to reduce the character set used for comments. For example, @, #, emoticons and URLs. Therefore the tagset used to annotate tweets will be different and will use special categories (Gimpel et al., 2011). However, the tagset for other social networking sites, for example, Facebook will be different as they do not have the same set of special characters and their data has no restriction on character limit.

Dependency grammar presents dependency-based representations of sentences in natural language parsing. There are two methods for dependency parsing: grammar-driven and data-driven. Dependency grammars are based upon the assumption that the syntactic structure of a sentence consists of lexical elements linked by binary asymmetrical relations called dependencies. This as compared to constituency based parsing lacks the phrasal structures. Therefore there are no phrasal nodes in dependency structures (trees) (Nivre, 2005). One of the recent and most utilised resources in dependency parsing is the Stanford Parser (Www.Stanford.Edu, n.d). There have been other types of representations of dependency relations such as the head-driven parsing models of Collins (Collins, 2003). In the Collin's parser decisions about dependency structures correspond to a head-centred, top-down derivation of the tree. Most parsers Collins Parser, Bikel's Parser (implementation of Collin's head driven statistical model) and Stanford Parser are claimed to be multilingual parsers but only accept training data in UPenn Treebank format (Chen et al., 2009a). They are different in terms of techniques, models and attributes.

The Stanford Parser is a probabilistic natural language parser. A natural language parser is a parser which works out the grammatical structure of sentences, for example, which groups of words go together in the form of 'phrases' like noun phrases, verb phrases etc.,. Whereas, probabilistic parsers use knowledge of the language gained from hand-parsed sentences, and produce their perception of the most likely analysis of a sentence. The original version of the Stanford Parser was mainly written by Dan Klein, with support code and linguistic grammar developed by Christopher Manning (Www.Stanford.Edu, n.d). Details about the Stanford Parser in specific to the current novel opinion analysis approach are discussed in Section 3.5.2.

4.2 Lexical Resources

As stated earlier the main approach for opinion analysis is the lexical based approach/machine learning approach. Both these approaches involve the annotation of textual resources by human annotators, either in the training of a machine learning algorithm or in the development of resources to be used during lexical analysis for identification of opinion bearing words. The words which are used to express positive opinions are known as positive words, whereas the words which are employed to express the negative opinions are known as negative words.

There are two ways of generating the lexicon for opinion analysis. One is manual and the other is automated/semi-automated.

Manual ways of generating lexical resources are very time consuming, slow and inconsistent (Morinaga et al., 2002; Greene, 2007; Devitt and Ahmad, 2008; Tadano et al., 2009; Lu et al., 2011). The accuracy of the lexical resource mainly relies on the understanding of the annotators of the language and terms used (Greene, 2007). If more than one annotator is involved the difference in understanding of the terms, language and the skills of the annotators brings inconsistencies and issues related to the precision of the annotation. Due to the slow nature of the process of manual annotation, the moods of the annotators may also bring in some inconsistencies.

All these issues support the perspective that manual annotation should not be used alone but in combination with automated approaches. Such semi-automated approaches generally begin with a manual selection of a seed set (set of words with known opinion orientation/strength), later this seedset is expanded by different methods like bootstrapping of the synonym and antonym structure of dictionaries (Liu, 2010). Similar to manual selection of seed set, the process of evaluation utilises a manual approach in order to check and rectify the mistakes made by an automated system (Yi et al., 2003).

4.2.1 Automated/Semi-Automated Approaches for Generation of Lexical Resources

As noted in Sections 2.8 and Section 2.9 lists are the most widely and commonly used lexical resources in the field of opinion analysis. The process of opinion analysis follows a simple approach. While parsing the textual data it parses the words and each word is checked in the provided positive and negative lists. When a word is found in a positive list it is marked as positive and vice versa. If the lists have scores attached to words the relative score is assigned to text. During the aggregation process if there are more positively marked words in the document, the document is classified as positive and vice versa. The process of finding words in lists is relatively simple and just employs standard search techniques; therefore list based resources gained popularity for opinion analysis. However there are some limitations identified in list based techniques. One of the limitations is that the same level of opinion (opinion score) is attached to the whole list, for example: all words in positive list are attached to score +1. This way of assigning polarity was a slow process as each word is to be searched and matched in lists. Therefore, the effective use of lists as a source has been substantial and many variations in these lists have been proposed. For example: the use of POS and the introduction of patterns with the help of more lists like, lists of adverbs of affirmation, adverbs of nouns, strongly intensifying etc. along with defined rules and patterns for aggregation and averaging are defined (Benamara et al., 2007; Subrahmanian and Reforgiato, 2008). Such lists are used in recent research, even sometimes using rules without identifying the syntactic information: POS. For example Li and Wu (2010) have used lists of positive and negative words along with five different lists of modifiers (with different emotional intensities) which if used along with some other opinion oriented words changes the strength of the word. This research was mainly based on the Chinese language (Li and Wu, 2010). One of the main limitations of lexical based opinion analysis is that if a word in textual data is not encountered within the lexical resource used for opinion analysis, then the word is marked as neutral. Therefore there is no polarity assigned to the word and it hasno effect on opinion analysis. Therefore there are two ways established for the development and extension of lexical resources: dictionary based; and corpus based.

4.2.1.1 Dictionary based Resources

Dictionaries are thought to be one of the most complete word based resources. Most words which one encounters in daily routine are considered to be found in dictionaries. The population of lists from dictionaries is achieved by identifying and using a number of seed words manually provided by researchers, and later the words are populated into the lists by using a reference resource (a dictionary). The relations of synonymy and antonymy are used for this purpose. For example a word ‘congratulate’ is manually placed into the positive verbs list. Synonyms of ‘congratulate’ are searched in any lexical resource (dictionary); for example, WordNet and all these synonyms are populated into positive lists based upon their relative POS (Hu and Liu, 2004; Kim and Hovy, 2004b). Similarly, antonyms of a seed word from a positive list can be found and populated into negative lists.

Kamps et al. (2004) have used synonymy relation in WordNet in another way, they have developed and used WordNet based distance measures for the semantic orientation of adjectives. WordNet glosses are another popular tool for the development of lexicon and many researchers have used them for extending and populating the resource (Esuli and Sebastiani, 2005, 2006a).

More recent work in the field of development of lexical resources based on WordNet is SentiWordNet (Esuli, 2008). SentiWordNet is an automatically generated lexical resource using a random walk algorithm. SentiWordNet is further analysed in Appendix A.

Sometimes the pre-processing techniques like spell checking, stemming etc. have to be implemented on the reference lists in order to get a maximum number of possible similar words to populate the list. An interesting technique is used by Neviarouskaya et al (2009); they have used manipulations of the morphological structure of known words. For example; by adding derivational prefixes or suffixes to the base words they have changed the meanings of words and tried to populate the maximum verity of words across each already available word. These derivational prefixes or suffixes can fall in many categories like: propagating; reversing; intensifying; and weakening. For example propagating: able, ful, ate; reversing: anti, de, dis; reversing: super, ultra, mega; weakening: semi, mini, let.

Using the word ‘love’, adding suffix ‘able’ makes it ‘loveable’, using the word ‘courage’ and adding prefix ‘dis’ makes it discourage, using the word ‘star’ and adding prefix ‘mega’ makes it ‘megastar’, and the word ‘book’ when added with the suffix ‘let’ gives ‘booklet’. All these words change their meaning and intensity just by adding an affix to them. This technique was used by Neviarouskaya et al (2009), in order to populate the lists automatically. This approach for population of list based resources using dictionary like resources differs from most widely used synonym/antonym approaches.

One of the main limitations encountered during the population of resources by using dictionary based resources is identifying the correct sense of words, WSD. Dictionaries like WordNet are based upon the analysis of written excerpts to find out the meanings associated with words. There can be more than one meaning of words based upon the context. WSD is a separate area of research where extensive work is being carried out, it has been mentioned in Section 3.5.7 in Chapter 3 and Section 8.2 in Chapter 8 as a limitation on the proposed opinion analysis approach in this Thesis.

4.2.1.2 Corpus Based Approaches

The methods of development and population of resources based on corpus mainly rely on heuristic rules, linguistic constraints, syntactic patterns and dependency parsers. They generally begin with a small seed list of opinion words.

A similar approach was used by (Hatzivassiloglou and Mckeown, 1997), for the population of adjective lists. They used constraints based on the connection/conjunction words (and, or, but etc.): for example; one of the constraints explores the use of the word ‘and’. The ‘and’ conjunction is used when conjoined adjectives provide the same opinion orientation in a sentence. For example; “The building was magnificent and glorious.” Here if ‘magnificent’ is a positive word, then ‘glorious’ should also be a positive word (Hatzivassiloglou and Mckeown, 1997). Liu (2010) named this idea ‘sentiment consistency’ and argues that in practice consistency is not possible (Liu, 2010). Hatzivassiloglou and Mckeown (1997) have used a log-linear model to determine the orientation of two conjoined adjectives and have used clustering to produce positive and negative lists.

Later Kanayama and Nasukawa (2006) have extended this approach in the Japanese language and introduced a concept of ‘context coherency’. They used the inter sentence sentiment coherency. Here, the idea is that when the same opinion is expressed in neighbouring sentences (consecutive sentences), the change in opinion is indicated by the use of terms, like; ‘but’, ‘however’, ‘whereas’ etc. (Kanayama and Nasukawa, 2006).

The requirement for a domain specific lexicon resource is heavily argued and supported in the literature. Qui et al (2009), introduced a bootstrapping approach for the population of a domain specific resource for product reviews. They used a propagation approach based on the relationship between an opinion word and topics of the opinion words (product features) by using dependency grammars. They considered the assumption that opinion words always associate with the topics by similar relations. They used the opinion word and the topic together to identify and extract new sentiment words and then have used these new words and features to find more words and features. As both words and features are used to extract new opinions, Qui et al (2009), named this a double propagation approach (Qiu et al., 2009).

Ding and Liu (2008) have taken the domain specific approach a step further, in their introduction of the concept of context. They presented that the same word can have a different orientation even staying in the same domain, if used within a different context: for example; in a camera review, the sentence: “The battery life is long.”: “It takes too long to focus.”. In the first sentence the word ‘long’ has a positive orientation whereas, in the second sentence the same word ‘long’ is used to show a negative aspect of the camera. Ding and Liu (2008) have presented that even in the same domain there is a need to use the product feature and opinion word together in order to identify context. They use the inter sentential and intra sentential rules with a consideration of context in order to develop the lexical resource (Qiu et al., 2009).

As the problems and issues encountered in the field of opinion mining get complex the resources develop accordingly.

Most of these resources were developed for some specific piece of research and are often not reused in any other research. Even where a researcher tries to use such an approach, the researcher has to reconstruct and redevelop it by understanding the literature. As additional understanding some of the widely used resources and a brief discussion about evolution of existing corpora are presented in Appendix A.

4.3 Existing Frameworks

As recognised in Chapter 2 Section 2.2 IR methods seek to reduce the overall search space for a user. IR selects the data that might be relevant according to a requirement and/or query. Therefore, IR methods help to make relevant information quickly available to users. There are multiple mechanisms through which analysed and processed information can be presented i.e., summarization, visualization, etc. The approach taken depends on the requirements of the system and needs of the users.

A large number of systems, models, and frameworks have been designed to facilitate textual opinion analysis. Many of these existing approaches are designed to perform a specific task, for a specific purpose within textual analysis. The number of frameworks and systems show that opinion analysis is a process which involves more than a singular task. Opinion analysis involves a number of different steps and tasks. Unification of this series of tasks can be performed in order to formulate an opinion analysis framework. As noted in Section 2.6, there are four tasks involved in the opinion analysis process (extraction, processing, analysis and presentation). Only a small number of existing researchers have sought to combine more than one of these tasks together (Yang and Liu, 2008; Torres-Moreno et al., 2009b) and swotti (http://www.swotti.com/). In addition, only a small number of researchers like Lloret et al. (2012) have provided an in-depth analysis of fully automated opinion analysis.

Yang and Liu (2008) have presented research which provided combined IR and summarization tasks into a system. They scored the summarized and retrieved data based upon queries and their structural context (Yang and Liu, 2008). However, their system needed several improvements in terms of search functionality and to the query based scoring technique which was applied (Yang and Liu, 2008). Torres-Moreno et al. (2009) presented an opinion mining and summarization based question answering system (CORTEX: COndensés et Résumés de TEXte (Text Condensation and Summarization)) (Torres-Moreno et al., 2009b). Their system is mainly based on and evaluated through text summarisation techniques. However, in the case of CORTEX there is a need to improve the calculation of the most appropriate Named Entity in the summarizing step, in order to improve the results (Torres-Moreno et al., 2009b). Swotti (http://swotti.starmedia.com/) is a system which retrieves and analyses opinion based web documents, found in specific topic areas (e.g. mobile phone based reviews). In this way swotti provides a limited search engine based upon a collection of product review data sources. However, swotti is not a result of academic research, and thus is not properly documented or evaluated. Therefore, it brings the need for an improved IR system which can collect, process and index opinion based textual data from web search engines making this available to users.

Combining one or more of the tasks involved in the opinion analysis process can improve the performance and capability of the overall system (Stoyanov and Cardie, 2006; Lloret et al., 2012). However, it is observed that approaches which employ more than one task can sometimes amplify the issues presented, and can be considered to be quite an ambitious aim (Lloret et al., 2012). For example, in the combination of an opinion analysis function and a presentation function (summarisation) the overall performance of the system is impacted by the performance of each individual function. If the opinion analysis function only provides a performance of x and the text summarisation function only provides a performance of y. Then the overall performance of the system is f(x, y) which may be a reduced performance on how these systems could perform individually dependent on the input data quality.

In the research based on opinion analysis there are many frameworks presented based upon one or more of the tasks identified (Consoli et al., 2008; Jin and Ho, 2009; Lloret et al., 2012).

4.3.1 Jin and Ho (2009)

Based upon the availability of the high number of product reviews for individual products available on the Web, coupled with the natural limitation to read all these reviews, it complicates the decision making process. Jin and Ho (2009), presented a lexicalized framework (shown in Figure 4-1), which aimed to extract product reviews for product categories, where an opinion is specified in relation to highly specific product related entities (the entities are defined in Table 4-1). The opinion analysis technique presented in Jin and Ho (2009) _ framework, classifies the positive/negative opinion for each recognized product entity. This opinion analysis technique is based upon linguistic features (POS tagging), for automated machine learning. Their lexicalized Hidden Markov Model (HMM) uses POS tagging and Named Entity Recognition (NER). POS tagging means the annotation of words encountered while parsing the dataset, with their corresponding POS i.e., noun, verb etc. NER is a process of identification and classification of a named person, location or organization.

Jin and Ho (2009) have adapted their technique from the work related to Korean POS tagging and Chinese Named Entity Relationships (NER) from Lee et al. (2000) and Fu and Luke (2005) respectively. An example for the definition of entity categories for a camera is shown in Table 4-1. This framework mainly uses manual annotation for the creation of a training corpus. The basic annotations for opinion classification are explicit and implicit opinion entities for positive and negative opinions. This framework mainly depends on the recognition of patterns and sequences in assigned tags (in the format of a pair (word, POS (word))). The identification of patterns and using them for machine learning reduces some natural language complexities, and the framework allows the system to self-learn new vocabularies. Manual annotation made it possible to recognize complex and infrequent entities. Jin and Ho's framework mainly emphasised the identification and retrieval of product entities and the opinions expressed within these. However, the main limitation of this framework is the need for a large dataset to establish refined training rules. The existing dataset is manually annotated, and manual annotation of large dataset with multiple products is a substantial challenge (as noted in Section 6.4.2). There is also a need for improvements to pronoun resolution for NER. These limitations strengthen the requirement for further research in the area of automation of opinion analysis, and highlight the need for an IR tool which can provide a way for the automatic generation of a repository (corpus), which can collect data from the Web and can be used in the context of an opinion based search engine.

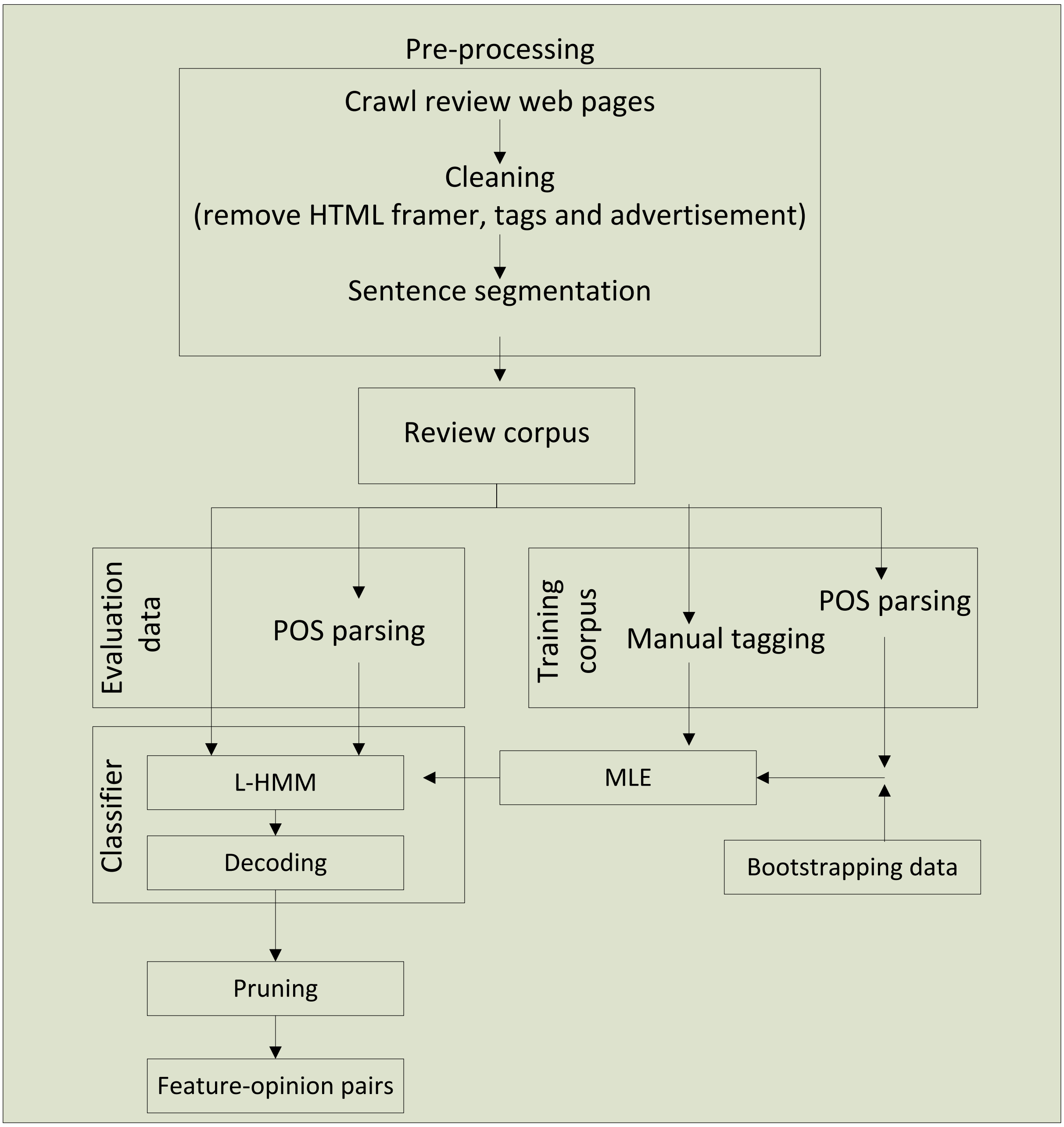


Figure 4-1: The System Framework by Jin and Ho (2009)

Table 4-1: Definitions of entity categories and examples, Jin and Ho (2009)

COMPONENTS	Physical objects of a camera including the camera itself, e.g., LCD, viewfinder, battery
FUNCTIONS	Capabilities provided by a camera, e.g., movie playback, zoom, automatic fill flash, auto focus
FEATURES	Properties of components or functions, e.g., colour, speed, size, weight, clarity
OPINIONS	Ideas and thoughts expressed by reviewers on product features / components / functions.

4.3.2 Lloret et al. (2012)

More recently, Lloret et al (2012), presented a framework in order to develop an integrated solution for IR, opinion mining and text summarization, as shown in Figure 4-2.

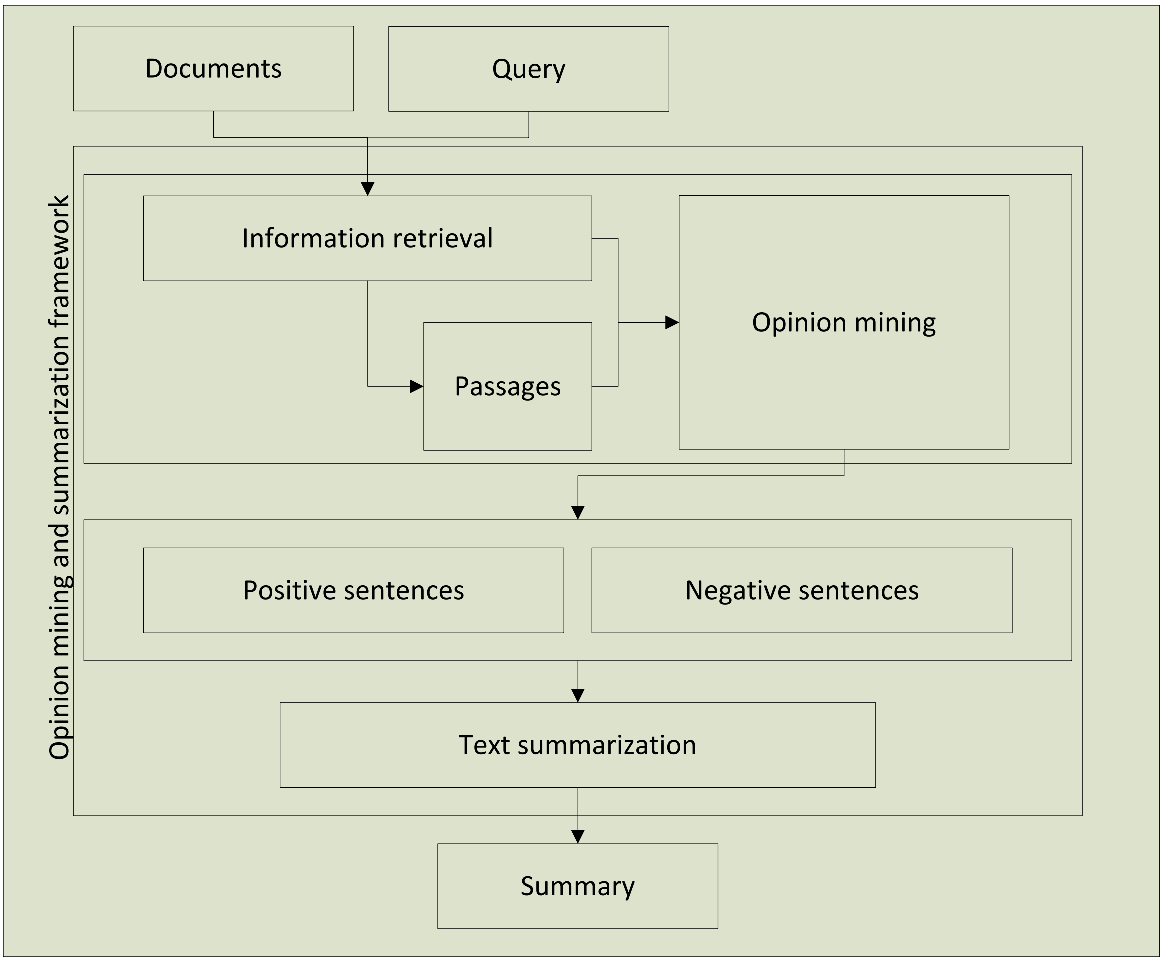


Figure 4-2: The proposed unified framework by Lloret et al. (2012)

They presented their framework in three distinctive stages and provided a detailed explication for each of these stages. Their first stage was an IR system. IR was based on the identification of fragments of text which answered questions provided by a user rather than returning whole documents. This fragment based approach was selected on a basis of its performance in international question answering competitions (Gómez-Soriano et al., 2006; Christensen and Ortiz-Arroyo, 2007; Buscaldi et al., 2010). Their approach mainly relies on the identification of the structure of a question and discovery of similar expressions in their document set. The similarity between the question structure and the expected answer establishes the relevance. They assigned a weightage for relevance based upon a distance based formula. WordNet Affect (Strapparava and Valitutti, 2004); SentiWordNet (Esuli and Sebastiani, 2006b); MicroWNOp (Cerini et al., 2007); and the JRC (Joint Research Centre) lists (Balahur et al., 2009). They mapped these four selected resources onto four classes: positive (1), negative (-1), high positive (4) and high negative (-4). This stage deals with the selection of opinion based sentences for opinion analysis. This level retrieves the thirty most relevant documents about a query using the Yahoo! Search engine. They employed Latent Semantic Analysis (LSA) through Infomap NLP Software, in order to identify topic related words. Sentences containing topic related words were scored for their opinion using the approach described in the first stage. LSA embodies words in a large corpus to define the similarity of the meaning of words by using statistical and mathematical calculations. It is important to understand the resulting similarity estimates in LSA are not contiguity frequencies, co-occurrence counts or correlations, it can deduce much deeper relations for meaning based judgments (Landauer et. Al., 1998).

The second stage in the framework is for opinion mining. Lloret et al (2012) used two different opinion mining levels. Their first level was employed from Balahur et al. (2009), it deals with opinion analysis. In this level Llorert et al. (2012), have developed their lexical resource based on four existing resources:

The next stage of the framework provides text summarization from the COMPENDIUM resource (Lloret et al., 2011). COMPENDIUM is selected as it was extensively evaluated and found to be acceptable over a wide range of domains.

Lloret et al. mainly evaluated the unification of tasks within their framework, rather than evaluating the respective opinion retrieval, analysis or summarisation techniques. They identified and brought forward the difficulties in bringing three application areas (opinion retrieval, analysis or summarisation) together with the aim of a coherent text fragment as an output. Their research employed all the respective retrieval, analysis and summarisation techniques from already established research. Therefore there was no requirement for these to be evaluated. However the dataset and data source (blogs), they used contained a lot of irrelevant and noisy information. Therefore the results for the identification of topic related information through their framework were not very high; only 30%. They identified that there may be better approaches than what they have employed in terms of each of the individual components (Opinion Retrieval, Mining and Summarization) and highlight in their conclusions that there is a need to improve individual components of the framework, and to extend the dataset for further evaluation.

4.3.3 Consoli et al. (2008)

Consoli et al. (2008) presented a conceptual framework for opinion mining on the Web. Their motive for opinion mining was to capitalise on customer opinion. They wanted to improve the products and reinforce customer loyalty by continually understanding customer opinion and behave accordingly. For this purpose they employed a Customer enterprise Customer (CeC) model, as illustrated in Figure 4-3.

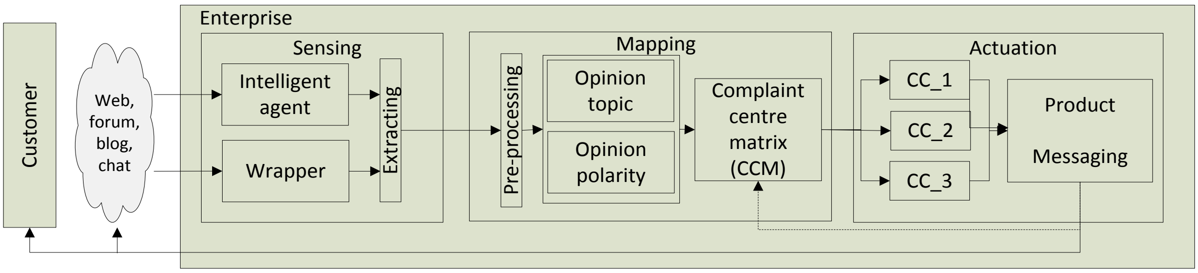


Figure 4-3: CeC Model presented by Consoli et al. (2008)

The CeC model is divided into three phases: sensing, mapping and actuation.

The objective of the sensing phase is to find and gather opinions about a product. However, most websites are structured and designed differently, there is a need to select the websites to be used to gather opinion data (Amazon, eBay, etc.), and to define wrappers in order to extract opinion from those websites. These wrappers are necessary, as the tools and agents (spiders) which extract opinion from the websites might need to go into the hierarchical structure of products, their ingredients, and components etc.

The Mapping phase includes opinion mining on a basis of the opinion and an opinion's topic. Once an opinion has been analysed based upon its topic, negative opinions (complaints) are extracted and routed to their respective departments to enable a response to be formulated.

The final phase of Actuation is basically the response phase. In this phase the complaints are responded to either by making improvements to the products, or by dealing directly with the customers.

Most work in opinion analysis is based upon adjective based opinion lexicons which specify positive and negative terms, or assign an opinion score to terms based upon a determination of positivity or negativity. However, Consoli et al. (2008) used an original algorithm based on emotional/affective values and in particular Ekman indexes (Grefenstette et al., 2004), in order to evaluate customer opinions about product. They used happiness, sadness, anger, fear, disgust, and surprise as the six most basic emotions.

None of the already existing and available lexical resources give opinion values based upon the spectrum of six emotions. Therefore, the reliability of Consoli et al.'s method mainly relies on the accuracy and consistency of the resources. Consoli et al. (2008), also brought forward the need for an IE and IR system similar to that of web crawlers. They highlight issues regarding the difference in the structure and design of websites, which can affect the overall performance of such an extraction system.

4.4 Justification for a New Framework Development

The examples above demonstrating a combination of more than one opinion analysis task (extract, process, analyse and present), and the discussion about opinion analysis frameworks, demonstrates there can be different motives to framework construction. However, the analysis of opinion based textual data always requires input text to be extracted from some source, and the analysed information needs to be presented, in any one of the following forms: summarisation, visualisation, corpus, lexical resource, machine learning rules, etc.

The discussed systems and frameworks specifically underline the need for an IE and IR component to be part of any effective opinion analysis system. The discussion also emphasised the need for a crawler system which can identify and extract opinion based textual data over the Web, analyse opinion from the extracted data, and present the analysed data in an easily retrieved form, for later use, in order to complete the cycle. The discussion also stresses the need for further research into the automation of opinion analysis.

Therefore, in order to unify the series of tasks and present the proposed opinion analysis approach (Chapter 3) as part of a process, a unified framework for opinion analysis is proposed in Chapter 5. This framework will extract textual data from the Web, identify, analyse, and save opinion related data. In addition it will present the analysed information into an automatically generated corpus for reuse. This corpus can be further used in the efficient retrieval of analysed opinion based data for search engines. In addition, this corpus can also be used as a training corpus for an extension of a lexical resource, into general or domain specific resources.

The main aim of this framework is to present a way to reduce manual processing for the extraction and interpretation of opinion from web based product review data. Therefore presenting the detailed opinion analysis and different parts of the opinion (Opinion-Opinion Topic) into a reusable and extendable corpus is determined to be a Thesis contribution.

As highlighted in this chapter the main limitation of corpora based resources in opinion analysis are: the size and diversity of corpus. This can be managed in the proposed corpus, as it will be able to be extended through the addition of extra materials, and its architecture will be extensible to deal with future challenges. Most corpus are manually generated therefore issues like inconsistencies and slow generation are observed, whereas the proposed corpus will be automatically generated. Finally existing corpora have problems regarding reusability as annotation schemes are not standard, and corpora are unstructured due to the unstructured nature of textual data. The proposed corpus provides a more structured approach with annotation occurring through the separation of files, creating more opportunities for re-usability.

On the whole the flexibility and scalability of the state-of-the-art corpora are their main limitations. Flexibility in terms of their reusability and scalability in terms of their append ability. As most of them are manually annotated, this manual annotation makes whole process slow and inconsistent.

Chapter 5 – Framework

Chapter 4 highlights a few of the available frameworks in the area of opinion analysis, which serve the requirement for combining a series of related tasks into frameworks. This study of the available frameworks, which were most related to the current study emphasised their limitations. These limitations provide a justification for the need for a new framework as a further contribution of the Thesis. The current chapter begins with the presentation of a framework. The output of the process presented in the framework is a corpus. In addition the chapter presents the design and novelty of the corpus.

5.1 Proposed Framework

As established in Chapter 4, the main purpose for this framework is unification of a series of tasks surrounding opinion analysis and providing an encapsulated process. The core of this framework is the novel opinion analysis approach presented in Chapter 3. The process of opinion analysis completes in three main steps, opinion extraction, opinion analysis and presentation of the analysed information. The opinion analysis approach presented in Chapter 3 is employed into the framework in combination with an information retrieval stage for providing input to the opinion analysis stage, before final stage presentation of the analysed information (output). The proposed framework as presented in Figure 5-1, has three main stages, IR, opinion analysis and corpus generation. It is a unified framework which encapsulates all three stages into a single process. The output of each of the stages is input to the next stage, i.e., output of information retriever is input to opinion analyser. The IR stage retrieves the subjective information from web documents and passes this onto the opinion analyser. The opinion analyser based upon the approach proposed analyses the retrieved information and saves the analysed information into a corpus. The design of the corpus is explained further in Section 5.1.4.

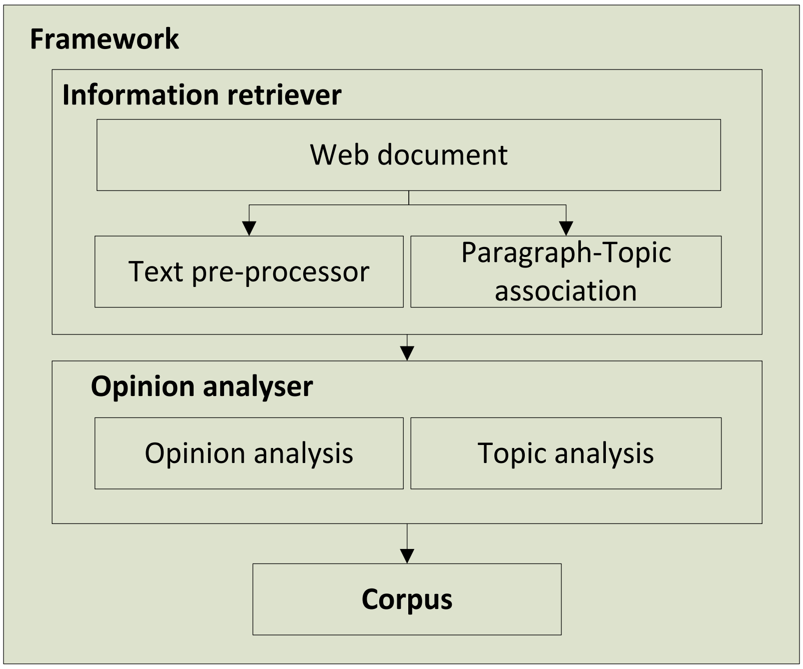


Figure 5-1: Block Diagram for Framework

5.1.1 Information Retrieval (IR) Stage

IR deals with the retrieval of information based upon the query. As discussed in Chapter 2, there is a tremendous amount of information on the Web and the sources of information (websites) are growing. Therefore there is a need for an automated system to handle the diverse types of data (images, text, audio, video, etc.) available online. However, the majority of current research focuses on textual data. The IR stage is presented into a block diagram as information retriever block in Figure 5-1.

The majority of subjective text resides as user generated text on blogs, chat rooms and social networking sites and forums. Such forums and blogs have their own specific structures, for example: http://www.medhelp.org/forums/list is a medical help forum where communities are defined based upon standard lists on forum. On choosing one of the topics in the list i.e., ‘eye care’, a list of sub-topics is displayed, these sub-topics are active threads which members of the community have created and are responded too by members. This gives each post in the thread two topics. One is the community name and the second is the topic assigned to the thread. Similarly each forum, blog, or social networking site can have its own structure, therefore when retrieving information from a website there is a need to write a wrapper which understands the structure of the information within the site.

The retrieval of textual data from the Web is not an easy task, as the structure and design of websites vary drastically. Therefore there is a need to develop ways of extracting and capturing the textual data from all required websites. In addition there is a lot of other materials like HTML Tags, scripts, advertisements and external links on each website, thus there is a need to make sure that only the user generated textual data from the website is retrieved, cleaning away all the external links and advertisements.

The output of this retrieval stage is provided to the opinion analyser stage, with this comprising of the opinion analyser use the opinion analysis approach presented in Chapter 3. One of the limitations of the opinion analysis approach is a lack of text pre-processing. Therefore the input to the opinion analyser should be cleaned and pre-processed. A pre-processing step is defined in the retrieval stage. Detailed pre-processing takes a pipeline like approach and is presented in Figure 5-2.

5.1.2 Pre-Processing

The retrieved data is user generated unstructured text, which holds a lot of diversity and is huge in size. The initial level of pre-processing starts in the IR stage when initial noise and uninformative elements of text such as HTML Tags, scripts, advertisements and external links are removed and user generated textual information is retrieved. One of the main limitations attached to user generated textual data is that it is not formal structured text. There can be many challenges attached to the analysis of user generated textual data, some of the challenges can be, spelling mistakes, typing errors, the use of multiple languages, and extensive use of pronouns etc. These are classed as challenges as the presence of spelling mistakes, typing errors, or use of multiple languages my lead to Out of Vocabulary (OOV) words, i.e., words which are not existent in the lexical resources or dictionaries used. These OOV words may lead to an unknown polarity or score and may mislead the overall opinion scoring and analysis. The extensive use of pronouns can also lead to misleading opinion holders and/or opinion topics. Therefore, textual data needs to be cleaned and filtered in the pre-processing stage. This pre-processing takes a pipeline like approach, where the retrieved text passes through a number of steps (stemming, stop word removal, and etc.) in order to make textual data ready for opinion analysis.

During this stage the paragraphs are parsed (tokenised) and a spell check is performed. The tokenisation separates all the words in the sentence. As the opinion analysis approach proposed in Chapter 2 and 3 is mainly used for user generated data, and utilises a dictionary like lexical resource. The opinion analysis approach relies on the availability of words in the lexical resources for opinion analysis. Therefore there is a need to perform a spell check in order to remove some of the spelling mistakes or typographical errors to improve the analysis. As the misspelled words can lead to inaccurate analysis i.e., if the words are misspelled they are not matched in the lexical resource, and therefore a neutral opinion is assigned to such words. This neutral opinion can result in misleading and an inaccurate representation of opinion.

In addition, the tokenised words are checked for stop words. Stop words are the words often used in textual data, but which do not add much semantic information to the written text. Words like ‘the’, ‘is’, ‘of’ etc. are classed as stop words; there is no standard list of stop words. Removing words classified as stop words may increase the efficiency of analysis as these words are not there to be looked up into resources for their meanings. Only a basic list of words with articles, ‘a’, ‘an’ and ‘the’ is used for stop words as all other words which are classed as stop words i.e., ‘is’ and ‘of’ are important for dependency parsing, and are screened during analysis of dependency structures.

The next step in text pre-processing is the identification of noun co-referencing. Noun co-referencing is detailed as one of the main limitations in the proposed opinion analysis approach in Chapter 3 and is detailed in Chapter 3 (other issues). It helps in the identification of a correct topic and correct opinion holder. In the last step, the pre-processor performs stemming on words. Stemming is the process of reducing the derived words to their stem or root word. In many cases the lexical resources do not contain the derived words in their repository. For example ‘argue’ is a word, which can have forms like ‘argued’, ‘argues’, and ‘arguing’. All these words mean the same which comes from their base/stem/root word: ‘argue’. However, the lexical resource might not realise this and return ‘argued’, ‘argues’, and ‘arguing’ as unidentified words being absent in the resource. Therefore stemming can help in discovering the words in the lexical resource.

The pre-processed text is then provided to the opinion analyser in order to perform opinion analysis.

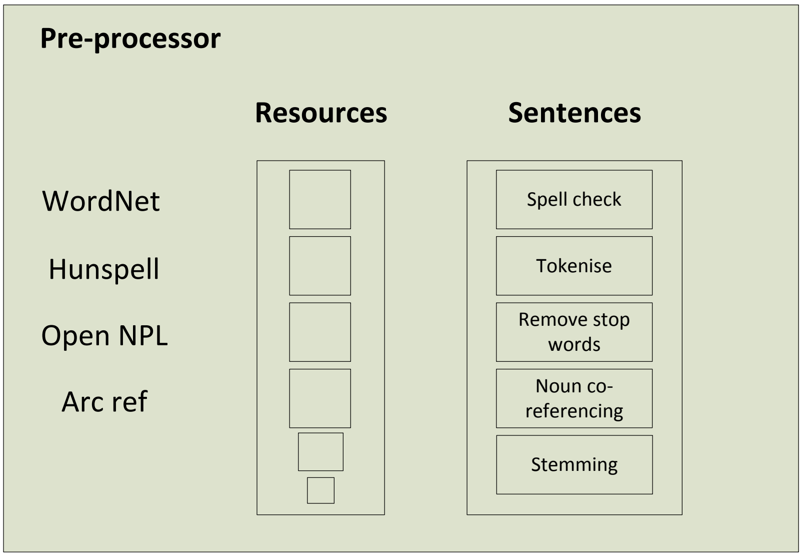


Figure 5-2: Pre-processing pipeline

5.1.3 Opinion Analysis Stage

The opinion analyser is based upon the novel opinion analysis approach presented and detailed in Chapter 3. The approach mainly relies on lexical and linguistic analysis of opinion in sentences. The output of opinion analysis mainly comprises of the opinion orientation and strength, Opinion Words in the sentences and Opinion Topics. The Opinion Topics when analysed with the sentence topics assigned in the IR stage provides insight into whether the opinion expressed is directly related to the topic of the post or not. This gives a guideline if the opinion gives any input to the overall opinion of the post, if it is drawing a comparison, or if it is tackling an irrelevant topic.

5.1.4 Corpus Generation

A corpus is a collection or a body of text, which can be used for linguistic analysis. There is a tremendous growth observed in interest and activity for corpus building and analysis (Atkins et al., 1992). There are many corpora available in English language and they are discussed in Appendix A. It is observed that the recent definition of the word corpus incorporates a feature of computer processing ability (Atkins et al., 1992; Bhattacharyya et al., 2009a; Bhattacharyya et al., 2009b). However, it is interesting to note that still most of the corpora designed and defined in the fields of linguistics and opinion analysis rely on the manual collection, and annotation of text (Wiebe et al., 2005, Kessler et al., 2010, FrameNet, n.d.).

Therefore there is a need to automatically collect and annotate textual information especially subjective textual information. Atkins et al. (1992) have presented criteria for designing and planning of corpus. Atkins et al. (1992) have considered generation of corpus based on both written and spoken sources i.e., books novels, conversation, lectures etc. Based upon their criteria and the system requirements of the proposed opinion analysis approach and framework the following design criteria are considered for corpus generation.

The proposed corpus is proposed for English Language textual data captured from web based user generated content (social networking sites, blogs, forums etc.). Based upon the opinion analysis the basic unit of the corpus is a word. The corpus saves words at a very basic level; however, all references of each word from textual data are saved so that the words can regenerate the phrases, clauses or even sentences, whenever required.

Mark-up and coding of the captured textual data (during information retrieval stage) is performed automatically during the opinion analysis stage, and thus does not consume a large amount of time. This corpus can help in the development of a web based search engine, where this corpus can be extended as a main repository for semantically intelligent opinion discovery.

The opinion analysis approach used in the framework is fully automated and annotates the text based upon the opinion analysed (opinion polarity and strength). In addition it identifies the opinion oriented words/phrase in the sentence and their respective opinion topic. Information is the information most frequently manually annotated in most used opinion based corpora MPQA (Multi-Perspective Question Answering) and JDPA (The J.D. Power and Associates Corpus).

The resulting corpus consists of a user generated web based content analysed for its opinion related information. The text is analysed and annotated at sentence/clause level for opinion and Opinion Topics. The opinion is aggregated and annotated on the basis of topics. POS information is also annotated for both phrase and word level.

The automated generation of a corpus through automatic annotation is performed with the help of an opinion analyser and its extendable nature makes this corpus unique. As Fillmore and Charles (1992) clearly made observation that there cannot be a single corpus which is large enough to contain information regarding all lexicon and grammar of any language. Therefore the append able and extendable feature in any corpus is very necessary in order to capture as much information as possible.

The main constituents of the corpus are topics, sentences, clauses and their respective phrases. The design of corpus is presented in Figure 5-3. The details of annotation are presented below:

Topic:

t_id - A unique id is attached to each topic. The assignment of id helps to reduce redundancy, as text already in the list of topics will not be repeated.

text - The textual representation of the topic.

Sentence:

s_id - A unique id is attached to each sentence.

text- The textual representation of the sentence.

score - Opinion orientation and polarity are saved as a score valued between [+1, -1], where + means positive and - means negative polarity, the value of number presents the strength of opinion, mapped on to (Strong, Mild and Weak). This gives Strongly (+/-), Mildly (+/-), and Weakly (+/-) as values of score.

t_id - the id(s)/list of topic corresponding to the sentence.

Clause:

c_id - A unique id is attached to each clause.

text - The textual representation of the clause.

Sentence - Clause

sc_id - A unique id is attached.

s_id - A unique id is attached to parent sentence.

c_id - A unique id is attached to each child clause.

Score **-** Opinion orientation and polarity are saved as a score valued between [+1, -1], where + means positive and - means negative polarity, the value of number presents the strength of opinion, mapped on to (Strong, Mild and Weak). This gives Strongly (+/-), Mildly (+/-), and Weakly (+/-) as values of score.

POS- Parts of Speech corresponding to each token: word/phrase in the clause.

Phrases:

There can be more than one type of phrases in a clause. The most basic phrases are Noun Phrases and Verb Phrases as shown in Chapter 3. Noun Phrases and Verb Phrases are generally constituent of (Nouns & Adjectives) and (Verbs & Adverbs) respectively. However there can be adverbial Phrases and Adjectival Phrases, which will be handled similarly.

Verb Phrase:

vp_id - A unique id is attached to each Verb Phrase.

text- The textual representation of the Verb Phrase.

Clause - Verb Phrase:

cvp_id- A unique id is attached.

vp_id - the unique id attached to each child Verb Phrase.

c_id - A unique id is attached to parent clause.

Score - Opinion orientation and polarity are saved as a score valued between [+1, -1], where + means positive and - means negative polarity, the value of number presents the strength of opinion, mapped on to (Strong, Mild and Weak). This gives Strongly (+/-), Mildly (+/-), and Weakly (+/-) as values of score.

Opinion (Y/N) - Generally opinion based words are found in verb phrases based upon adverbs therefore a Boolean value for the identification of opinion bearing phrases is saved.

O_id- A unique id is attached to each opinion bearing phrases if it is an opinion.

Topic_id - id of the corresponding topic of opinion (identified in Noun Phrases) is attached in order to create a link between opinion bearing words and the Opinion Topics.

Noun Phrase:

np_id - A unique id is attached to each Noun Phrase.

text - The textual representation of the Noun Phrase.

Noun:

N_id - A unique id is attached to each Noun.

Text - The textual representation of the Noun.

Score - Opinion orientation and polarity are saved as a score valued between [+1, -1], where + means positive and - means negative polarity, the value of number presents the strength of opinion, mapped on to (Strong, Mild and Weak). This gives Strongly (+/-), Mildly (+/-), and Weakly (+/-) as values of score.

Noun Phrase - Noun:

N_id - A unique id is attached to each Noun.

NP_id - A unique id is attached to each Noun Phrase.

Clause - Noun Phrase:

cnp_id - A unique id is attached.

np_id - the unique id attached to each child Noun Phrase.

c_id - A unique id is attached to parent clause.

score - Opinion orientation and polarity are saved as a score valued between [+1, -1], where + means positive and - means negative polarity, the value of number presents the strength of opinion, mapped on to (Strong, Mild and Weak). This gives Strongly (+/-), Mildly (+/-), and Weakly (+/-) as values of score.

Topic (Y/N) - Generally corresponding topics of the opinion based words are found in Noun Phrases therefore a Boolean value for the identification of topics associated with opinion bearing phrases is saved.

topic_id - A unique id is attached to each opinion topic if it is a topic.

o_id- id of the corresponding opinion bearing words the Opinion Topics and their corresponding words.

Adjective:

a_id - A unique id is attached to each Adjective.

word- The textual representation of the Adjective.

Noun Phrase - Adjective:

npa_id - A unique id is attached.

a_id - A unique id is attached to each Adjective.

np_id - A unique id is attached to each Noun Phrase.

Score - Opinion orientation and polarity are saved as a score valued between [+1, -1], where + means positive and - means negative polarity, the value of number presents the strength of opinion, mapped on to (Strong, Mild and Weak). This gives Strongly (+/-), Mildly (+/-), and Weakly (+/-) as values of score.

5.2 Summary

The proposed framework use similar attributes as the majority of other opinion based corpora (Kessler et al., 2010). However most other corpora available are manually generated and are mostly used only as training corpora to train machine learning systems (Wiebe et al., 2005; Wilson et al., 2005; Kessler et al., 2010). Most of the existing corpora are generally subject/topic based and as they are manually generated they become difficult to train machines for datasets from any other domain than that of the domain of the training data. Whereas the current corpus is not domain specific, it can be used across multiple domains. It has topic as a class which can capture any domain. The automation of the opinion analysis approach, automated generation of the resulting corpus, and its structure to be append able are major strengths of the corpus.

The framework proposed follows the general flow of an online opinion analysis system, to capture, process, present and save information for later usage. An application of the framework could see it used in the context of an intelligent spider for online subjective information. Capturing, analysing, presenting and storing such information for use by search engines and subjective repositories. Such automated systems can also help for the automated creation of corpus for machine training purposes. The usage of automated corpus for machine training can help in prevention of human errors and bias spawned from manual generation of training corpus.

The next chapters of the Thesis present a plan and process for evaluation of the proposed opinion analysis, framework and to determine the effectiveness of the resulting corpus. On the basis of the evaluation plan the evaluation process is achieved by implementation of a prototype system based upon the proposed opinion analysis approach.

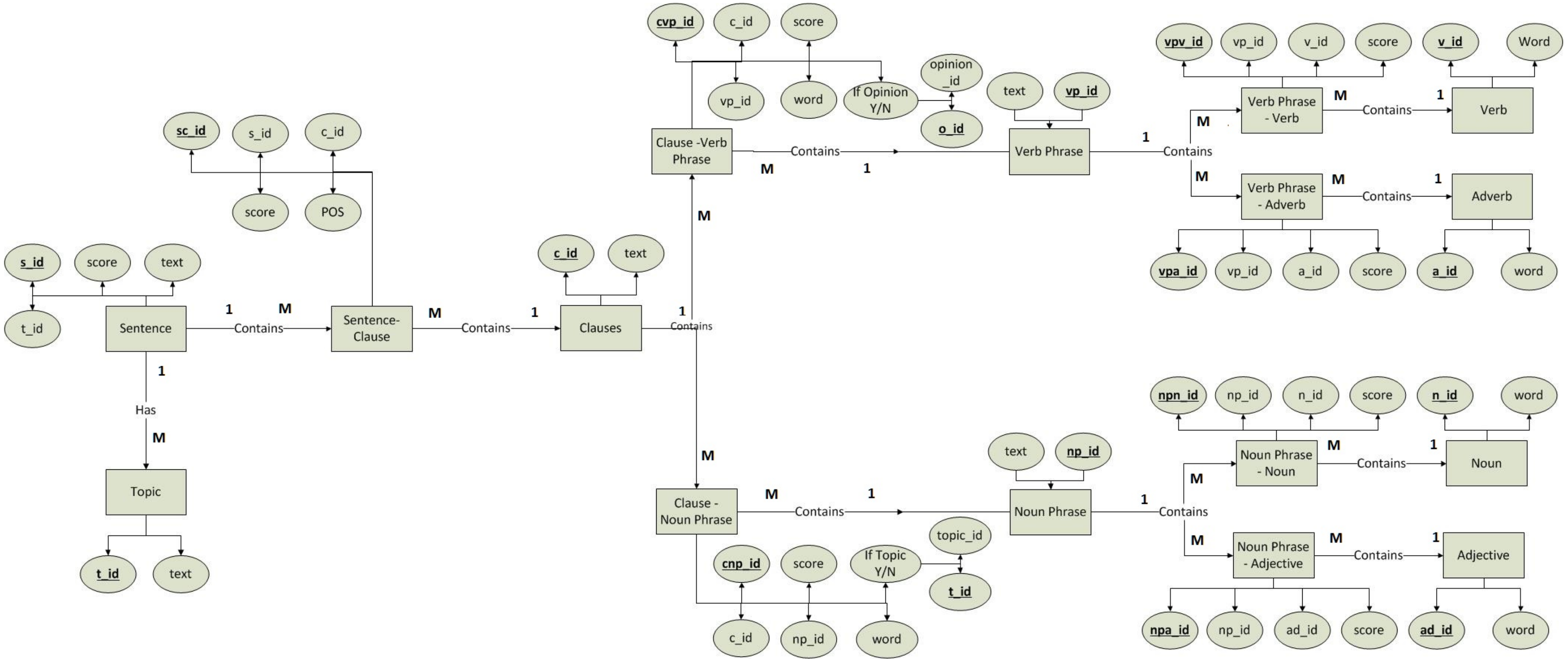


Figure 5-3: Corpus Design

Chapter 6 – Evaluation Plan

The main aim of the current research is to provide a novel opinion analysis approach which contributes to the existing state-of-the-art in improving automatic semantic intelligent opinion analysis at a clause level. This opinion analysis approach is employed into a framework, which is designed to capture user generated data from the Web, to perform opinion analysis on the textual data captured, and to generate an easy to retrieve opinion based corpus. This corpus is a resource which captures and stores information regarding the opinion, opinion topic, opinion orientation and opinion polarity scores at a clause level.

There is a need to evaluate the reliability and verify the precision (linked to expert and non-expert human annotation) of the opinion analysis approach. There is also a requirement to assess the improvements in opinion analysis resulting from the identification of phrases in sentences and their usage in the process of opinion analysis. This chapter proposes a plan for the evaluation of the presented opinion analysis approach (see Chapter 3) and a plan to assess the utility of the corpus generated through a focus on its applicability and reusability.

This chapter begins in Section 6.1 with a review of methods used during the evaluation of existing state-of-the-art opinion analysis resources and approaches. The next section in this chapter provides a critique of existing evaluation techniques used in relation to opinion analysis. This is followed by presentation of established goals for the evaluation of the proposed opinion analysis approach. These goals are articulated into a detailed evaluation plan. Finally, the limitations of the evaluation plan and techniques used conclude the chapter.

6.1 Evaluation Process Followed in the State-of-the-Art

6.1.1 SentiFul

SentiFul (Neviarouskaya et al., 2011), a recently developed sentiment lexicon, is a word based sentiment lexicon which is generated and scored automatically. The use of direct synonymy, antonymy, and hyponymy relations (from the WordNet resource) form fundamental lexical units at the core of the SentiFul resource. SentiFul extended the word set available in the existing resource Affect database (Neviarouskaya et al., 2007), by scoring words based on sentiment scored lemmas (words), and different types of affixes. An innovative algorithm was used in order to discover new lexical units and to score these using defined rules and patterns based on derivation and the compounding of lemmas. Derivation is the process of creating combinations of lemmas and affixes on a basis of different features (e.g. propagating, revising, intensifying, and weakening). For example, affixes that have intensifying features on a word e.g. ‘super’ + ‘hero’ gives ‘superhero’ and, ‘over’ + ‘awe’ gives ‘overawe’. Compounds are words that contain at least two roots. Therefore the algorithm used combinations of two independent existing words to form new words or phrases. For example, ‘risk’ + ‘free’ gives ‘risk-free’.

In order to evaluate the accuracy of the methods used to generate this lexicon, a two-step evaluation procedure was adopted. The first step in the evaluation process of SentiFul involved the design of two sets of Gold Standards (GSs). These GSs were used to compare and evaluate the accuracy for the scoring of SentiFul. In this process, authors randomly collected 1000 terms from SentiFul and two non-native English speakers were chosen in order to annotate those randomly selected 1000 terms. The purpose of this annotation was to assign a polarity (positive, negative or neutral) and a numeric polarity score to each word. The first GS (GS-1) was defined where both the annotators agreed. The manual annotation by two human annotators assigned polarity over the range of positive, negative and neutral. However, SentiFul lexicon’s architecture did not contain mechanisms to represent neutrality i.e., to distinguish between neutral and sentiment conveying words. Therefore, another GS (GS-2) was constructed excluding all the words with neutral labels in GS-1.

Pearson’s Correlation (Chung, 2007) was used in order to evaluate the relatedness of polarity scores assigned by the system; with the GS scores assigned by human annotators, and analysis were undertaken.

The second step in the evaluation of SentiFul involved analysing the comparison of the sentiment scores of the SentiFul lexicon with the already established lexical resource: General Inquirer (GI) (Harvard.Edu, n.d). Neviarouskaya et al (2011), collected 4002 terms from the GI (1813 positive and 2189 negative) in order to generate a GI based GS. They calculated Cohen’s Kappa Coefficient (Cohen, 1960) between SentiFul and

GI based GS in order to find the similarity between the SentiFul opinion score annotation and GI opinion annotation. They measured precision, recall and F-score for positive and negative words separately. Precision, recall and F-score measures are further discussed in Section 6.2.3 and Section 6.1.3.

6.1.2 SentiWordNet

SentiWordNet (Esuli and Sebastiani, 2006) is one of the most popular lexical resources for opinion analysis. It assigns three sentiment scores: positivity, negativity, and objectivity, to each synset of WordNet (Miller, 1995). In order to evaluate the reliability of the opinion oriented scores attached to each WordNet synset, the authors tested the accuracy of their tagging method. Firstly, they did this through evaluating their system against GI, and secondly, by using the more direct route of generating a GS annotated by human participants.

They selected GI because it was a lexicon which was fully tagged according to three opinion related labels positive, negative and objective. In the first phase, the authors compared SentiWordNet scores across GI labels at two stages: positive, negative and objective scores; and subjective and objective scores. The reason behind the choice of GI for benchmark selection is that the labelling of GI is very similar to that of SentiWordNet, i.e., providing positive, negative and objective scores.

The GS for the second phase was generated by selecting a subset of 1000 words from WordNet and separately labelling (annotating) these words by five different annotators. Each annotator assigned three scores (positive, negative and objectivity) to each word (synset) such that all three scores summed to 1.0. Comparisons among the scores assigned by different annotators to each word (synset) gave an inconsistency score (defined as an inter indexer inconsistency score). Special training was given to the five annotators where the meanings of all the words were clarified to them, in order to keep inter indexer inconsistency within specified limits. The comparison of the scores assigned by different evaluators to the same synset was analysed and provided an understanding of overall inter indexer inconsistency.

6.1.3 Thet et al. (2010) Clause Level Opinion Analysis

Thet et al. (2010) presented an opinion analysis approach based upon clauses. Their approach considered the opinion orientation as well as the strength of the opinion by associating opinion scores to the clauses. Their approach considered the grammatical dependency structure for clause level analysis and related opinion scores to a set of pre-identified aspects (overall, director, cast, story, scene, and music). These aspects are recognised as opinion topics.

Thet et al. (2010) divided the evaluation of their clause level opinion analysis technique into two phases. In the first phase, they used a GS approach and in the second phase, they implemented two baseline techniques in order to give a benchmark for comparison with their approach.

They used two datasets across both phases of their analysis. In the first dataset, they selected 34 movies in total (17 positive and 17 negative movies) based on user ratings. For the 34 movies discussion threads were selected randomly, and the positive and negative sentences in the post were collected manually. The data set contained 1000 manually collected sentences (500 positive and 500 negative movie review sentences) from the discussion board of a movie review site (www.imdb.com). The dataset contained groups of sentences which linked to each of the six pre-defined aspects. However, they believed that the manual construction of the dataset could have brought in some bias; therefore, they implemented a systematically constructed second dataset using a data crawler that filtered irrelevant discussion threads by analysing thread titles and the first post of each thread. Using this technique they generated a dataset containing 500 sentences. Following completion of this dataset construction, two annotators manually screened each sentence and the clauses contained based upon their relevance with the pre-identified aspects. This process resulted in the filtering of 158 sentences out from the automated dataset, to leave 324 relevant sentences. For each dataset, two annotators independently read the sentences and manually classified the opinion orientation towards each target aspect on the basis of positive, negative, and neutral.

Cohen’s kappa coefficient was used in order to find agreement between two independent annotators. It was found that the calculated values showed very good agreement for both datasets. Conflicting labels were reviewed and manually re-classified by one of the authors. This re-classification was used as an answer key (GS) for both datasets.

Precision, recall and F-score were determined for the review aspects of the clauses in datasets of 1000 sentences, where equation 6.1 – 6.3 are used.

$$\text{Precision} = \frac{\text{number of correctly tagged clauses}}{\text{number of automatically tagged clauses}} \qquad \text{Equation 6.1}$$

$$\text{Recall} = \frac{\text{number of correctly tagged clauses}}{\text{number of manually tagged relevant clauses}} \qquad \text{Equation 6.2}$$

$$\text{f} - \text{score} = 2 * \frac{(\text{precision} * \text{recall})}{\text{precision} + \text{recall})} \qquad \text{Equation 6.3}$$

In the second phase, they implemented two baseline approaches. Both their baseline approaches were very basic and did not consider any syntactic parsing, or grammatical dependencies. They only considered co-occurrence of opinion based words and aspects in a single sentence, in order to connect opinion based words to their respective aspect. If more than one aspect in a sentence existed they used a shortest distance algorithm to determine the relativity of opinion and aspect.

The first baseline approach used in the second phase; a word count approach, involved counting the number of positive and negative terms in individual sentences. If the count of the number of positive terms was higher than the count of the negative terms, the sentence/clause was considered as positive, or vice versa. The second base line approach; a sentiment score based approach, utilised the previously assigned scores to individual words in each sentence. The associated scores of positive/negative terms were aggregated across each sentence/clause. If the overall score was positive, the sentence was deemed to be positive, otherwise it was deemed to be negative. They calculated precision, recall, accuracy and f-score across each of the aspects. This provided them with benchmark values for comparison of their approach.

Where they defined precision, recall and accuracy as shown in equation 6.4 – 6.6,

$$\text{Precision} = \frac{\text{number of correctly classified positive or negative sentences or clauses}}{\text{number of automatically classified positive or negative sentences or clauses}} \qquad \text{Equation 6.4}$$

$$\text{Recall} = \frac{\text{number of correctly classified positive or negative sentences or clauses}}{\text{number of manually classified relevant positive or negative sentences or clauses}} \qquad \text{Equation 6.5}$$

$$\text{Accuracy} = \frac{\text{number of correctly classified positive and negative sentences or clauses}}{\text{number of manually classified relevant positive and negative sentences or clauses}} \qquad \text{Equation 6.6}$$

Both test phases (i.e., the GS experiments and the baseline implementation experiments) were repeated with the second dataset. The comparison of both sets of evaluations showed that in the first phase (using GSs) the manually selected dataset results outperformed the semi-automated dataset results, whereas, during the second phase of evaluation, the proposed system performed better with the semi-automated dataset. Both the datasets consistently yielded better results than the base line implementations. Error analysis was presented for detailed and careful analysis of sentences where system generated results differed from the GS results.

These results support the argument that the evaluation results are highly dependent on the datasets used, therefore direct comparisons of any opinion analysis technique with any existing research is not feasible without access to existing algorithms or experimental datasets. There is always a need to re-implement a comparative technique, in order to generate an acceptable benchmark.

6.1.4 Critique

It is observed from recent research (Das and Martins, 2007, Esuli, 2008) that the most acceptable way for testing the accuracy of annotation experimentally (whether manual or automated) is through the manual tagging of a complete dataset, based upon the requirements of the selected system. Manual annotation is a mechanism providing a representation of human performance. However, manual annotation is a very time consuming process (Tadano et al., 2009, Lu et al., 2011) which can also bring in bias based upon the knowledge base, and the background of the annotator (Greene, 2007). Manual annotation also brings in inconsistencies and ambiguities based on the understanding of the sentences, situations, and the mood of the annotators (Greene, 2007).

Apart from all the limitations associated with the manual annotation, it is determined that the most frequently used approach still centres on evaluation of performance against manually annotated datasets. The most preferred approach to manual annotation comparison is through the generation of GSs (Wilson et al., 2005, Esuli, 2008, Lu et al., 2011, Neviarouskaya et al., 2011, Lebanon et al., 2012). A GS reduces the size of the dataset (to be manually annotated) and focuses primarily on comparison with quality data. In order to minimise the effect of bias and to take steps in improving the consistency of data an approach using more than one expert annotator is preferred. Consistency can be improved through analysing the relationship between annotated datasets for each expert, determining solutions for levels of disagreement between the experts, and potentially establishing an agreed dataset as a GS.

However, perhaps due to the limitations of manual annotation, the GS approach is not used independently. It is always used in combination with another set of benchmark values to evaluate the system (Baccianella et al., 2010; Thet et al., 2010). These evaluation techniques also have their own share of limitations and strengths as presented in Table 6-1.

One of the approaches used alongside a manually annotated GS approach is to evaluate the system against a similar existing resource/research or resources/research with similar features (Esuli, 2008; Neviarouskaya et al., 2011). An example of such evaluation is the evaluation of SentiWordNet against the GI. However, the availability of similar resources is not always easy to obtain (Thet et al., 2010) as there is no standard forum/body which monitors, standardises, and publishes the research as standards in the field of opinion analysis.

Therefore, another way to evaluate the system is to evaluate it against a simulation of an existing resource or system; however, this process also raises many questions, as there is a need to align any simulated test directly with original test scenarios and resources. Original resources and datasets for regeneration of the research can prove problematic to obtain (Thet et al., 2010). As soon as resources and datasets change, the results can vary dramatically. Thet et al., 2010 determined that there can be substantial variation in system performance due to changes in datasets.

This leads to consideration of an evaluation technique of using other similar research and implementing their proposed systems by using the same test dataset(s). This approach can result in a removal of differences and issues raised by the use of different dataset(s). However, as noted above, the resources originally used in the research cannot always be reproduced (and are not always available e.g. algorithms and systems). Therefore the accuracy of any reproduction is always highly dependent upon the quality of the resources used, the quality of the implementation, and the amount of available system knowledge.

The complexity of using an implemented technique as a benchmark is also questionable. The technique should not be too simple that it may not generate a good benchmark. The manual selection of datasets for reproduced experiments and evaluation purposes may also bring bias to the datasets (Thet et al., 2010). However it is observed in the research by Thet et al (2010) that it might be necessary to manually select or manually screen automatically gathered datasets, otherwise the relevance of the data to the aspects, research question, or research goals can be questioned.

The most common measures used for the evaluation of IR systems and research are precision, recall and accuracy (Powers, 2011). However, there have been arguments that these measures are biased, and there is a need to properly define and gauge their measures (Alvarez, 2002; Powers, 2011). Thet (2010) has argued that recall and accuracy rely on the relevance of sentences. The relevance of sentences is not very easy to interpret, and is a challenging problem. Many researchers (Pang and Lee, 2005; Thet et al., 2010) identify that determination of relevance is an issue to resolve, with different approaches utilised. For example, in Thet et al. (2010) relevance is determined by manual selection of items in the construction of their dataset. Whilst in Pang and Lee (2005) relevance is determined by identifying the subjectivity of the dataset, as they only performed the sentiment analysis on sentences which have subjective information. Thet et al., 2010 determine that relevance still needs to be addressed and is a topic for independent further research. The relevance of sentences used in the evaluation of the novel approach presented in this Thesis is through manual selection of dataset.

As established earlier GS approach is most widely used approach for giving a benchmark for the evaluation purposes. The GS are always generated with set(s) of annotation performed by humans. There is always a need to measure the agreement level between annotators especially when annotators have freedom to categories text into more than one class. This gives a reliability measure for quality of annotation. Many researchers (Thet et al., 2010; Neviarouskaya et al., 2011) have used Cohen's kappa coefficient statistics to assess inter-rater reliability for establishment of GSs (the level of agreement between the annotators/raters/observers for assignment of categories to a categorical variable), orin order to find the level of agreement between two resources. Kappa has a value range from 0 to 1.00, where higher values indicate better reliability. However, the measure of Cohen's Kappa used quite extensively in literature is argued to be a biased measure (Powers, 2011).

Kappa can be used as a measure of independence between the annotators and as a measure to quantify the level of agreement between annotators. The Kappa measure takes into account the fact that sometimes agreement can only occur by chance; therefore, the use of kappa as a measure of independence is accepted. However, the use of kappa as a measure for the level of agreement between coders is not always reliable (Uebersax, 1987; Viera and Garrett, 2005) as sometimes a low value of kappa does not mean a low level of agreement between coders.

Kappa is only used for nominal levels of data, and it only gives a measure for agreement which does not consider various types and levels of disagreement. Kappa measure is not comparable across studies, procedures, or populations (Thompson and Walter, 1988; Cicchetti and Feinstein, 1990). For assessment of the ordinal level of data, researchers often select and associate weights to categories in order to calculate a weighted kappa score (Cohen, 1968; Uebersax, 1987). There has been work conducted in order to draw equivalence between weighted kappa and interclass correlation coefficient as a measure of reliability (Fleiss and Cohen, 1973).

Table 6-1: Approaches used for evaluation of opinion based research

Approaches to Evaluate Opinion Analysis Systems				
	Manual Tagging Of Complete Dataset (Das and Martin, 2007)	Gold Standard (Wilson Et Al., 2005)	Evaluate Against Existing Resource With Similar Features	Evaluate Against The Simulation Of Existing Resource
STRENGTHS	• More accurate annotation in reference to human performance • More complete annotation as whole dataset is annotated	• Minimises limitations associated with manual annotation. • A way to achieve standard Statistical approach of sample collection,	• Pre-evaluated resources	• Pre-evaluated resources
LIMITATIONS	• Time consuming • May involve bias based upon the knowledge base and mood of individuals. • Inconsistent • Ambiguous	• Reduced size of dataset. • Used with other comparative approaches. • Can still involve bias based upon the knowledge base and mood of individuals.	• Harder to find research with similar features. • No regulation/ standardizing authority. • No direct comparison is possible as datasets and resources used for evaluation of earlier research are different. • Issue of relevancy of importance.	• Need to align simulation test directly with original resources used. • Need to use same dataset for comparison • Issue of relevancy of importance.

6.2 Evaluation Goals

The research questions outlined in Chapter 1 were:

- Are there improvements targeted at clause level opinion analysis, making use of phrases that can be made to existing state-of-the-art systems, which can bring the process of automated opinion analysis closer to manual 'expert' performance levels?
- Does clause level analysis provide opportunities for the identification of additional information (like phrases and targets of opinions) which can be used to support opinion analysis?

A novel opinion analysis approach is proposed in Chapter 3, which posits improvements based upon the research question presented in Chapter 1. There is a need to evaluate the proposition; this evaluation process needs to focus on a determination of the distance between the performance of the proposed opinion analysis approach, and the performance of human experts for opinion analysis. In addition, a technique can be used to gain an understanding of acceptance of system generated output. Therefore the initial evaluation goals set are to:

> E1 - Evaluate system performance in relation to expert human performance.
>
> E2 - Evaluate system performance (opinion identification in relation to general human performance).

In addition to the evaluation of the proposed approach against the performance of human experts, there is a need to establish the improvements ascertained by the realisation of internal structures (phrases). Therefore, another evaluation goal is to:

> E3 – Determine if the system performance is improved by the introduction of internal structures (phrases) to opinion analysis.

In addition to the proposal of a novel opinion analysis approach, the author has also presented a framework, which provides a supportive infrastructure for the generation of a corpus. As the framework is mainly comprised of two parts, the opinion analysis technique and the generation of a corpus. The opinion analysis technique is evaluated through E1-E3. The evaluation of the second part of the framework (corpus) is evaluated through E4. There is a need to evaluate the potential usability of the corpus, though an evaluation of its applicability and re-usability. Therefore, evaluation goal E4 is:

E4 - To determine the applicability of the corpus in relation to IR.

The following section provides a clear description of the evaluation plan which enables data to be collected and analysed, to support the achievement of the evaluation goals.

6.3 Evaluation Plan

It is observed that all the evaluation goals rely on the assessment of the performance of proposed opinion analysis technique with human expert performance (E1), general human performance (E2) and improvements introduced by the enhancement of phrase analysis in the proposed approach (E3). In order to assess the performance of the proposed approach, there is a need to establish the results based upon the approach; therefore, the most appropriate mechanism is through experimental prototype development. Therefore, a proof of concept prototype (P1) enables the opinion analysis approach to be evaluated (meeting E1and E2) and provides a test corpus which later help for E4. The proof of concept prototype is discussed in Section 6.4.1.

There is a need to establish the improvements in the opinion analysis approach, brought in by the introduction of phrase level analysis into the opinion analysis process (E3). The direct comparison of the results with other similar systems is not possible as, the author was unable to identify any system having a similar approach in combination with phrase level analysis. The simulation of previously conducted research, which is closely related to the proposed innovative approach, is also not possible as it would require access to the various resources originally used to implement existing opinion analysis approaches. However, these resources are not readily available (Thet et al., 2010). The majority of resources used while implementing related research works are not publically available and there is no standard body or forum to publish the research and resources. Therefore, instead of choosing to simulate any existing research, a basic opinion analysis approach without the enhancements of phrase level analysis is implemented (Experimental Prototype P2) in order to give a benchmark for comparison (meeting E3).

The primary output of the experimental prototype (P1) is the establishment of opinion frames (Section 3.4.2 provides details of the opinion structure). These opinion frames are the core input for the proposed corpus design (provided in Section 5.1.4). Therefore, these opinion frames can further be used in order to verify the reusability of the corpus and to help determine the accuracy of information retrieved from the proposed opinion based corpus (meeting E4).

The second experimental prototype P2, follows the same flow for the analysis as that of P1. It takes data from an established dataset, analyses opinion, and presents the output in a similar tuple as that of P1 (Subject, Object, Verb, Opinion Words and Opinion Topics). However, the opinion analysis approach used for P2 is different from that of P1. P2 takes input in the form of sentences, uses a word based dependency structure and identifies the clauses. In order to calculate the opinion scores attached to each word P2 uses the same resources as that of P1, i.e., WordNet and SentiWordNet. P2 does not identify and use phrases for opinion analysis, opinion scores are calculated using SentiWordNet and are attached to each clause/sentence. The output is generated as a tuple, which is also based upon individual words. The difference in process for opinion analysis in both P1 and P2 is shown in Figure 6-1. The analysed tuple along with opinion orientation will be saved into XML files for evaluation of analysed information with respect to P1 output.

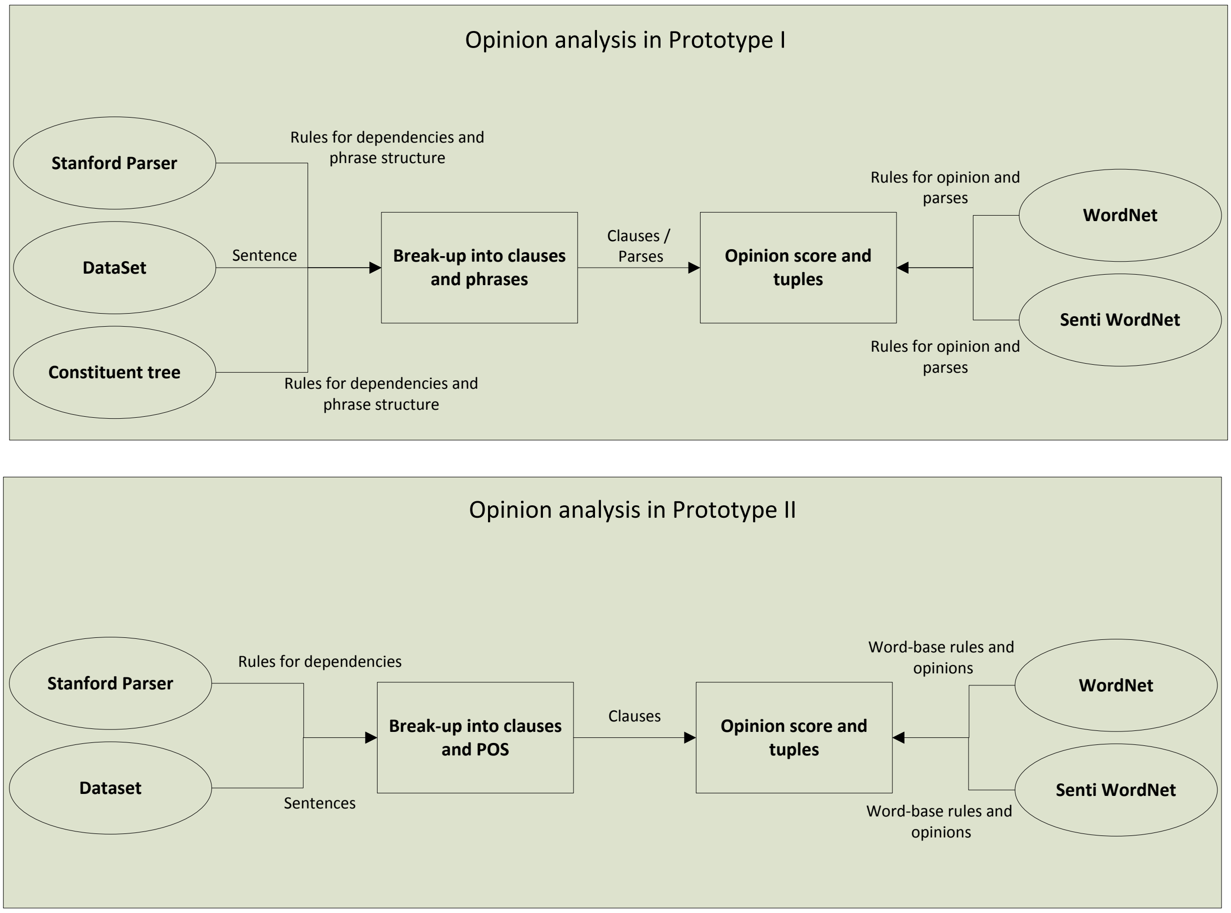

Opinion analysis in Prototype I
Stanford Parser
DataSet
Constituent tree
Rules for dependencies and phrase structure
Sentence
Rules for dependencies and phrase structure
Break-up into clauses and phrases
Clauses / Parses
Opinion score and tuples
Rules for opinion and parses
Rules for opinion and parses
WordNet
Senti WordNet
Opinion analysis in Prototype II
Stanford Parser
Dataset
Rules for dependencies
Sentences
Break-up into clauses and POS
Clauses
Opinion score and tuples
Word-base rules and opinions
Word-base rules and opinions
WordNet
Senti WordNet

Figure 6-1: Comparison of opinion analysis in P1 and P2

The evaluation and validation process for the proposed opinion analysis approach follows a GS approach along with the implementation of a second prototype P2 in order to provide a benchmark performance. The GS approach requires the development of single or multiple GSs involving manual annotation of the dataset by experts. Manual annotation has been previously criticised as inconsistent, ambiguous, time consuming, and as a process which can contain bias, in Section 6.1.4. Therefore, there is a need to establish mechanisms to minimize the limitations of the approach. As outlined in Table6-1, the GS approach when compared to the other approaches identified in the analysis in Section 6.1.4 has multiple benefits which outweigh the issues involved in reducing the limitations associated with it. Aside from these limitations, it is observed that the GS approach is the most appropriate and most frequently utilised approach for the establishment of a benchmark of human performance. However it is established in Section 6.1.4, that it is generally used with another approach to generate a benchmark. Therefore as explained earlier a system based comparison using a second prototype system, P2, is also implemented.

The next section explains in detail the design of the evaluation process enabling the evaluation of the proposed analysis approach and corpus design.

6.4 Core Components of the Evaluation Process

The execution of the evaluation plan is dependent on the following five agents.

6.4.1 A Proof of Concept Prototype System

As outlined in the evaluation plan above, a proof of concept system integrating the novel opinion analysis algorithm with existing resources is required to meet defined evaluation goals. This proof of concept system is outlined in Figure 6-2. The system can essentially be broken into three stages: I, II and III.

A proof of concept prototype requires the utilisation of existing resources like, Stanford Parser, WordNet, SentiWordNet, Penn Tree Bank (Penn_Treebank, 1992; Miller, 1995; Esuli and Sebastiani, 2006b; Www.Stanford.Edu, n.d). All these resources are the results of independent existing research; therefore, validation is not required to verify their reliability. However, the rules and conditions bringing these resources together and the approach itself require *validation* and *verification*. This can be achieved through verification/validation of syntactic analysis; establishing specification of the subjects (Nouns/Noun Phrases), objects (Nouns/Noun Phrases) and verbs within provided sentences (meeting E1).

A dataset is established based on the specifications and requirements of the research as explained in Section 6.4.2. In stage I, each sentence is taken from the dataset in order to identify the clauses and phrases within the sentence. This identification of clauses and phrases is performed by using and analysing the output from the Stanford Parser and the constituent trees. The Stanford Dependencies and constituent tree are brought together with the help of some rules established in order to understand the internal relations of words to make phrases and identify dependencies based upon phrases. In stage II, these phrases are further used for the identification of Subjects, Objects and Verbs within the clauses. Opinion words and phrases are identified and phrase level opinion scores are calculated with the help of the SentiWordNet lexical resource using the algorithm presented in Chapter 3. Opinion phrases and the dependencies help in the identification of opinion topics.

SentiWordNet is a word based lexical resource, having word based polarity scores. The algorithm presented in Chapter 3 is used to calculate phrase level scores. These phrase level opinion scores are mapped on a seven level opinion based scale which is also explained in Chapter 3. As a result of the opinion analysis, values for opinion polarity and opinion intensity are assigned to each clause. Later in stage III these values are aggregated at sentence level, based on the rules developed from the relations, the conjunctions and the Opinion Topics. Sentence level opinion analysis generates opinion tuples which are saved into XML files in the form of corpus. The design of the corpus is presented in Chapter 5.

There is a need to *verify* the values of polarity and intensity, in order to find out the *validity* of these values. Therefore, the scope of Subjects and Objects are correctly identified for the opinion based words within the clauses, and sentences, in order to generate the opinion frames. These frames are used to populate the opinion based corpus, based upon the corpus design as presented in Chapter 5.

Netbeans IDE 7.0.1 is used to implement the prototypical solution of the proposed approach. Java is used as the programming language to construct the prototype. As Java is open source and provides compatibility with most of the available resources, which might be needed later for future work. Sentence structure and de-constructing sentences into basic units (words) as well as understanding their relations within and between sentences is important in an opinion analysis approach. Therefore, implementation of the proposed approach employs the Stanford Parser. The Stanford Parser API is used as it provides the complete sentence structure and dependencies within a sentence. Another important building block for the opinion analysis approach is the semantics of the words used. Most existing implementations require manually annotated corpora to understand the meanings and senses of the words encountered during analysis. In order to give the proposed approach a more standard approach. Two lexical resources: WordNet and SentiWordNet are used to establish the polarity and score associated with different words. The polarity and score are totally dependent on the POS tags associated to a particular word within a sentence. A dataset of 600 test sentences is used; the details of this dataset are given in Section6.4.2. The proposed algorithm (mentioned in Chapter 3) apply after retrieving all the required information from the resources mentioned above. This algorithm produces the information that is stored finally into XML files. The stored information is in a structured format (given in Chapter 5) that can be queried to discover the accuracy of the evaluation process.

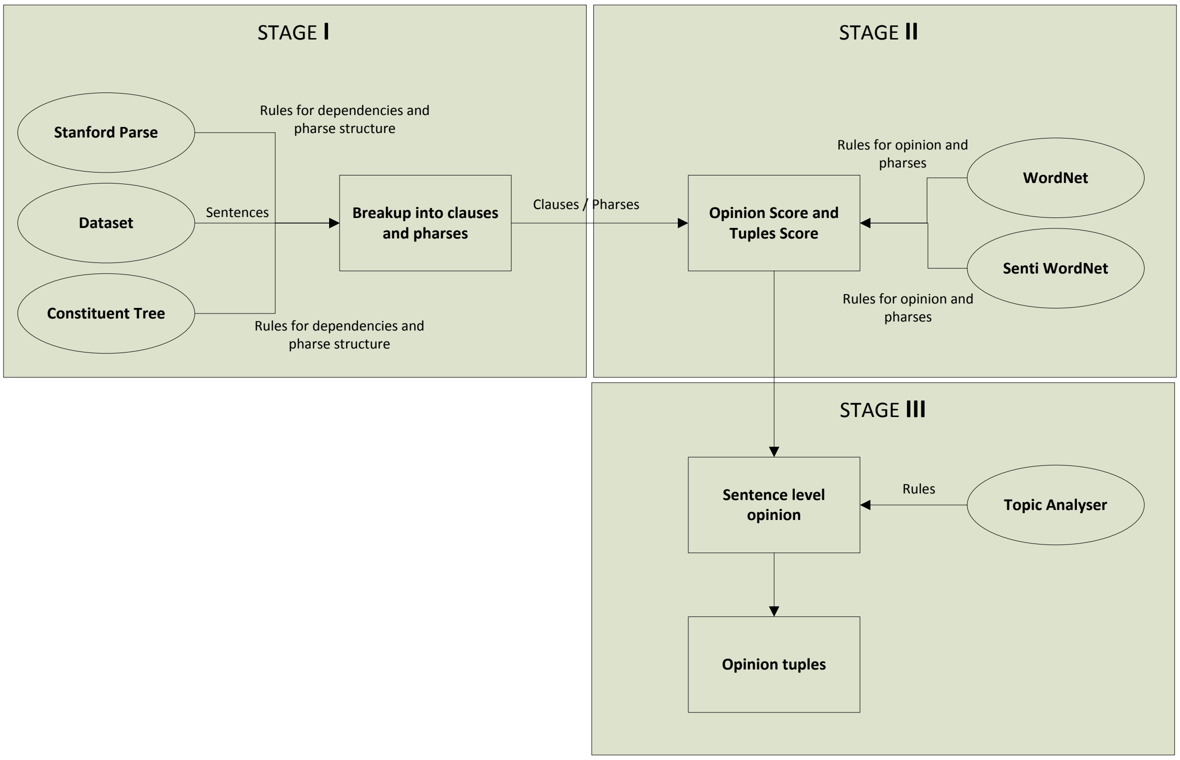

STAGE I
Stanford Parse
Dataset
Constituent Tree
Rules for dependencies and pharse structure
Sentences
Breakup into clauses and pharses
Rules for dependencies and pharse structure
Clauses / Pharses
STAGE II
Opinion Score and Tuples Score
Rules for opinion and pharses
WordNet
Senti WordNet
Rules for opinion and pharses
STAGE III
Sentence level opinion
Rules
Topic Analyser
Opinion tuples

Figure 6-2: Outline of Proof of Concept Prototype P1 & Prototype P2

6.4.2 Dataset

In order to evaluate different aspects of opinion analysis and to generate an opinion based corpus, a dataset is needed to be used during the implementation of the prototypical solution. Therefore, a dataset of 600 manually gathered sentences from existing research datasets (Bethard et al., 2005, Jindal and Liu, 2006, Kim and Hovy, 2006, Kessler et al., 2010, FrameNet, n.d.) is selected. Manual selection of sentences for the dataset is undertaken including identification of different aspects i.e., negation, domain independent, conjunction, comparative and complex sentences. Manual construction of the dataset also helps to filter out any non-relevant sentences, opinion spam (non-independent online posts that give positive reviews about some product or movie in order to promote it; or unfair negative reviews provided in order to damage reputation), and non-subjective sentences etc. The automated detection of opinion spams and non-relevant sentences is not part of the current research. However, research currently exists which focuses on issues like opinion spam (Jindal and Liu, 2008) and subjective/objective sentences (Pang and Lee, 2005; Wiebe and Riloff, 2005). A topic is also attached to each sentence, thus enabling the assigned topic to be used for the testing of IR from the corpus (meeting E4).

Admittedly there may be a small level of bias introduced due to manual selection and construction of the dataset; however, as discussed in Section6.1.4 manual selection and/or screening of sentences can be used to improve the overall quality of the dataset by providing a higher relevance of the sentences to the overall requirements of the system.

6.4.3 Gold Standard Dataset

A benchmark GS is generated over a smaller subset of the larger dataset. This subset is generated by automatic random selection of 50 sentences from the initial dataset. This automated random selection of a subset of sentences helps to control the bias introduced with the initial manual selection of sentences for the dataset.

This subset of 50 sentences is annotated by two native English expert annotators. The expert annotators are provided with clear guidelines and instructions for the annotation of sentiment orientation, and the structure of sentences. Conflicting labels are observed by the researcher while reviewing the annotation. These conflicting labels are reviewed and discussed during the sessions with both the experts in order to comprehend their understanding. However, during these discussion sessions, it is realised that there is a level of disagreement between the experts understanding for the structure of the sentences. Therefore, in order not to spoil the original understanding of the experts about the opinion and structure of sentences, two GSs are defined: GS1 and GS2. GS1 and GS2 are two annotations performed by each of the experts.

Determination as to the expert status of the two individuals is made through evaluation of their prior experience. However, it should be noted that they are not professional annotators, and that training is required in how to use any annotation scheme. The experts are chosen such that there is a good level of confidence in their decisions. Having two expert evaluators helps to limit the bias, level of inconsistency, and ambiguity within the annotated materials produced. However, it could be argued that having even more experts can further minimise the limitations, however having more experts can bring issues related to their background knowledge, experience and diversified background knowledge. Therefore the decision about experts is balanced against issues which can be raised by increasing number of experts. There is no appropriate number for expert participants presented in the literature that can be considered to be suitable to minimise the limitations.

6.4.4 Participants

150 human participants from different backgrounds and fields are requested to annotate (opinion orientation and strength) the complete set of 600 sentences as the annotation process does not depend on any formal training or professional background. However, a number of examples and guidelines are provided before the participants are involved in annotating the sentences. The diversity in the backgrounds and knowledge of this group of participants bring an understanding of how people from different backgrounds can interpret opinions from given textual data. A snowball sampling technique is used (Atkinson and Flint, 2004). The request is forwarded to a number of research groups in the areas of computer science, data mining, opinion analysis and psychology at Hull, Manchester, Sheffield, Salford, and Beijing. The group members are further requested to re-transmit the request to other individuals.

The main reason behind the selection of the use of a snowball technique for accessing participants is to limit restrictions on the background of the participants. However, through association of the participants with research groups the participants should have a fair level of understanding about the nature of the research, delivering better quality data. It is observed that like most survey oriented techniques there can be issues (such as low response rates, and lower quality responses) related to participant take-up particularly in circumstances where the survey requires an extensive thought process. Therefore, the survey is kept simple and only asks participants about their understanding of opinions in written text. They are provided with radio buttons in order to make their selection of opinion. Each participant is sent 20 random sentences. This gives a response of 3000 sentences, i.e., 600 sentences each annotated by 5 participants or 150 participants each responding to 20 sentences (600*5 / 150*20). Similarly 150 participants are sent with the same 600 sentences for validation of the proposed opinion analysis research, (output of P1). This time again they are provided with examples and only asked for their level of agreement over a likert scale, (by selection on radio buttons).

There is no explicit effort made to guarantee that both sets of participants remain mutually exclusive. However the chances of the same sentences being sent to the same participant for both stages (evaluation of opinion and validation of opinion) are very rare. As each sentence is evaluated by five participants and each participant has 20 randomly selected sentences out of a pool of 600 sentences.

No personal information is collected for each of the participants, therefore it is difficult to make sure that the same sentence is not evaluated and validated by the same participant.

6.4.5 Ethics Permission

No personal data is gathered in the evaluation process. The participants are only asked for their understanding of opinion and their level of agreement with the system output (P1). This data input helps in analysing the captured information and understanding the trends in the data.

6.5 Evaluation Test Bed

The evaluation process for the opinion analysis approach is divided into four phases based upon the evaluation goals defined in Section 6.2.

The first phase as shown in Figure 6-3, provides a focus on the evaluation of the system performance against expert performance (with the help of P1) on a basis of syntactic analysis, opinion analysis, and the generation of opinion frames (meeting E1). Both GSs are used as benchmarks for the syntactic and lexical analysis of the opinion frames generated using the defined novel opinion analysis approach, with the help of the proof of concept prototype system (P1). P1 is explained in detail in Section 6.4.1.

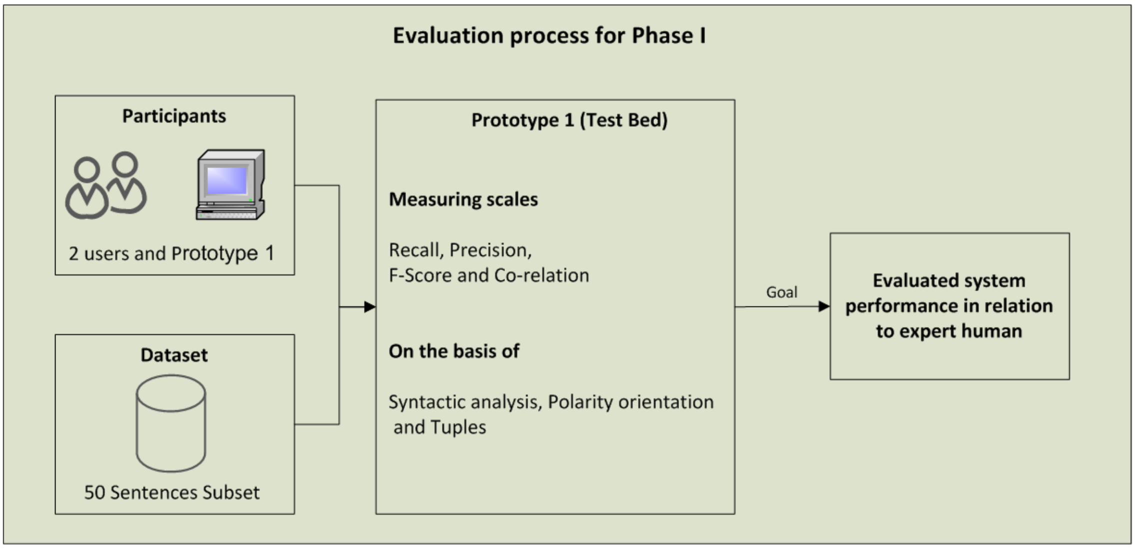


Figure 6-3: Process for Phase 1 of Evaluation Process

In order to give an acceptable level of disagreement (in terms of linguistic and opinion analysis) between the GS(s) and the output generated using prototype (P1), the same subset of the dataset is annotated by two experts. Dissimilarities in the annotation for opinion orientation and strength by experts with each other are analysed (using correlations) to provide an understanding of the level for the acceptable difference between the opinion identified by the proof of concept prototype system and the GSs. Then correlation of the system with each GS is calculated and analysed with the benchmark set by the correlation of both expert GS(s).

The proposed opinion analysis approach also performs the syntactic deconstruction of sentences by identifying Subjects, Objects and Verbs as parts of a sentence and the opinion words and their opinion topics as components of an opinion. The P1 output is compared with GS1 and GS2, and then measures of recall, precision and f-score are calculated. The analysis of P1 output with respect to GS1 and GS2 gives a level of performance for the proposed technique with respect to expert human performance.

The second phase mainly *evaluates*, *validates* and *confirms* the opinion related information associated with each sentence by the novel opinion analysis approach (meeting E2). During this phase an evaluation and verification process is adopted. First of all, the complete dataset of 600 sentences is sent to 150 participants in the form of a web based survey. Each participant is sent with 20 sentences and each sentence is sent to 5 participants for annotation of opinion polarity and strength. The dataset captured in this phase is compared with system P1 generated output. The participants are provided with seven level Likert scale polarity options (Strongly Positive, Mildly Positive, Weakly Positive, Neutral, Weakly Negative, Mildly Negative, Strongly Negative). They are asked to select the most appropriate value based upon their interpretation of opinion in sentence.

There are six opinion values assigned across each sentence, five from non-expert participants and one system P1 output. The main reason behind selection of five participants is to understand, how human understanding for opinion in written text varies. This variation in non-expert understanding gives an acceptable level of variation between human response and P1 output. Therefore for each sentence the mean for non-expert opinion values are calculated. This mean value of opinion is used to calculate Standard Deviation (SD), of non-expert opinion annotation, for opinion across each sentence. This gives an acceptable range of values for P1 output. i.e., [Mean + SD, Mean –SD]. So all the system P1 responses within the range of [Mean + SD, Mean – SD] are acceptable.

The percentage of the total of 600 sentences falling within the range of [Mean + SD, Mean –SD] gives the performance measure of the system in comparison to non-expert performance

In the verification stage, the dataset is analysed using the proof of concept prototype system (P1). The opinion scores calculated during the analysis are mapped on to a seven level scale for opinion polarity and strength as shown in Chapter 5. Later, these sentences along with their calculated opinion polarity and strength are sent to 150 participants, using a web based survey. A survey is constructed in a way similar to first stage, with each sentence verified by 5 different participants. Each participant is sent with 20 sentences in a web based survey. This provides 3000 sentences verified for opinion polarity and strength as calculated by the prototype system (P1). Participants are asked for their level of agreement with the system performance (P1 output) and to select the most appropriate value from the likert scale (Strongly Agree, Agree, Indifferent, Disagree, and Strongly Disagree).

The second set of responses gives 3000 responses from 150 participants and a percentage level of agreement is calculated in order to understand the proportion of participants which agree/disagree with P1 output.

The process and its participants for second phase are depicted in Figure 6-4.

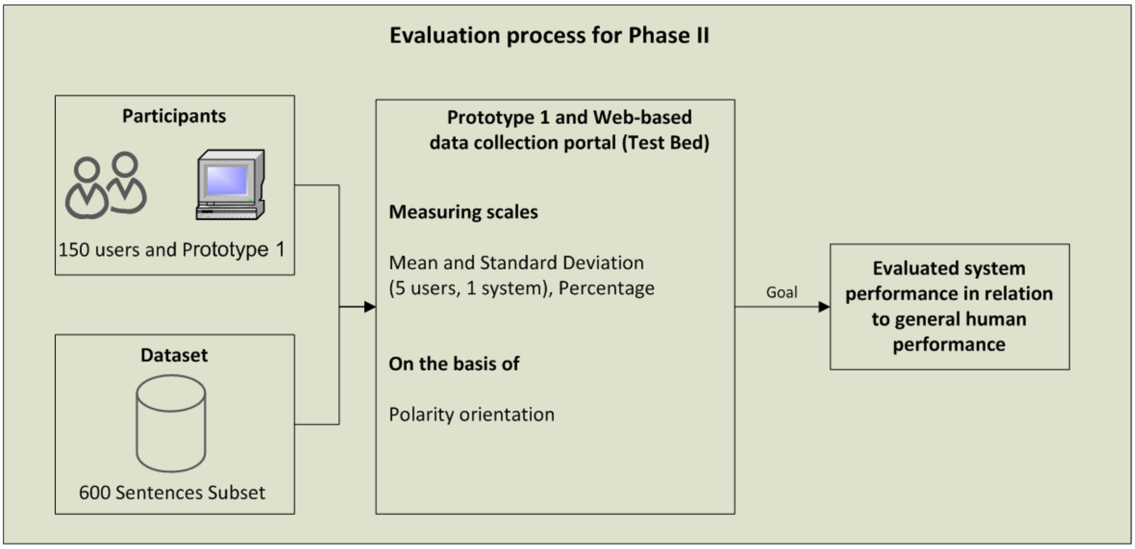

Evaluation process for Phase II
Participants
150 users and Prototype 1
Dataset
600 Sentences Subset
Prototype 1 and Web-based data collection portal (Test Bed)
Measuring scales
Mean and Standard Deviation (5 users, 1 system), Percentage
On the basis of
Polarity orientation
Goal
Evaluated system performance in relation to general human performance

Figure 6-4: Process for Phase 2 of Evaluation Process

The third phase evaluates the improvements in the proposed opinion analysis approach based upon the identification and analysis of phrases (meeting E3). A prototype representation of the novel opinion analysis approach without the consideration of phrases (P2) is implemented and evaluated against benchmarks GS1 and GS2. This process provides an in-depth analysis of a smaller subset of data using outputs from both prototype systems (P1 and P2). This smaller subset is captured by randomly selecting sentences to create a subset of 50 sentences. The process is described in Figure 6-5. Both prototypes P1 and P2 are implemented using the same resources for lexical and linguistic analysis. They only vary on a basis of rules identifying phrases and understanding phrase level analysis. The opinion analysis approach proposed in this Thesis mainly relies on lexical resources, and the lexical resources used for both the prototypes P1 and P2 are the same. Therefore there might not be any significant difference in the opinion analysed by both prototypes. However, the structure and dependency analysis may be significantly affected by the phrase level analysis, and the main motive of E3 is to explore the structural deconstruction of the sentences and comparing both P1 and P2 outputs with GS(s) to understand the difference captured by phrase level analysis in P1.

.

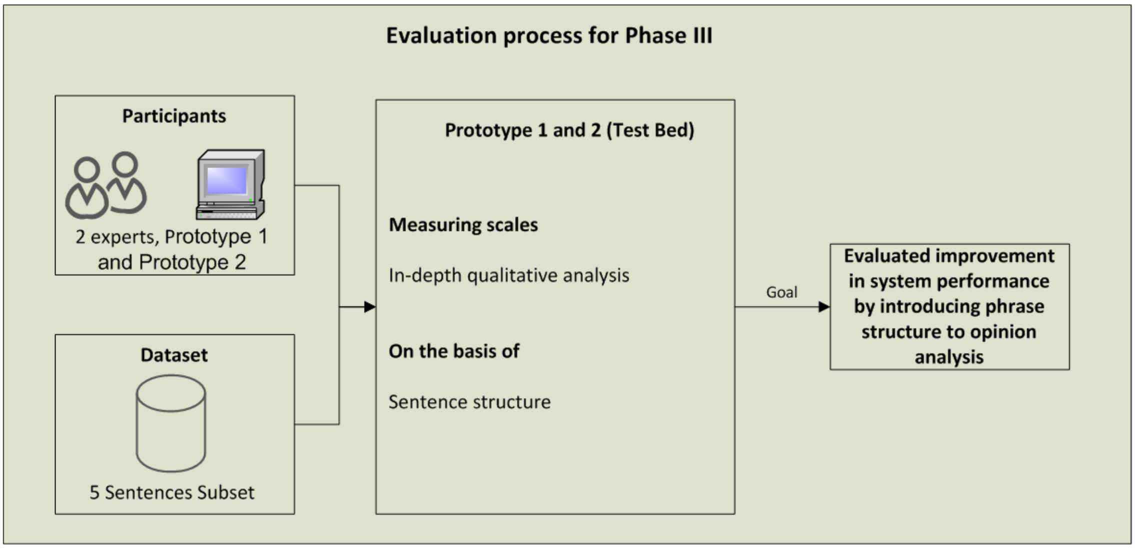

Evaluation process for Phase III
Participants
2 experts, Prototype 1 and Prototype 2
Dataset
5 Sentences Subset
Prototype 1 and 2 (Test Bed)
Measuring scales
In-depth qualitative analysis
On the basis of
Sentence structure
Goal
Evaluated improvement in system performance by introducing phrase structure to opinion analysis

Figure 6-5: Process for Phase 3 of Evaluation Process

In order to verify the IR abilities, the corpus generated using the proof of concept prototype (P1) is used in the fourth phase of evaluation (meeting E4). The corpus based upon the proposed design is constructed with the help of P1. This corpus is the resulting output of the opinion analysis conducted through P1. Data in the corpus is sorted for topics, opinions, and different parts of sentences. In order to accomplish E4, to verify the reusability and applicability of corpus, a small search engine is implemented. The primary reason for this search engine is to measure the effectiveness of the corpus by measuring the retrivability of data for later use. This implementation is based upon an XML based search, however this implementation only uses basic XML based searching techniques, and no emphasis is placed on the efficiency of search performance (searching algorithm). Therefore the effectiveness is only measured based upon the relevance of the retrieved information and other features like optimization and ranking of retrieved information are not implemented or considered during evaluation.

The process for the fourth phase is shown in Figure 6-6. Queries are processed for the search against keywords and the presence of opinion in them. Search results are calculated against assigned topics and opinions, on the basis of IR values calculated (precision, recall, fallout and missout) as suggested by Egghe (2008).

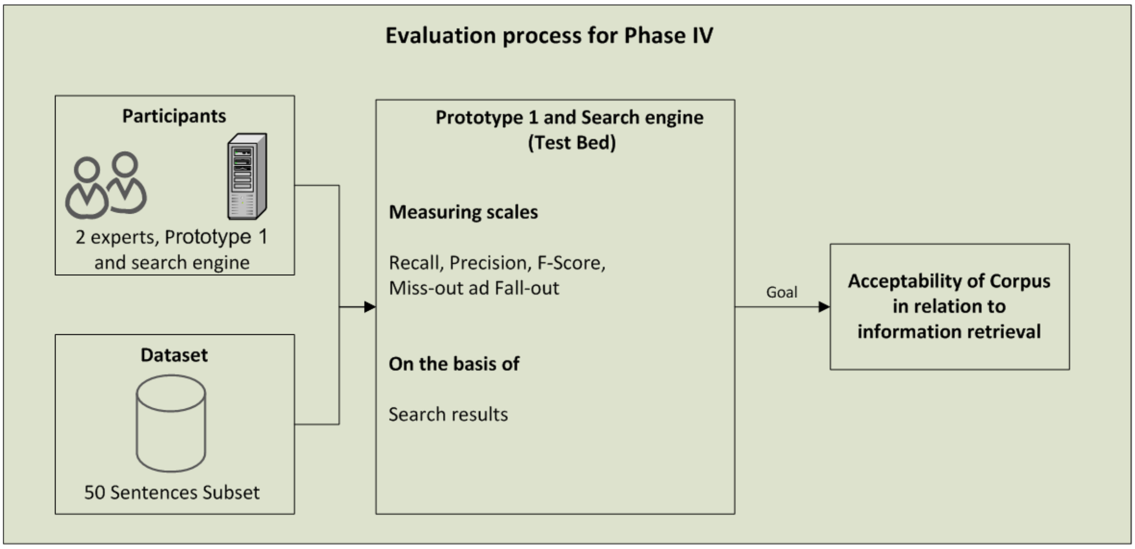

Evaluation process for Phase IV
Participants
2 experts, Prototype 1 and search engine
Dataset
50 Sentences Subset
Prototype 1 and Search engine (Test Bed)
Measuring scales
Recall, Precision, F-Score, Miss-out ad Fall-out
On the basis of
Search results
Goal
Acceptability of Corpus in relation to information retrieval

Figure 6-6: Process for Phase 4 of Evaluation Process

The *tractability* and *retrievability* of the required information in the corpus is calculated, in order to test the *reusability* and *applicability* of the corpus (E4). The efficiency and speed of the search engine are not considered. This search engine is based on two types of queries.

1. Queries based on objects (opinion topic) or any feature or aspect of them, for example: a product (e.g. the iPhone) and associated opinions with this. So, as an example the complete query could be "positive reviews about the iPhone".
2. Queries based on the subjects (opinion holders), person or organisation holding opinions about any particular topic. For example: Steve Job's opinion, about the failure of Apple I and Apple II computers.

The queries are constructed with phrases: generally noun phrases (opinion topics/opinion holders) and opinion words/phrases, as the corpus contains the results of phrase level analysis.

6.6 Measurements

The measures of precision, recall and f-score (Thet et al., 2010) for each opinion frame (i.e., opinion words, and opinion topic, opinion intensity and polarity) generated by the prototype system are calculated against both GS1 and GS2. Fallout and missout along with precision, recall and f-score are calculated to determine the quality of IR from the designed Corpus.

Precision can be defined as a measure of confidence (Powers, 2011). It is used to specify the proportion of retrieved items that are judged by the experts to be relevant (Alvarez, 2002). Therefore it is determined to be the proportion of real positive cases identified and annotated by the system (automatically) that are correctly predicted as positive by the experts (GS) (Powers, 2011). Recall measures retrieval coverage defined as the proportion of the set of relevant items that is retrieved by the system (Alvarez, 2002), which means recall is a measure of sensitivity (Powers, 2011). It is observed that precision and recall have a reverse relation therefore they are not discussed in isolation. Instead they are measured in comparison with a fixed level of measure, or are combined into one measure such as f-score or accuracy. F-score is the weighted harmonic mean of the precision and recall, where both recall and precision are evenly weighed (Manning et al., 2009). Accuracy measures the fraction of correct classifications. It is argued by Manning et al (2009) that accuracy is not an appropriate measure for IR as in almost all cases the data is skewed. Almost 99.9% of the documents are from a non-relevant class (Manning et al., 2009).

The first evaluation goal (E1) examines system performance in relation to human expert performance. In order to do this the system output and expert annotation for opinion orientation and sentence decomposition are evaluated. For opinion orientation, performance can be analysed against the likert scale values associated with each opinion score, and can be generalised on the basis of polarity.

Therefore, measures of precision and recall for opinion orientation are calculated on two levels: one is the exact system match to the benchmark GS according to the likert scale. For this purpose the correlation between experts' results and each of the GS with system output are calculated. The correlation between the GSs gives a benchmark value for the correlations between the experts and the system.

The algorithms for Precision, Recall and F-Score for opinion orientation are given below in equation 6.7 – 6.9:

$$\text{Precision} = \frac{\text{No. of correctly classified opinion orientation for sentences}}{\text{No. of all automatically classified opinion orientation for sentences}} \qquad \text{Equation 6.7}$$

$$\text{Recall} = \frac{\text{No. of correctly classified opinion orientation for sentences}}{\text{No. of all manually classified opinion orientation for sentences}} \qquad \text{Equation 6.8}$$

$$\text{F} - \text{score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \qquad \text{Equation 6.9}$$

On the second level, the opinion orientation scale can be simplified to Positive, Negative and Neutral values. This revised scale is used to map the opinion orientations of the dataset for the system, and both GSs. This results in a requirement for differences in the precision and recall values as shown in equation 6.10 and 6.11.

$$\text{Precision} = \frac{\text{No. of correctly classified positive/negative sentences}}{\text{No. of all automatically classified positive/negative sentences}} \qquad \text{Equation 6.10}$$

$$\text{Recall} = \frac{\text{No. of correctly classified positive/negative sentences}}{\text{No. of all manually classified positive/negative sentences}} \qquad \text{Equation 6.11}$$

For the analysis of sentence decomposition the calculation of precision and recall are a little more complex. For each classification (e.g. verbs, opinion oriented words etc.), the GS results and the system results are categorised into one of four different forms: True Positive; True Negative; False Positive; False Negative as shown in Table 6-2. These forms can be more easily understood through an abstract example, as below:

Sentence = a b c d e f g - where each letter corresponds to a word.
GS = {a, b, c, d, f}
SR = {a, b, c, d, e}

In all circumstances where the system results (SR) contain words which are a match to the GS they are regarded as True Positives. In the above example the set {a, b, c, d}.

In cases where words are contained within the SR but not contained within the GS these are regarded as False Negatives. In the above example the set {e}.

In cases where words are contained within the GS but not contained within the SR these are regarded as False Positives. In the above example the set {f}.

In cases where words are not contained in either GS or SR but are part of the sentence these are regarded as True Negatives. In the above example the set {g}.

Table 6-2: Confusion Matrix

	SystemOutput (Observations)	
Benchmark Expert Annotations (Expected Results)	True Positive	False Positives
	False Negative	True Negative

These measurements can then be mapped onto precision and recall calculations using the algorithms detailed below in equations 6.12 and 6.13:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \qquad \text{Equation 6.12}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \qquad \text{Equation 6.13}$$

Which can further be defined as equation 6.14 and 6.15:

$$\text{Precision} = \frac{\text{No. of correctly classified subjects/objects/verbs/opinion words/opinion topics}}{\text{No. of all automated classified subjects/objects/verbs/opinion words/opinion topics}} \qquad \text{Equation 6.14}$$

$$\text{Recall} = \frac{\text{No. of correctly classified subjects/objects/verbs/opinion words/opinion topics}}{\text{No. of all manually classified subjects/objects/verbs/opinion words/opinion topics}} \qquad \text{Equation 6.15}$$

For the second evaluation goal (E2), there are two requirements; firstly, there is a need to verify the overall opinion orientation assigned by the system to the dataset, and secondly there is a need to understand an allowed level of disagreement among the non-expert human evaluators. These measures are calculated through measurement of correlation between the human participants (in order to identify an acceptable level of disagreement for understanding of opinion orientation among human beings) and the overall level of percentage for agreement and disagreement (over a range of Strongly Agree, Agree, Indifferent, Disagree, and Strongly Disagree) in order to measure the level of system performance for identification of opinion orientation.

The third evaluation goal (E3) analyses the improvements in system performance for the decomposition of each sentence structure, as introduced by the in-depth analysis of phrases. This goal is achieved by comparing the detailed breakdown of each sentence as performed by the system, and analysing these against the benchmark GSs. This analysis takes place through a qualitative narrative evaluation of the comparative datasets.

The fourth evaluation goal (E4), also makes use of the measures of precision and recall, along with the addition of the measures of fallout and missout (Egghe, 2004). All measures are used for the evaluation of the corpus through search engine retrieval. The measures of precision, recall, fallout and missout are calculated based upon the opinion topics and opinion orientation data within each of the GSs. Formulas for the application of these measures in this analysis are provided below in equations 6.16-6.19:

$$\text{Precision} = \frac{\text{Retrieved} \cap \text{Relevent}}{\text{Retrieved}} \qquad \text{Equation 6.16}$$

$$\text{Recall} = \frac{\text{Retrieved} \cap \text{Relevent}}{\text{Relevent}} \qquad \text{Equation 6.17}$$

$$\text{Fall out} = \frac{\text{Retrieved} \cap \text{Non Relevent}}{\text{Non Relevent}} \qquad \text{Equation 6.18}$$

$$\text{Miss out} = \frac{\text{Not Retrieved} \cap \text{Relevent}}{\text{Not Retrieved}} \qquad \text{Equation 6.19}$$

In this context, recall measures how well the search engine performs in finding all the relevant data items for the query in the search space (corpus). Precision measures how well the search engine performs in rejecting non-relevant data items. Finally, fallout is the measure for non-relevant data which is retrieved also known as false trues. Fallout always is very small as the search task is skewed in nature and there are more non-relevant data items than relevant in the search space (corpus). Missout is a measure of relevant but non-retrieved data items.

For example, in a result set if there are 20 relevant data items and 14 out of 20 are retrieved. Given a total number of data items of 1000, then fallout would be 6/1000.

Mostly search engine effectiveness is measured on the basis of recall and precision to summarise results.

The evaluation concludes by providing an in-depth analysis of any unexpected results through detailed error analysis. This error analysis identifies the source of errors within the result set, which provides guidance for future work and improvements into the algorithm and future research.

6.7 Limitations of Evaluation Plan

Evaluation of the proposed novel opinion analysis approach and corpus design is proposed to be undertaken through the development of a proof of concept prototype system. The proof of concept implementation is just a prototype system; therefore the prototype can be improved and enhanced for further refined application of the proposed approach. Further refinement of the implemented prototype could result in enhancements to the results. Similarly the choice of the resources used can also be improved, and improved resources can be generated. The use of improved and refined resources can improve the effectiveness of the opinion analysis approach as the resulting corpus is heavily dependent on the available resources used.

A GS approach is proposed to be used in evaluation. A small dataset GS (in the proposed evaluation plan of 50 sentences) is annotated by experts. The choice of 50 sentences as a sub dataset is made based upon the recommendation made by Manning et al (2009), for the minimum size of any GS dataset to be fifty items. This process of generation of the GS relies on the experts providing a form of absolute knowledge to provide the benchmark. The proposed evaluation plan uses two experts in order to minimise the limitations raised by the process of manual annotation. The level of personal bias and inconsistencies introduced by the experts' personal information level and understanding of words can be reduced by having more than one expert annotator. However, there can be an argument that two is not an appropriate number. While reviewing the state-of-the-art it is observed Thet et al (2010) and Neviarouskaya et al (2011) have used two annotators, whereas Baccianella et al. (2010) has used five annotators. There is no best number of annotators identified in literature; therefore it has to be a balance. The five participants (non-experts) reviewing each sentence have helped with the validation of the system output, and there can be an argument about five being an appropriate number or not in relation to this evaluation.

The dataset is collected manually, which can raise the argument, that the manual selection of sentences can bring bias to the dataset. Therefore, the subset from the dataset for the generation of the GS is gathered automatically. This automatic selection of sentences for the GS helps to minimise the effect of biases introduced due to manual selection of the dataset.

Although precision, recall, and accuracy are the most widely used measures in the fields of IR, machine learning, and computational linguistics, they are believed to be biased measures (Powers, 2011). For example, the precision measure has been indicated to penalize system retrieval of irrelevant items (false positives) but not to penalize failures by the system to retrieve items that the user considers to be relevant (false negatives) (Alvarez, 2002). In addition, the recall measure penalizes false negatives, but not false positives (Alvarez, 2002). However, the relevance of sentences does not come under the scope of this research. The current research is based upon the assumption that the objective sentences do not contain sentiment expressions.

6.8 Summary

This chapter has presented the proposed process of evaluation for the novel opinion analysis approach and corpus design. This process concentrates on the development of a proof of concept implementation prototype to be used in order to meet specified evaluation goals. The chapter began with a presentation of the description and critique of the evaluation techniques used by other researchers. The next section presented the goals of evaluation process. Followed by a presentation of the evaluation plan and design respectively with a description about the datasets, participants, and details about each stage and phase. Finally, the chapter discussed the limitations of the evaluation plan.

Chapter 7 – Evaluation

The previous chapter presented a detailed evaluation plan in order to analyse the contributions of this Thesis. The evaluation plan included details about the evaluation goals; choice of datasets; participant sample; tools; and techniques used as well as the experimental test beds designed.

This chapter provides a detailed analysis of the results delivered as an outcome of the evaluation plan and discusses the impact of these results on assessment of the presented novel opinion analysis algorithm and corpus design. The chapter presents analysis of the results aligned with each evaluation goal and concludes with a thorough review of any unexpected results and outliers. Results from this chapter also help to establish the formulation of future work for this research in Chapter 8.

7.1 Evaluation Process

Table7-1 presents a short synopsis of the evaluation plan as detailed in Chapter 6. Evaluation goals are provided matched to the appropriate datasets, the focus of the analysis, the measurements to be determined, the number of participants involved and the tool used.

Table 7-1: Synopsis of Evaluation Plan

Goal Number	Goal	Dataset	Basis	Measurements	Participants	Test bed
E1	Determine system performance in relation to expert human performance	50 Sentence subset	Polarity orientation, Sentence (Subject, Object, Verb, opinion words, topics)	Recall, Precision, F-Score (opinion Seven scale, opinion Three scale, Subject, Verb, Object, Opinion Words, Opinion Topics), co relation	Two experts and system	Prototype P1
E2	Validate system performance (opinion identification) in relation to general human performance	600 Sentence dataset	Polarity orientation	Mean and Standard Deviation (Five users and systems), Percentage (Five scale agreement level)	150 participants and system	Prototype P1, web based data collection portal
E3	Determine if the system performance is improved by the introduction of phrase level opinion analysis	5 sentences	Structure (Subject, Object, Verb, opinion words, topics)	In-depth qualitative analysis	Two experts and system	Prototype P1 & P2
E4	Determine the applicability of the corpus in relation to IR	50 sentences subset	Sentence (topic-objects)	Recall, Precision, F Score , Missout, Fallout	Two expert systems and search engine	Prototype P1 & search engine

7.2 Analysis of Results

7.2.1 Analysis Based Upon Evaluation Goal 1

E1 determines system performance in relation to expert human performance, where the system performance covers both the opinion analysed, as well as the structural deconstruction of sentences (Subjects, Objects, Verbs, Opinion Words and Opinion

Topics). In order to achieve E1 a subset of 50 randomly selected sentences from the initial dataset is used, and human performance is benchmarked through the use of two expert GSs. The system output to be evaluated is generated by using prototype system P1, which utilises the novel opinion analysis algorithm using phrase level analysis for the determination of opinion orientation, and the structure of sentences, by syntactic analysis.

As explained in the Section 6.6, the evaluation of opinion orientation is performed on two scales. First, the system performance is evaluated for the opinion orientation and strength using the seven level likert scale values (Strongly Positive, Mildly Positive, Weakly Positive, Neutral, Weakly Negative, Mildly Negative, Strongly Negative) and second, the opinion polarity is evaluated with opinion polarity redefined over three values (Positive, Negative, Neutral).

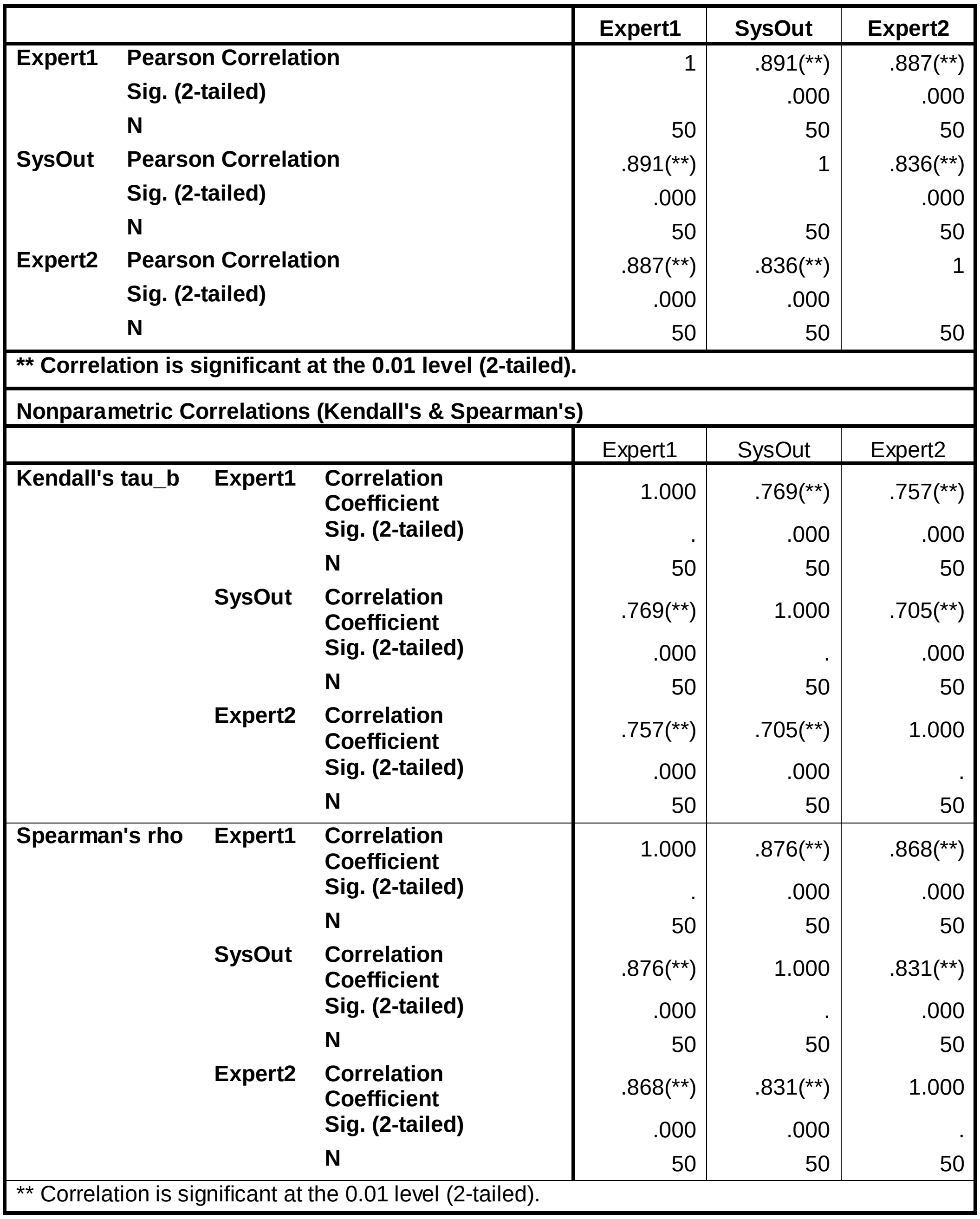
For the first scale, three values are analysed across each sentence in the dataset, i.e., the opinion orientation assigned by the system, opinion orientation in GS1 by expert 1, and opinion orientation in GS2 by expert 2.

7.2.1.1 Evaluation Goal (Opinion Orientation)

Correlation is calculated (see Table 7-2) among all three values in order to identify the level of closeness between the opinion orientation assigned by each of the experts and the system. For this purpose Pearson coefficient correlations, Spearman’s rho and Kendal's tau are used. The main reason behind using three different correlations is that Spearman’s rho and Kendal’s tau are less sensitive to non-normality of distributions. Whereas, Pearson correlation is good for linear relations.

Table 7-2: Correlation for opinion orientation: where Expert 1, SysOut, and Expert 2 are GS1, System (P1) output and GS2 respectively

Parametric Correlations (Pearson Correlation)



		Expert1	SysOut	Expert2
Expert1	Pearson Correlation	1	.891(**)	.887(**)
	Sig. (2-tailed)		.000	.000
	N	50	50	50
SysOut	Pearson Correlation	.891(**)	1	.836(**)
	Sig. (2-tailed)	.000		.000
	N	50	50	50
Expert2	Pearson Correlation	.887(**)	.836(**)	1
	Sig. (2-tailed)	.000	.000	
	N	50	50	50

** Correlation is significant at the 0.01 level (2-tailed).

Nonparametric Correlations (Kendall's & Spearman's)

			Expert1	SysOut	Expert2
Kendall's tau_b	Expert1	Correlation Coefficient	1.000	.769(**)	.757(**)
		Sig. (2-tailed)	.	.000	.000
		N	50	50	50
	SysOut	Correlation Coefficient	.769(**)	1.000	.705(**)
		Sig. (2-tailed)	.000	.	.000
		N	50	50	50
	Expert2	Correlation Coefficient	.757(**)	.705(**)	1.000
		Sig. (2-tailed)	.000	.000	.
		N	50	50	50
Spearman's rho	Expert1	Correlation Coefficient	1.000	.876(**)	.868(**)
		Sig. (2-tailed)	.	.000	.000
		N	50	50	50
	SysOut	Correlation Coefficient	.876(**)	1.000	.831(**)
		Sig. (2-tailed)	.000	.	.000
		N	50	50	50
	Expert2	Correlation Coefficient	.868(**)	.831(**)	1.000
		Sig. (2-tailed)	.000	.000	.
		N	50	50	50

** Correlation is significant at the 0.01 level (2-tailed).

Correlation values between 0.75 and 1 mean highly correlated values, this provides an understanding that the predictability of one dataset from another is higher, and the error of prediction is low. The Pearson's correlation between expert 1 (GS1) and expert 2 (GS2) is 0.887, Kendall's tau is 0.757 and Spearman's rho is 0.868, which shows a high correlation (demonstrating limited disagreement) and strong association between both GS(s). This level of correlation between the two GS for each of the measure establishes a benchmark for system performance in comparison to both GS(s). As highlighted in Table 7-2, system performance is evaluated based on two GS. In the comparison with GS1 system shows Pearson's correlation as 0.891, Kendall's tau is 0.769 and Spearman's rho is 0.876,in the comparison with GS2 system performance with Pearson's correlation as 0.836, Kendall's tau is 0.705 and Spearman's rho is 0.831. In both cases these values are highly correlated and in the case of GS1, the match between the system and expert performance, outperforms the benchmark value.

It is observed from Table 7-3 that in the current case there is no difference between the percentages reported for recall and precision, as the classification process results in a binary value.

When investigated in detail it is found that less than 50% of the values assigned by both experts are exact matches. The remaining values are highly correlated. This shows that expert human understanding across fine grained opinion based scoring does not result in an exact match based upon a range of annotator considerations (e.g. their background knowledge, understanding of different words etc.) (Bhowmick et al., 2008).

Table 7-3: Recall, Precision and f-score for opinion orientation over seven scales

	System P1 Generated Opinion Orientation over Seven Scale		
	Recall	**Precision**	**F-score**
GS1	44%	44%	44%
GS2	30%	30%	30%
The values for Recall, Precision and F-score between GS1 and GS2 =42%			

It is observed that the percentage of exact matches for classified opinion orientation between both experts is only 42%. This gives an insight to how two human experts can disagree in identifying opinion orientation on a fine scale, rather than just identifying the polarity of being positive or negative sentences. This also provides a benchmark for any automated system to demonstrate agreement of 42% or more. The performance of the prototype system P1 gives a 44% level of agreement with GS1, and 30% level of agreement with GS2.

7.2.1.2 Evaluation Goal 1 (Opinion Polarity)

When the opinion orientation is mapped to a basic scale of Positive, Negative and Neutral, then the percentage match between both GSs is 92%. The values for opinion polarity are generated by mapping the system result of 'Strongly Positive', 'Mildly Positive' and 'Weakly Positive', to 'Positive'; system results of 'Strongly Negative', 'Mildly Negative' and 'Weakly Negative', to 'Negative'; and the system result of 'Neutral' to 'Neutral'. The system results in comparison with GS1 and GS2, are shown in Table 7-4. Table7-4 exhibits a system performance of 86% and 82%, this is lower as compared to the benchmark performance established between both GSs (92%). Similar to earlier the values of recall and precision are merely the percentages of correctly annotated opinions. Therefore the f-score value does not give any further insight for analysis. The values of precision and recall are 86% and 82% for both GS1 and GS2 respectively, which shows high values, hence establishing good system P1 performance.

Table 7-4: Recall, Precision and f-score over the opinion polarity scale

	7.2.2 System P1 Generated Opinion Polarity over three scale		
	7.2.3 Recall	**7.2.4 Precision**	**7.2.5 F-score**
7.2.6 GS1	86%	86%	86%
7.2.7 GS2	82%	82%	82%
The values for Recall, Precision and F-score between GS1 and GS2 =92%			

7.2.7.1 Evaluation Goal 1 (Sentence Decomposition)

As discussed during the plan and later observed in the evaluation of polarity and the strength of opinion, three measures of precision, recall and f-score are calculated. High values of precision and recall i.e., close to 1.0 or 100% are considered better.

Table 7-5: The values of Recall, Precision and F-Score for matches in Subjects, Objects and Verbs in Sentences

	System P1 Generated Subjects			System P1 Generated Objects			System P1 Generated Verbs		
	Recall	Precision	F-Score	Recall	Precision	F-Score	Recall	Precision	F-Score
GS1	89%	92%	90%	74%	89%	81%	86%	88%	87%
GS2	90%	95%	93%	76%	95%	84%	87%	91%	89%

Table 7-6: The values of Recall, Precision and F-Score for matches in Opinion Words and Opinion Topics

	System P1 Generated Opinion words			System P1 Generated Opinion topics		
	Recall	Precision	F-Score	Recall	Precision	F-Score
GS1	60%	93%	73%	39%	98%	55%
GS2	72%	78%	75%	28%	88%	43%

A precision value 1.0 means every item labelled in a category ‘A’ actually does belong to ‘A’, and is correctly annotated. However the precision scores provide no information about words which actually belonged to one of the classes (e.g. ‘A’) but are wrongly classified into another. Therefore a recall measure is calculated. The recall value of 1.0 or 100% means that every word which (according to the GS(s)) belongs to a class ‘A’ is correctly classified in ‘A’, and is not wrongly classified into another class.

Often there is a trade-off between values of recall and precision, and improvement (getting close to 1.0/100%) in either of them can impact the other. However, this trade off does not always exist. In an ideal case there are systems which give 100% or 1.0 for both precision and recall, without giving any trade-off behaviour. Therefore another measure f-score is used. All these measures are discussed in detail in Chapter 6.

It is observed from Table7-5 that all the precision and recall values for the restructured sentences (Subjects, Objects and Verbs) show slightly higher values for GS2 as compared to GS1, i.e., both precision and recall values show results ranging between 74-92% for GS1 and 76-95% for GS2.

Both precision and recall values are based upon a measure of relevance. In the current research the overall relevance of data is controlled by the manual generation of a dataset. This manual generation means that only those sentences are captured which have close relevance with system requirements. Relevance in specific with precision and recall means that the annotation is correctly performed. So the annotation by the prototype system (P1), which is matched correctly with a GS is classified as relevant. The precision value of 92% for annotation of Subjects according to GS1, means that 92% of the words classified as 'Subjects' (by P1) are correctly classified in relation to the GS1 benchmark, and only 8 % of words annotated as subjects do not belong to the 'Subject' category. Similarly precision values of 89%, 88%, 93% and 98% in relation to GS1 for 'Objects', 'Verbs', 'Opinion Words' and 'Opinion Topics' means the classification of (Objects, Verbs, Opinion Words and Opinion Topics) performed by P1 is close to the GS.

The recall value of 89% for the annotation of 'Subjects' according to GS1, means that 89% of the words are correctly classified as 'Subjects', according to GS1, and only 11% are incorrectly classified into other categories than 'Subjects'. Recall values of 74%, 86%, 60%, and 39% respectively gives the percentage of for 'Objects', 'Verbs', 'Opinion Words' and 'Opinion Topics' in the dataset which are correctly classified with respect to GS1. It is observed for 'Opinion Word' and 'Opinion Topic' classification that the precision value is fairly good (i.e., as high as 93% and 98%) but, the recall is calculated very low (60% and 39%). This means that not all the words classified as 'Opinion Words' and their respective topics belong in the 'Opinion Words' and 'Opinion Topics' classes respectively. It is observed that the recall value for 'Opinion Words' according to GS2 is high i.e., 72%however the recall for 'Opinion Topics' reduces to 28%.

The above suggests that for the more standard classifications, syntactic analysis in the system (Parts of Sentence (Subjects, Objects and Verbs)) results in very good (high) precision and recall values (81%-95%). This improves with the benchmark GS2. The values of 'Opinion Words' and their respective 'Opinion Topics' are based more on personal understanding of the words, therefore the results for recall can be very low.

The calculations summarised in Table7-6, show an altogether different observation, system identified 'Opinion Words' in GS1 show a recall value of 60%, and precision value of 93%, this means that according to GS1 only 60% of the relevant opinion based phrases are identified by the system. Whereas for the retrieved 'Opinion Words' in relation to GS1, 93% are correctly identified by the system. This depicts the existence of trade-off between recall and precision values for ‘Opinion Words’.

Comparison of P1 with GS2 provides an average result of 72% for recall and 78% for precision values. The f-score values of both GSs are very close to each other 73% (GS1) and 75% (GS2). The results for the ‘Opinion Topic’ classification are very different and they show a clear inverse relationship of the recall and precision values. Recall shows only a 39% and 28% performance for GS1 and GS2 respectively. This recall score is very low and demonstrates a poor performance of the system in terms of identification of relevant ‘Opinion Topics’. Precision is calculated for both the GSs and demonstrates a precision value of 98% (GS1) and 88% (GS2). Precision scores for ‘Opinion Topics’ show that most identified ‘Opinion Topics’ are correctly identified by the system. The f-score value (55% for GS1 and 43% for GS2) shows a poor performance of the system for the identification of ‘Opinion Topics’.

7.2.8 Discussion

As detailed earlier the evaluation goal is carried out at two levels 1) opinion (polarity and strength) and 2) sentence structure. Opinion polarity between both the experts (GSs) is observed to have a high level of agreement i.e., 92%, which sets a benchmark performance between two human experts. The low level of disagreement shows the distinctive understanding of both experts for opinion based words and phrases and their usage in sentences. The system performance for opinion polarity when measured against GS1 and GS2 is 86% and 82% respectively, which is lower than the benchmark value.

When discussed with both experts in discussions and meetings the level of agreement raised to 100%. This 100% agreement of polarities is not considered as an agreed GS as it sacrifices the natural course of human understanding. The main reasons for disagreement between both the experts were found to be in the use of ambiguous terms, and the lack of agreement in contextual knowledge and understanding. In order to agree both experts had to make assumptions about the sentences. For example: “Before continuing one sentence further, am I aware that I sound like the world’s most spoiled rotten brat?”, Expert 1 classed it as ‘Weakly Negative’ whereas Expert 2 classified as ‘Neutral’ and the proposed system has assigned ‘Mildly Negative’ for its polarity.

While discussing this sentence Expert 2 justified the sentence to be ‘Neutral’, as the author is just presenting an assumption not any opinion about any particular opinion (topic).

A low level of agreement is observed for opinion polarity and strength over the seven scales measurement, as results show only 42% of exact matches between both GSs. This behaviour demonstrates a low level of agreement between experts in the understanding of polarity in relation to a finer level of granularity. This 42 % of agreement demonstrates a high level of variability in human understanding of opinion strength. However, the high level of correlation between both experts GS(s), gives a good benchmark, to prove that even if there are slight disagreements over a seven point opinion scale for polarity and strength, their opinions do not vary drastically, and the strong correlation expresses their movement together, i.e., the opinion polarity and strength do not exactly match, but are closely related which may mean if one assigns ‘Weakly Negative’ as an opinion then the other might have assigned ‘Mildly Negative’. For example: “Image quality was not as good as expected.” expert 1 gave a polarity and strength of ‘Mildly Negative’, whereas expert 2 gave ‘Weakly Negative’. Therefore the measure of exact match fails and this disagreement adds up to a low level of agreement. However, it is further explored that both experts agreed that the sentence is expressing a ‘Negative’ opinion. So, both experts agree on the polarity of opinion but the disagreement is about the strength of opinion expressed. Similar to the previous example, the use of recall, and precision only gives a percentage as the result, therefore f-score (harmonic mean) does not suggest anything different.

In order to analyse the results for the evaluation of the sentence decomposition, there is a need to establish a benchmark as achieved for opinion polarity and strength. Therefore the results for both GSs were compared with each other in order to establish an acceptable level of disagreement between system results with GSs. The measures of precision, recall and f-score are discussed in detail earlier in Chapter 6. From Table7-7 and Table 7-8 it is clear that the precision values for ‘Subjects’, for ‘Objects’, ‘Verbs’, ‘Opinion Words’ and ‘Opinion Topics’ is very high, i.e. closer to 100%. The values range between 89% and 100%, which means that most of the annotation labelled in each category are correctly annotated. Similar results are observed in Table7-5 and Table7-6 where precision is providing high results. It means in both the GSs if the system annotates anything into identified classes and labels it does it correctly and the chances of error (misclassification) are low ranging between 2% to 22%.

In the case of recall different trends are observed. The benchmark shows a high level of recall for ‘Subjects’, ‘Objects’ and ‘Verbs’ (syntactic and structural part of sentence), the results remain in range of 81% to 88 %. A similar trend is observed in Table7-5, where system (P1) shows recall values ranging from 74%- 90%.

It is observed that the recall value for ‘Objects’ is 74% and 76% for both GS1 and GS2 respectively, which means 24% to 26% of ‘Objects’ were either misclassified or are altogether missed during annotation. When further analysed it is observed that in the user generated content available online the structure of language is not perfect with authors tending to write incomplete sentences, missing objects, punctuation marks especially full stops. For example one of the sentences selected from the online forums is *“I am worried my 7yr old boy is bully at and they told me if he doens ' t stop it they are goin to expell him from , please help me how do i stop him from being bully ?”*. This sentence is incomplete; it has missing objects, spelling mistakes and missing punctuation marks. When such sentences which depict lack of structure and grammar are deconstructed, the correct identification and classification of POS, dependency objects, and opinion based words are difficult. While reading the sentence a human expert or even a general human participant can interpret and understand that the author might be writing about some institute (school) where a boy is bullying, however, for the automated system it is not difficult to display this level of intelligence and therefore the structural decomposition of the sentence does not match the benchmark.

Table 7-7: The values of Recall, Precision and F-Score for matches in Subjects, Objects and Verbs in Sentences between both GSs

	Subjects in GS2			**Objects in GS2**			**Verbs in GS2**		
	Recall	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**	**Recall**	**Precision**	**F-Score**
GS1	88%	100%	94%	81%	100%	89%	86%	89%	87%

Table 7-8: The values of Recall, Precision and F-Score for matches in Opinion Words and Opinion Topics between both GSs

	Opinion words in GS2			Opinion topics in GS2		
	Recall	Precision	F-Score	Recall	Precision	F-Score
GS1	55%	94%	69%	40%	98%	57%

As presented in Table 7-8, the recall values between both expert GSs based on 'Opinion Words' and 'Opinion Topics' are 55% and 40%, which again do not show very promising results, and is a clear indication of an inverse relation between precision and recall as the precision values are 94% (Opinion Words) and 98% (Opinion Topics). This demonstrates that there is very small chance that any opinion word or opinion topic is wrongly classified into 'Opinion Words' and 'Opinion Topics'. However, the low recall rate highlights missed or wrongly classified 'Opinion Words' and 'Opinion Topics' into other classes.

'Opinion Words' and 'Opinion Topics' are more semantic by nature and mainly rely on the context and understanding of the annotator. It is observed that after discussion between both experts, the value of recall improved significantly for both 'Opinion Words' and 'Opinion Topics', i.e., 88% and 84%, however, there were still disagreements, and the benchmark values are not changed based upon their discussions.

Table7-6 shows similar results for recall related to 'Opinion Words' in the case of GS1 and GS2 this is 60% and 72% respectively. P1 is a lexical based approach and only one resource (SentiWordNet (Esuli, 2008)) is used for the development of the prototype system. It is believed that with more refined resources, and an implementation of a form of WSD the results can improve. It is observed that SentiWordNet is based upon WordNet and has more than one sense for each word. These different senses are assigned with different opinion scores which sometimes show strong variation. For example: horror and praise are two simple words, when they are analysed in SentiWordNet 3.0 it is observed that based upon WordNet 'horror' is returned with three senses and out of them the first sense is weakly positive, whilst the second and third senses are mildly negative. Similarly the word 'praise' has two senses as a noun, where in SentiWordNet 3.0 both senses are neutral. The word 'praise' as a verb has only one sense and still shows a neutral opinion. When such words that have multiple senses in the resource are encountered during analysis, then the identification of the correct sense becomes complicated and the inability to identify and use the correct sense may lead to inconsistent opinion analysis.

Therefore the refinement of resources and implementation of WSD is very important.

The high correlation of P1 output with each of the expert annotations (GSs) as shown in Table7-2, shows a strong relationship between expert performance and the implementation of the proposed opinion analysis approach. High precision values for ‘Subjects’, ‘Objects’, ‘Verbs’, ‘Opinion Words’ and ’Opinion Topics’ show a higher ratio of correct annotation. The high recall for ‘Subjects’, ‘Objects’ and ‘Verbs’ classifications shows limited misclassifications and miss outs, which means the f-score value is also calculated to be a high score i.e., closer to 1. This demonstrates accurate deconstruction of sentences on the basis of syntactic and linguistic analysis. The low recall values for ‘Opinion Words’ and ’Opinion Topics’ have been discussed earlier and the need for better lexical resources and in-depth implementation techniques are highlighted.

Overall, precision and recall calculations for ‘Subjects’, ‘Objects’ and ‘Verbs’(syntactic decomposition of sentences) provide better results with GS2, however, calculations for ‘Opinion Words’ and ’Opinion Topics’ provide better results with GS1.

7.2.9 Analysis Based Upon Evaluation Goal 2

The second evaluation goal provides information regarding the validation of system performance in terms of the opinion analysed in relation to the non-expert human performance. For this purpose the complete dataset of 600 sentences is used. System performance is captured by obtaining the opinion orientation of all the 600 sentences in the dataset with the help of prototype P1.

For this purpose a two stage process is adopted. In the first stage, a Web based questionnaire is sent to 150 participants, so that each sentence and its opinion orientation is reviewed by 5 individual participants. Each participant is sent 20 questions to annotate. The participants are general non experts i.e., people with no particular restrictions about their background and with no training in opinion analysis or annotation. A brief description of the system and a set of examples are presented to them before presenting them with the questionnaire in order to give them some brief training and to provide them with an overview of what they are required to do including how their annotation is going to be used for the current research. The participants are presented with a sentence in an order and are provided with the seven scale measurement tool. They are asked to select the most appropriate opinion orientation for each sentence provided.

The system output and the level of closeness of the participant's results are evaluated. For this purpose the data is organised in a way that each sentence has six values of opinion associated to them, one is system P1 generated and five are assigned by five general human participants. The process explained in Chapter 6 is followed, first the mean for all non-expert human participants across each sentence is calculated, and SD for all five human assigned values for opinion orientation and strength is calculated. This process provides a value of mean and SD across each sentence for the five non-expert human values captured. The next step establishes if the P1 output for each sentence falls within a range of 1SD of the mean, i.e., within [Mean – 1SD, Mean + 1SD]. It is observed that out of the 600 sentences 541 P1 outputs fall within the acceptable range of [Mean – 1SD, Mean + 1SD].

Stage 2 is based on sending the results about opinion orientation and strength, from system P1 to the participants. The participants respond with the level of agreement or disagreement with an orientation value using a five point scale (Strongly Agree, Agree, Indifferent, Disagree, Strongly Disagree). During the data collection phase the participants were allowed not to respond to any question. This provides some missing values in the dataset captured at this stage (Stage 2). These missing values are only 1% of the dataset as presented in Table7-9.

There were 600 sentences and each sentence was sent to 5 respondents, therefore, there is a total of 3000 responses. The results are presented in Table7-9. Out of 3000 responses, 96 % have agreed with the system generated opinion whereas, almost 2.5 % have stayed indifferent or neutral about the response and less than 1 % have disagreed at any level.

Table 7-9: Percentage results for Stage 2

Total Responses	3000	%age
Strongly Agree	2096	69.87%
Agree	797	26.57%
Indifferent	74	2.47%
Disagree	3	0.10%
Strongly Disagree	2	0.07%
No Answer	28	0.93%

7.2.10 Discussion

In stage 1 of the second evaluation goal more than 90% of system responses fall within the range of 1SD of Mean. These results are extremely positive as they provide an understanding that the prototype system P1 is 90% as good as non-expert human understanding of opinions, from within the written text. In this sample set similar results were observed when the system was observed with both GSs in E1, where Table7-2 shows a 0.887 correlation between the GSs.

The results obtained at stage 2 are also interesting. It was observed that there was a high positive correlation of all participants with the system generated output. However, that high correlation is not as high as 96.44% (level of agreement as shown in Table7-9, 69.87% + 26.57%).

The results obtained as a result in this stage may be biased as it is observed that many people tend to agree more often when they are provided the likert scale of agreement and disagreement (Johns, 2010). It is also observed that respondents tend to give neutralising responses, i.e., selecting middle values on the likert scale, especially if the questionnaire demands a lot of thinking and decision making for responses (Johns, 2010; Boyer and Stron, 2012). The above observations about general human behaviour can be the reason that in online reviews neutral comments are thought to be negative (Liu, 2010).

As the questionnaire was sent to general human non-expert participants it requires a high level of analytical skills, requiring participants to spend quite a substantial amount of time in reading and understanding the sentences, many participants might have agreed more often, instead of challenging the system generated response. This might be the reason for such a high agreement rate. However, the system also has shown a high correlation between P1 responses and each of the participant responses at the first stage of the evaluation of E2. This shows that the opinion analysed by using the proposed opinion analysis approach is closely related to that of general human participants and shows good and acceptable results.

7.2.11 Analysis Based Upon Evaluation Goal 3

The third evaluation goal provides the information regarding the improvements introduced by using the phrase level structures into opinion analysis. For this purpose a small subset of the complete dataset is used. In order to provide an in-depth analysis of the system performance at phrase level structure, the opinion analysis, and sentence deconstruction using the prototype system P1 is used. Another prototype P2 as explained in Chapter 6 is also used for comparison, the second prototype P2 does not consider phrase level analysis and only uses word based opinion analysis and sentence decomposition. The outputs from both P1 and P2 are analysed in relation to both GSs: explained in Chapter 6.

When the analysed subset is examined in detail, it is observed that overall opinion analysis for both word based and phrase based analysis give exact results, when analysed for the whole of GS (50 sentences), it is observed that they remained constant. Therefore it can be stated that overall opinion analysis is not effected by using word based or phrase based analysis. However, this is not a conclusive result, as when further analysed in depth it is observed that as the resources used for development of both prototypes (P1 and P2) are only word based (SentiWordNet and WordNet), and both prototypes employed the same techniques for identification of opinion based words and modifiers (intensifiers, diminishes and negation words). Therefore both prototypes (P1 and P2) identify the same opinion based words, and handle intensifiers, diminishers and negations in the same way. There are some instances where opinion phrases are different than opinion based words. However, the opinion scores attached to those words do not present any significant difference and therefore the overall opinion (opinion orientation and strength) for the sentence is not affected. For example, one of the sentences has a phrase ‘especially low’ where ‘low’ is the opinionated word, but especially has no opinion value assigned by the resource therefore the phrase has made no difference for opinion analysis. Another example is ‘jaw dropping’ where P2 only has identified ‘dropping’ as an opinion based word, in both P1 and P2 this is identified as a negative opinion, as ‘jaw’ holds no value in the resource. So the lack of a phrase based resource does not allow prototype P1 to capture the actual meaning associated with the phrase’ jaw dropping’, as surprise and a positive opinion evaluation.

When analysed further for opinion based sentence deconstruction the results are different. First, the results for sentence decomposition from both prototypes P1 and P2 are captured and compared against each other. It is identified that for sentence structure deconstruction in terms of Subject, Object and Verb as well as Opinion Topics, there is a very noteworthy difference in both outputs when both the outputs from P1 and P2 are compared against both GSs. P1 output is closer to both GS performance and P2 output misses a large amount of information. For example in a sentence, Opinion Topic is identified as ‘Ford Interceptor Concept’ in P1, whereas in P2 only ‘concept’ is annotated as the Opinion Topic, giving only part of the overall topic.

7.2.12 Analysis Based Upon Evaluation Goal 4

The fourth evaluation goal provides the evaluation of the framework by determining the applicability and reusability of the resulting corpus. The framework itself employed the opinion analysis approach which is evaluated during the evaluation of E1, E2 and E3. The resulting corpus is evaluated for its application based upon its ability to effectively trace and retrieve information. Therefore a small search engine is implemented in order to evaluate the quality of information retrieved.

Search engines generally can be evaluated for their effectiveness and efficiency. Effectiveness is an ability of a search engine to find the right information i.e., relevant information with respect to query. Whereas efficiency measures how quickly the search is done. Efficiency is defined in terms of time and space. Sometimes it is argued that a search engine that is extremely fast is of no use unless it produces good results (Croft et al., 2009). In order to evaluate efficiency in search engine, there is a requirement for huge investment in processor, memory disk and networks. In addition to all these arguments the focus of the current evaluation goal (E4) is based upon IR and generally IR techniques focus on improving the effectiveness of search.

Earlier in the 1960s and 1970s large scale evaluation for search performance was performed, generally referred to as the Cranfield experiments (Voorhees, 2001). The test corpus has changed over years and more recently relatively smaller corpora are used as it becomes easier to manually identify the relevance of the retrieved data.

The current implementation of a search engine is performed over the test corpus generated during the evaluation of the opinion analysis approach using prototype P1 over the dataset of 50 sentences. The corpus is designed in such a way that it can be used as a data repository for a search engine or where data can be stored for later use. One major reason for using the 50 sentence dataset is that this dataset is annotated for topics (sentence topics) as well as Opinion Topics and the relevance can be calculated based upon the annotations.

The query based upon the corpus for measurement of effectiveness of the search engine is based on phrase combinations.

- Phrases combining opinion words with Subjects/Objects (opinion topics/opinion holders)
- Phrases combining opinion words with positive/negative (opinion orientation).

The measures used for this level are again precision, recall, fallout and missout (Egghe, 2004) as discussed in Chapter 6. Recall and precision are measures for the effectiveness of the search retrieval based upon the relevance of the retrieved data result of each query. Five queries are processed and based upon the results and topics originally assigned to the sentences values across precision, recall, fallout and missout are calculated as presented in Table7-10.

Table 7-10: Search results

	Query 1	Query 2	Query 3	Query 4	Query 5
Precision	0.857	0.909	1	1	1
Recall	0.75	1	1	1	1
Fallout	0.002	0	0	0	0
Missout	0.002	0.002	0	0	0

The values of precision and recall for all five queries show results ranging between 0.75 to 1.0, whereas the values of fallout and missout are very low ranging between 0 to 0.002.

High values of precision and recall between 0.75 to 1.0 show that most of the relevant results are retrieved and most of the retrieved results are correctly retrieved, which means the retrieval results are very good. Low values of fallout and missout ranging between 0 to 0.002shows that most of the non-relevant results are not missed, and none of the relevant results are missed. High values for recall and precision show good results for the information retrieved. This shows the applicability of the corpus as an information repository which can be extended on a larger scale. The scalability of the corpus is not evaluated in this current research.

7.3 Limitations

The evaluation process adopted to analyse the evaluation goals established in Chapter 6, may suffer some limitations based upon the nature of the research as well as the evaluation process adopted. Some of the limitations are already identified in Chapter 6 based upon the evaluation plan. The discussions with both the experts and some of the non-expert participants have brought forward the following limitations about the nature of the research, data and process into consideration.

Complexity of language

Language itself is a complex phenomenon, which has a set of words, rules and grammar, which continue to develop on an ongoing basis. It is difficult to capture the complete structure of any language, into a computer program, and/or linguistic or lexical resource. It becomes even more difficult to integrate the development of language into an automated system.

Human factor

The complexity of language is further enhanced with the use of user generated content on the Web. The Web is an uncontrolled world where users from all across the world generate content; some of them have English as their native language, whereas the majority of them are from non-English speaking countries, which can make the usage of the language less accurate and more complex. Further use of slang, emoticons and multi lingual discussions (bi lingual discussions); make the availability of words difficult in most of the lexical resources. The typo-graphical errors increase the complexity of the task. For example, one of the sentences in the dataset describing a camera is expressed as "*No, I am not talking about cheapo pocket digital cameras that everyone carries these days.*" Where cheapo is not a word in formal English and this brings in issues like informal and unstructured user generated content.

Expression of opinion

The identification and retrieval of opinion in written text becomes further complex; especially when care is not taken about the structure of language, or the author's cultural preferences or moods impact on their expression of opinion. The expression of multiple opinions in a sentence sometimes makes it difficult to interpret the opinion even for human beings. Sometimes, while reading a sentence it is not very clear whether a sentence is positive or negative, however the sentence may definitely be determined to be not neutral. Sentences can have mixed or inconsistent opinions expressed within the sentence, and there is the need for topic to be attached to the sentence in order to know whether the sentence is expressing a positive opinion about one or other object. A lack of knowledge about the context of written text, including the mood and cultural norms (the Web is a multicultural forum) of the author further enhance the difficulty.

Understanding of opinion

The interpretation of opinion at a fine grained level of Strong, Mild and Weak can also be subjective to personal experiences and the understanding of the context. Therefore, the evaluation against human performance can turn out to be inconsistent and sometimes even biased.

Size of Corpus

Experiments for IR based on the search criteria may differ by extending the corpus to a larger scale, as the scalability of the corpus is not evaluated in the current Thesis. The increase in the size of the corpus can also introduce issues in terms of the measurement of missout and prior assignment of opinion topics which also captures the relevant but not retrieved options.

Chapter 8 – Conclusion and Future Directions

The focus of this Thesis has been to examine the limitations of existing state-of-the-art approaches to opinion analysis with the intention of applying knowledge from other related research areas to the issues captured, in order to improve the process of automated opinion analysis. The growth in user generated data on the WWW has resulted in a situation (through information overload) where at present current search engines are not fit for purpose with regards to subjective information. Individuals and organisations cannot easily translate this plethora of content into knowledge that can be used to support organisational change processes through regular search technology.

In the context of this Thesis the spotlight has fallen on two particular challenges in relation to opinion analysis. The first of these has been understanding levels of granularity and their application in the opinion analysis process towards establishing whether any gaps remain in this area. The second challenge has been the investigation of current approaches to corpus design in order to determine whether there are any alternative approaches which may involve greater automation.

It was posited in Chapter 1 of this Thesis that phrase level analysis within the context of existing opinion analysis approaches was an area which required further investigation (Rill et al., 2012b). This position was formed as it is observed that the majority of existing approaches in opinion analysis focus primarily on word based lexical resources. Phrase level analysis offers opportunities to understand the interaction between words within sentences. This is important because the interaction between words in sentences can significantly impact the meanings of words and as a result the sentences themselves (Marneffe et al., 2006; Tan et al., 2011c). Exploring phrase structures within sentences gives us the opportunity to examine this impact and how this impact can be used to better inform the opinion analysis process.

Given the motivation captured above in Chapter 1 two research questions were formulated. These were:

- Are there improvements targeted at phrase level that can be made to existing state-of-the-art systems that can bring the process of automated opinion analysis closer to manual 'expert' performance levels?

- Does phrase level analysis provide opportunities for the identification of additional information that can be used to support opinion analysis?

The initial stage of answering the above questions involved gaining a significant appreciation of the historical development of approaches in the opinion analysis area. This fulfilled O1 and O2 identified in Section 1.3. Identifying the overuse of BoW approaches in this context. This overuse has resulted in a large body of research being constructed around the generation of word based resources (including lists) (Hatzivassiloglou and Mckeown, 1997; Strapparava and Valitutti, 2004; Subrahmanian and Reforgiato, 2008; Abdelrahman and Moustafa, 2010; Baccianella et al., 2010; Li and Wu, 2010). Much of this research has been constructed independently resulting in many pockets of word based resources for use in multiple different contexts. Often the challenge for researchers has been to improve the efficiency of the resource generation process or to improve the completeness of the resources themselves. Unfortunately, this activity has been completed generally in a context of a lack of sharing of resources until recently with the production of WordNet and SentiWordNet.

Approaches to opinion analysis do not sit in a vacuum; improvements have been made in other related research areas (for example NLP, computational linguistics and IR) that could transform existing opinion analysis processes. As an example the Stanford Parser and OpenNLP that have been constructed in the area of NLP can be used to better understand word based dependency structures within sentences (Www.Stanford.Edu, n.d; Apachi.Org, n.d.). However, phrase level analysis does not form a part of most of the accepted research (Stanford Parser and other resources) discussed in Section 2.8.4. It is suggested in this Thesis that whilst there are limitations linked to phrase level analysis resources (L2, L6 and L8 in Section 2.10) produced from other research areas, it can be used to better understand the interrelationship between words in sentences.

A small number of existing approaches exist in relation to the use of phrase level analysis in opinion analysis processes (Marneffe et al., 2006; Takamura et al., 2007; Agarwal et al., 2009; Rill et al., 2012b). However; these approaches currently suffer from a series of limitations. The major limitations are detailed in Chapter 2 achieving O2 in Section 1.3.

Taking into account the limitations of the state-of-the-art with existing opinion analysis systems (including those utilising phrase level analysis) there is a need to continue to formulate innovative approaches to improve this process. There is scope for bringing improvements from other related research areas into the area of opinion analysis in order to help make further improvements. The researcher in this Thesis has taken as a basis these concerns and opportunities in order to explore answers to the research questions posed.

In relation to research question 1 the approach taken has been to explore the proposal of a novel opinion analysis approach presented as an objective O3 of research in Section 1.3. This approach builds on the existing state-of-the-art regarding phrase level analysis and extends these systems through original thinking. The emphasis of this approach has focused on dealing directly with: the challenge of identifying/de-constructing phrases within textual data; limited automation; and the lack of phrase level resource production.

The novel approach for opinion analysis, takes care of the identification of opinion based words (lexical analysis) and the use of these words into syntactic constructs: phrases, and clauses (syntactic analysis). The opinion analysis approach utilises NLP and computational linguistics techniques in order to interpret opinion and its related target in a clause. The opinion is analysed at a finely grained level and opinion scores are mapped onto one of seven values (Strongly Positive, Mildly Positive, Weakly Positive, Neutral, Weakly Negative, Mildly Negative, Strongly Negative) these help to handle the limitations highlighted in L8 in Section 2.10. The proposed approach is unique in its analysis which utilises the structural hierarchy of sentences, as it determines opinion at a word level, based upon the lexical resource used, and aggregates the opinion based upon phrases utilising the constituent structure of the clause/sentence, and identifies the relationships between Opinion Words and Opinion Topics (targets) based upon the dependency structures. The approach provides steps towards improvements in relation to limitations L2, L3, L5 and L6 as presented in Section 2.10. The opinion analysed is presented into a frame of (Opinion Words, Opinion Topic , Opinion Polarity and Strength) based upon the sentence structure of Subject, Verb and Object, which expands on L2 and L7 in Section 2.10. The English Language sentence structure of Subject, Verb and Object help in current opinion analysis as Subjects and Objects are identified as Noun Phrases whereas Verbs are analysed into Verb Phrases which at times are nested into other phrase structures i.e., Adverbial Phrases, making sentence structure more complex.

Based upon objective O5 and O6, the evaluation of the opinion analysis approach gives an insight into the closeness of the performance of the developed prototype system, with the performance of experts and non-expert humans in terms of analysing and understanding the opinion communicated in written text as suggested in the second half of research question 1. Similar to observations in Section 6.1 developed within the literature review it was observed in this study that observations regarding how individual annotators (both expert and non-expert) deconstruct and interpret sentences is variable. Thought was placed into the selection of experts for the development of a GS. Thought was also placed into how these expert observations would be used in the context of the testing given the apparent variability and initial issues regarding resolving the differences between the GSs through conversation. A decision was made to use two GSs rather than compromising either or both based upon analysis of state-of-the-art.

The level of disagreement experienced between GS annotations was highly variable when experts were annotating the semantic roles (Opinion -Opinion Topic) over the structure of a sentence. The structure of a sentence can be thought to be more standard and therefore easily defined. Therefore both experts seem to agree more on the identification of Subjects, Verbs and Objects within sentences. In identifying the opinion and opinion strength, the words communicating these aspects are based upon an individual's understanding of the meaning, coupled with their interpretation of opinion strength. This results in a significant level of disagreement between both the expert annotations for the identification of Opinion Words, their respective Opinion Topics and in the granularity of an association of opinion score. However the generation of GS is the most accepted way used for evaluation in the fields of opinion analysis, NLP, IR and corpus/resource generation. In order to minimise the effects of GS approach as discussed in Section 6.1.4 and summarised in Table 6-1, a simulation/prototype approach is also used.

It is observed that there is a slight difference between the results of both expert annotations (GSs). The Subject, Object and Verb identification gave a high level of precision and recall for both GSs. However the results for GS2 show slightly higher values than that of the results of GS1. The results of recall for 'Opinion Words' and

‘Opinion Topics’ are not high. Especially in analysis of the Opinion Topics the recall values are very low which means that the system has many Opinion Topics, which are not annotated as Opinion Topics, or are missed, or are annotated under another class. Whereas in the analysis of the results for precision the system shows high values for both GSs for ‘Opinion Topics’, but the precision values for Opinion Words show contrasting results. For GS1, the precision value for Opinion Words is very high, but GS2 shows lower values. This contrasting behaviour depicts the difference in understanding of opinion and sentence structures even in the case of experts.

Whilst the significant issue of human variations in performance is highlighted above, it is also useful to critique evaluation instrument performance in Section 6.1.4. In this area we can talk about concerns regarding the prototype including the experimental set-up, in addition, to talk about issues regarding the novel opinion analysis approach. In relation to the experimental set-up it is observed that any change in the dataset used during evaluation or any changes in the GSs established as a benchmark can change the evaluation results. The prototype system is primarily based on only one lexical resource (SentiWordNet) although more could potentially be utilised. The proposed opinion analysis approach utilises lexical resource(s). Therefore there is a need for a very wide range of resources (complete vocabulary to be encountered during analysis) otherwise the opinion could be determined to be biased with respect to OOV terms. Similarly there is a need to consider WSD, where words are present with multiple senses (SentiWordNet). SentiWordNet is a resource developed in extension to WordNet, which captures multiple senses across each word, therefore SentiWordNet assigns a different opinion orientation and scores to each sense. This raises the need to use the WSD in order to improve the identification of opinion based words, their respective topics and overall opinion analysis process.

Taking into account the novel opinion analysis approach explained in Chapter 3 and the further extension of the approach with the identification of an extended framework and corpus design in Chapter 5, it has been posited that there is further information that can be gathered from the use of phrase level analysis. This provides evidence in relation to answering the second research question posed within this Thesis.

In analysis of the literature it is observed that opinion analysis systems generally use more than one task to represent opinion analysis into a system and/or process. Therefore generally systems utilize two or more of IR, opinion analysis, opinion summarization, opinion visualization, opinion regeneration etc., together within the contexts of delivered applications (Jin et al., 2009; Torres-Moreno et al., 2009a; Lloret et al., 2012). In order to give the process of opinion analysis a unified approach and promote the standardization and reutilization of the research for further use the proposed opinion analysis technique is employed into a framework. This framework is proposed as a unified process for opinion analysis. The framework uses IR, opinion analysis, opinion aggregation based upon the opinion topic, and saves the output into a corpus. The evaluation of the opinion analysis framework and its integration with proposed opinion analysis approach achieves O4 as presented in Section 1.3. It is completed through the evaluation of the novel opinion analysis approach and the original corpus design.

Phrase level opinion analysis, in addition, to the structure of the opinion used in this research (opinion - opinion topic) gives an opportunity to aggregate opinions based upon the Opinion Topics. This opinion aggregation process is only utilised at a basic level of WordNet relations between Opinion Topics, and rules based upon conjunction words between clauses within a sentence. However, there is scope for detailed opinion topic analysis to be used in the construction of a phrase based resource for domain specific topics (products). This potential aggregation falls out of the scope of Thesis and is not evaluated explicitly. However reusability of the framework and its output is assessed in order to achieve O8 from Section 1.3.

The output of the framework is a corpus. This corpus is designed based upon the requirement of the reusability of the analysed data generated as result of opinion analysis. Most of the other corpora in the field of opinion analysis are manually annotated which brings in issues related to manual annotation discussed in Section 4.2 and Section6.1.4. Many of the existing corpora are used only for the training of systems based upon machine learning approaches. The corpora resulting from research in opinion analysis are not designed with the flexibility to be utilized in other areas of research e.g., IR etc.

The proposed design of corpus is novel as it is generated as a result of an automated opinion analysis approach, which can easily be appended and therefore can be extended. The corpus is tested on its utilization and reusability into an IR system (search engine) and therefore is determined to be flexible and automated (E4 in Section 6.2). The provision of the automated corpus is a step towards overcoming L1 in Section 2.10.

As established earlier opinion analysis is a very active area of research and there is a large amount of research currently ongoing, opening new opportunities for extensions to current research. It is observed that some of the recent work in the area of opinion analysis is closely related to research presented in the current Thesis (Sykora et al., 2013a; Yadav et al., 2013) adding strength to the presence of the initial research question and later findings.

8.1 Closely Related Recent Research

EMOTIVE (Sykora et al., 2013a; Sykora et al., 2013b) means Extracting the Meaning of Terse Information in a Geo-Visualisation of Emotion. The group of researchers working on the project have experience in diverse backgrounds of knowledge management, IR, Computer Science, text mining, linguistics and discourse analysis.

EMOTIVE focuses on monitoring fine-grained emotional responses relating to events of importance for national security. EMOTIVE features in three areas: emotions; geo location; and filtering phrases. EMOTIVE is based upon a fine grained representation of emotions and does not rely on opinion centric resources. It proposes a lexical based approach therefore adopts an ontology engineering approach. Ontologies are rules based databases. On the basis of an extensive research process, a more complex and closer to natural cognition process structure for emotion is adopted for EMOTIVE. Sykora et al. (2013) uses a set of eight emotions, i.e. six Ekman's emotion (Ekman, 1985): Anger, Disgust, Fear, Happiness, Sadness, Surprise +Shame, and Confusion. An NLP pipeline is proposed to clean and pre-process the data as user generated textual data does not follow the rules of natural language. They explored a number of lexical resources and dictionaries and employed them into the process of the generation of a strong ontology based resource for emotions.

Another feature of EMOTIVE is geo location of hotspots, as EMOTIVE is mainly developed to patrol emotionally charged Web traffic, it identifies and detects key phrases which show high levels of negative emotions (Anger, Disgust etc.) and highlights geo-hotspots in communication, which can raise an alert regarding any unwanted accident, fight, or issue. Later all this information is visualised on an interactive system in order to provide an easy way to interpret and interact with the system.

Recent research by Thelwall et al. (2012), provides an improvement on their initial SentiStrength research (Thelwall et al. 2011). SentiStrength is a lexicon-based classifier which also uses linguistic information and rules to detect sentiment and displays strength in analysing informal English text. The SentiStrength output gives two scores, a positive and a negative score for each text. Both scores have values between 1-5, where 1 represents no sentiment and 5 represents strongest sentiment. The neutral text is depicted as 1, 1.

8.1.1 Similarities between Approaches Used

The similarities of the current Thesis with EMOTIVE can only be drawn on levels of emotion detection and phrase identification. The use of geo-location within EMOTIVE has no comparison with the research presented within this Thesis. The structure of emotion and background research of EMOTIVE is very different from that of the research in this Thesis; this is explored further in the next Section 8.1.2. The emotion extraction and interpretation part of EMOTIVE uses a similar approach to that of the proposed approach in the current Thesis. Both approaches use linguistic based opinion analysis approaches using lexical resources. Both emphasise the requirement to go to a level of refinement beyond word based granularity and have used phrases, i.e. EMOTIVE proposes as a key feature 'filtering key phrases'.

Both approaches establish the fact that the user generated textual data available online is not very structurally constructed and does not follow the rules for Natural Language. Therefore both approaches introduced a stage of pre-processing after retrieval of data in order to remove basic syntactic errors and spellings mistakes. Even in the evaluation stage both use the f-score measure (based upon precision and recall).

Both, the approach presented in this Thesis and SentiStrength added pre-processing features like spell checking and sentiment intensifiers, as well as having paid special consideration to negation handling. In addition, both have used linguistic based information to identify sentiment and its strength. Both approaches provide sentiment strength over a wider range of values rather than just providing polarity.

Both have used the Gold Standard approach in the generation of benchmarks for evaluation.

8.1.2 Differences between Approaches Used

In spite of similarities in the overall process followed for emotion extraction and analysis by EMOTIVE and the opinion analysis approach proposed in the Thesis, there are many differences in the way both are carried-out and implemented. One of the main differences between both research projects are, the way they have interpreted opinions and emotions. Sykora et al. (2013) have adopted a complex emotion based classification. They do not agree with simplified opinion based classification (positive, negative and neutral). They argue that simple classification of opinion/emotion over positive and negative does not fulfil the requirements of emotion analysis as emotion is a complex phenomenon.

The review of the state-of-the-art in the current Thesis, while analysing the structure of opinion, established that the classification of an opinion into categories other than (positive, negative and neutral) as those pursued by (Neviarouskaya et al., 2007; Neviarouskaya et al., 2009; Neviarouskaya et al., 2011) etc., makes the classification task further complex. It is further drawn from Neviarouskaya et al. (2011) that all these categories can simply be mapped onto positive/negative classes. Further to this there are categories in emotion classifications which are not very straightforward detailing complex emotions, i.e., depression; human interpretation of emotion at a refined level in order to differentiate ‘happiness’ from ‘surprise’. Standardising this process across human interpretation is difficult, if not impossible.

The structure of opinion for the current Thesis is based upon the comparison with the English sentence structure in Natural Language form. However, the structure of an opinion is independent of the rest of the analysis technique, therefore an emotion based structure can easily be employed with the same analysis approach, as emotions can also be expressed and analysed in the flow of Natural Language (phrase structures).

The proposed opinion analysis approach in the current Thesis does not emphasise an in-depth implementation. One of the major strengths of EMOTIVE is the extensive research in existing resources available online (https://sites.google.com/site/lboroemotive/resources/sentiment-analysis). Sykora et al. (2013) proposed an ontology engineering technique and employed a large number of diverse set of resources, listed on (https://sites.google.com/site/lboroemotive/resources/sentiment-analysis) in order to engineer the emotion based ontology as a lexical resource to be utilised during analysis. One of the main limitations of the evaluation of proposed opinion analysis technique presented in Thesis is the use of only one lexical resource. The absence of even a single opinion based word in the resource used can lead to a neutral opinion assignment, which can be an inaccurate analysis of opinion. This inaccuracy is purely on the basis of incomplete resource and hence inaccurate implementation, and has nothing to do with inaccuracy within the opinion analysis approach.

EMOTIVE only closely relates to the current Thesis in terms of opinion analysis, evaluation of opinion analysed, and the extraction of opinion based phrases. Other features of EMOTIVE i.e., geo location and emphasis on implementation (implementation details) are not comparable as they do not fall in the scope of current Thesis.

The research by Sykora et al. (2013) provides good variations to extend and build upon current research. Use of psychological emotion based models to redefine the structure of opinion/emotion and the use of more complex, complete, and rich lexical resources may contribute in terms of changes (improvements) in the results. However the emphasis of the work of Sykora et al. 2013 for lexical and linguistic based analysis and in considering phrases as a unit of analysis supports the contributions of the current Thesis.

SentiStrength mainly is based on list based lexicon resources, like, lists of negation words, lists of intensifiers, lists of idioms, whereas the approach presented in this Thesis is not dependent on any particular resource. It is an approach to calculate opinion, and its implementation is based on dictionary based resources.

The use of linguistic information for SentiStrength is mainly rule based which uses manually annotated data for training purposes, whereas, the linguistic information in the approach in this Thesis is generated and calculated at run-time and it heavily relies on a more standard dependency parsing system.

SentiStrength is a closely related classifier of sentiment polarity and strength providing a resource dependent lexical approach for sentiment analysis, which strengthens the research question and requirement of work presented in the Thesis.

8.2 Limitations of Current Research

In Chapter 3, 5, 6 and 7, the chapters conclude with discussion of the limitations for the opinion analysis approach, proposed framework, evaluation plan and process. In the current section some of the overall limitations in the current research are presented.

- The structure of opinion presented in Chapter 3 can be further analysed and improved with further research in areas of psychology and cognition in discourse analysis. This in-depth research can improved through a more comprehensive range of opinion/emotions and can improve the quality of the opinion analysed in the written text which, hence can improve the overall quality of opinion in opinion analysis.
- The textual analysis in written text is very complex task and research is being pursued in different areas based upon different utilities, i.e., NLP, WSD, noun co-referencing, resource development in terms of corpora and dictionaries, relationship finding within and between written texts etc. All this research has to be brought together in order to improve the quality of opinion analysis. Therefore the research in opinion analysis cannot be done in isolation. If all these research areas would have been studied in detail the overall opinion analysis approach could have been improved.
- The current opinion analysis approach is based upon lexical analysis which is preliminary performed on a basis of words, and later the scores are aggregated based on the phrase structures using the equations presented in Chapter 3. The sentence structure of Subject, Verb and Object is used. This approach can be improved by assigning the thematic roles to words while parsing the sentences.

i.e.: an Agent (corresponding to a Subject); a Theme (an object that has a particular location); Recipient (the person receiving the theme) etc. (Harley, 2007).

- Existing resources are used for the implementation of the approach at the evaluation stage. Online user generated textual content contains slang, short-hand text, incorrect spellings, inconsistent punctuations, emotions and many other OOV words. Therefore the lexical resources used for opinion analysis might not be good enough for opinion analysis. This limitation is observed in the evaluation of E1, where both precision and recall scores for ‘Opinion Words’ are not very high for both GSs.
- The evaluation phase can extend into the evaluation stage for the framework. This can improve the results for overall opinion analysis.
- An increased size of dataset for E3 and corpus for E4 (in Section 6.2) can improve the evaluation results and give a better insight into results.
- The evaluation of the opinion analysis approach was mainly dependent on the implementation prototype. The opinion analysis approach is mainly based upon the opinion identified through lexical analysis, which depends upon the quality of the lexical resource. There is a need to improve the quality of the lexical resource, either by using multiple resources or by generating a new resource. The use of an improved lexical resource can improve the opinion analysed.
- As evaluation is based upon the details implemented in the prototype system. The improvement in heuristic rules, resources and integration of other areas can give an improvement in the overall evaluation process.

8.3 Future Directions

Opinion analysis is a relatively new area of research and is showing a high level of activity in research. As more research is generated, it raises new questions and opens up new areas for exploration. Similarly the research presented in this Thesis opens some avenues for future work in the area that can be built upon the research conducted.

8.3.1 Level of Granularity

As identified in the literature within Chapter 2, there can be multiple levels of granularities, in terms of text (document, sentence, clauses, phrases, and words) and opinion (positive/negative, positive/negative/neutral, anger/disgust/fear/happiness/… etc.).

- The proposed opinion analysis approach can be applied for other levels of opinion classification and structure, just by redefining the opinion and resources for lexical analysis, For example see the work by Sykora et al. (2013).
- An implementation using the proposed approach can be extended by evaluating the approach with pattern based heuristic rules, which can also extend the opinion categories.
- The proposed opinion analysis approach extracted opinion words and opinion topics. The analysis of opinion topics is used in the aggregation of opinion at sentence level. The extraction of opinion topics can be further analysed for cross sentence boundaries in opinion aggregation.
- The proposed opinion analysis approach can be improved by developing the lexical resources especially resources utilising rules for WSD as these inform polarity calculation for the words encountered during analysis. Improvements in contextual analysis can also help in analysing irony and sarcasm.

8.3.2 Utilise the Diversity of User Generated Data

User generated data is very diverse by nature. This diversity is based upon the geo-location, ethnicity and culture of authors, the native language of authors (non-English speaking background, can affect the quality of language used), and the use of multiple languages etc.

- There is a need to use multi-lingual opinion analysis as mostly non-English speaking people use multiple languages while writing in SNS (social networking sites), blogs or forums (informal communication). The use of multi-lingual opinion analysis can help in reducing the number of unidentified lexical units (words) in resources and may improve overall opinion analysis.

- The opinion analysis approach can be extended for other languages by considering the structures of other languages and utilising resources for other languages.

8.3.3 Utilisation and Evaluation

The proposed opinion analysis approach is mainly used for review based data. There can be other areas of research where the proposed approach and corpus can be utilised.

- The opinion analysis approach and corpora can be used and extended for online patrolling systems, automated customer services, and online opinion based semantic search etc.
- The opinion analysis approach and resulting corpora can be employed for user profiling systems which can be used for e-learning, customer profiling and online recommender systems, etc.
- The proposed opinion analysis approach can be evaluated with different resources (ontologies, corpora) and datasets (more complex sentence structures). This re-evaluation can determine any further limitations of the approach.
- The corpus design is only evaluated for effectiveness, the efficiency of the design and utility can also be evaluated. This would help in understanding how to extend the corpus on a larger scale. It may also be implemented in the context of an opinion based search engine and/or web spidering technology.

References

Abbasi, A. (2007) Affect Intensity Analysis of Dark Web Forums. In: Intelligence and Security Informatics, 2007 IEEE. 282-288.

Abbasi, A., Chen, H. & Salem, A. (2008) Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Trans. Inf. Syst., Vol. 26, 3, pp. 1-34.

The Arabic Langauge Technology Center ALTEC (2010) The Pre-SWOT Analysis Opinion Mining.

Agarwal, A. & Bhattacharyya, P. (2005) Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified. In: The International Conference on Natural Language Processing.

Agarwal, A., Biadsy, F. & Mckeown, K. R. (2009) Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams. In: The 12th Conference of the European Chapter of the Association for Computational Linguistics. Athens, Greece. Association for Computational Linguistics,

Airoldi, E., Bai, X. & Padman, R. (2006) Markov blankets and meta-heuristic search: Sentiment extraction from unstructured text. Advances in Web Mining and Web Usage Analysis. Carnegie Mellon University, Pittsburgh, PA USA, School of Computer Science, School of Public Policy and Management.

Akkaya, C., Wiebe, J., Conrad, A. & Mihalcea, R. (2011) Improving the impact of subjectivity word sense disambiguation on contextual opinion analysis. In: The Fifteenth Conference on Computational Natural Language Learning. Portland, Oregon. Association for Computational Linguistics, 87-96

Alvarez, S. A. (2002) An exact analytical relation among recall, precision, and classification accuracy in information retrieval.

Apachi.Org (n.d.) Apache OpenNLP [WWW]. Available from: http://opennlp.apache.org/documentation.html [Accessed November 26, 2013].

Athar, A. (2011) Sentiment analysis of citations using sentence structure-based features. In: The ACL 2011 Student Session. Portland, Oregon. Association for Computational Linguistics, 81-87.

Atkins, S., Clear, J. & Ostler, N. (1992) Corpus Design Criteria. Literary and Linguistic Computing, Vol. 7, 1, pp. 1-16.

Atkinson, R. & Flint, J. (2004) Snowball Sampling. The SAGE Encyclopedia of Social Science Research Methods. SAGE.

Baccianella, S., Esuli, A. & Sebastiani, F. (2010) SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: The Seventh conference on International Language Resources and Evaluation LREC'10. Valletta, Malta. European Language Resources Association ELRA (May 2010)

Bai, X., Padman, R. & Airoldi, E. (2004) Sentiment Extraction from Unstructured Text using Tabu Search-Enhanced Markov Blanket. In: 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA.

Baker, C. F., Fillmore, C. J. & Lowe, J. B. (1998) The Berkeley FrameNet Project. In: The 17th international conference on Computational linguistics - Volume 1. Montreal, Quebec, Canada. Association for Computational Linguistics, 86-90.

Balahur, A., Hermida, J. M. & Montoyo, A. (2012) Building and Exploiting EmotiNet, a Knowledge Base for Emotion Detection Based on the Appraisal Theory Model. IEEE Transactions on Affective Computing Vol. 3, 1, pp. 88-101.

Balahur, A., Steinberger, R., Goot, E. V. D., Pouliquen, B. & Kabadjov, M. (2009) Opinion Mining on Newspaper Quotations. In: The 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology IEEE Computer Society, 523-526.

Balog, K. & Rijke, M. D. (2007) How to Overcome Tiredness: Estimating Topic-Mood Associations. In: International Conference on Weblogs and Social Media. Boulder, Colorado, U.S.A. 199-202.

Baxendale, P. B. (1958) Machine-made index for technical literature: an experiment. IBM Journal of Research and Development, Vol. 2, 4, pp. 354-361.

Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D. & Subrahmanian, V. S. (2007) Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In: The International Conference on Weblogs and Social Media (ICWSM). Boulder, Colorado, U.S.A.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V. & Jurafsky, D. (2005) Extracting Opinion Propositions and Opinion Holders using Syntactic and Lexical Cues. Computing attitude and affect in text: Theory and applications Vol. 20, pp. 125-141.

Bhattacharyya, D., Das, P., Mitra, K., Ganguly, D., Swarnendumukherjee, Samir, K. B. & Kim, T.-H. (2009a) A Novel Approach for Refinement of Corpus in the Field of Opinion Mining. In: The 2009 International Conference on Future Networks. IEEE Computer Society, 281-285.

Bhattacharyya, D., Das, P., Mitra, K., Mukherjee, S., Ganguly, D., Bandyopadhyay, S. K. & Kim, T.-H. (2009b) Refine Crude Corpus for Opinion Mining. In: The 2009 First International Conference on Computational Intelligence, Communication Systems and Networks. Washington, DC, USA. IEEE Computer Society, 17-22.

Bhattacharyya, D., Mitra, K., Choi, M. & Robles, R. J. (2009c) An Approach of XML-ifying the Crude Corpus in the Field of Opinion Mining. Distributed Computing, Vol. 2, 3, pp. 13-24.

Bhowmick, P. K., Mitra, P. & Basu, A. (2008) An agreement measure for determining inter-annotator reliability of human judgements on affective text. In: The Workshop on Human Judgements in Computational Linguistics. Manchester, United Kingdom. Association for Computational Linguistics, 58-65.

Bhuiyan, T., Xu, Y. & Jøsang, A. (2009) State-of-the-Art Review on Opinion Mining from Online Customers ’ Feedback. In: The 9th Asia-Pacific Complex Systems Conference. Tokyo. 385-390.

Binali, H., Potdar, V. & Wu, C. (2009) A state of the art opinion mining and its application domains. In: The 2009 IEEE International Conference on Industrial Technology. Gippsland, Australia. IEEE Computer Society, 1-6

Bird, S., Klein, E. & Loper, E. (2009) Analysing Sentence Structure. 1st edition. O'Reilly Media.

Bloom, K. (2011) Sentiment Analysis Based On Appraisal Theory And Functional Local Grammars. Doctor of Philosophy in Computer Science, Graduate College of the Illinois Institute of Technology.

Boiy, E. & Moens, M. F. (2009) A machine learning approach to sentiment analysis in multilingual Web texts. Journal Information Retrieval, Vol. 12, 5, pp. 526-558.

Bollen, J., Mao, H. & Zeng, X. (2011) Twitter mood predicts the stock market. Journal of Computational Science, Vol. 2, 1, pp. 1-8.

Boyer, S. & Stron, M. (2012) Best Practices for Improving Survey Participation An Oracle Best Practice Guide. Oracle.

Branthwaite, A. & Patterson, S. (2011) The power of qualitative research in the era of social media. Qualitative Market Research: An International Journal, Vol. 14, 4, pp. 430 - 440.

Brill, E. (1992) A Simple Rule-Based Part Of Speech Tagger. In: The third conference on Applied natural language processing. Trento, Italy. Association for Computational Linguistics, 152-155.

Brown, G. I. (2011a) An error analysis of relation extraction in social media documents. In: The ACL 2011 Student Session. Portland, Oregon, USA. Association for Computational Linguistics, 64-68.

Brown, G. I. (2011b) Relation Extraction on the J.D. Power and Associates Sentiment Corpus. Master of Science, University of Colorado.

Buscaldi, D., Rosso, P., Gómez-Soriano, J. M. & Sanchis, E. (2010) Answering questions with an n-gram based passage retrieval engine. Journal of Intelligent Information Systems, Vol. 34, 2, pp. 113-134.

Calvo, R. A. & D'mello, S. (2010) Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. IEEE Transactions on Affective Computing, Vol. 1, 1, pp. 18-37.

Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M. & Gandini, G. (2007) Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. In: ANGELI, F. (ed.) Language resources and linguistic theory: Typology, second language acquisition, English linguistics Milano, IT,

MSRA. (2009a) Better Parser Combination.

Chen, Y. C., Shang, R. A. & Kao, C. Y. (2009b) The effects of information overload on consumers' subjective state towards buying decision in the internet shopping environment. Electronic Commerce Research and Applications, Vol. 8, 1, pp. 48-58.

Chesley, P., Vincent, B., Xu, L. & Srihari, R. (2006) Using Verbs and Adjectives to Automatically Classify Blog Sentiment. In: The Spring Symposia on Computational Approaches to Analyzing Weblogs. Stanford, US.

Choi, Y. & Cardie, C. (2008) Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: The Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii. Association for Computational Linguistics, 793-801.

Choi, Y., Cardie, C., Riloff, E. & Patwardhan, S. (2005) Identifying sources of opinions with conditional random fields and extraction patterns. In: The conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada. Association for Computational Linguistics, 355-362.

Christensen, H. U. & Ortiz-Arroyo, D. (2007) Applying data fusion methods to passage retrieval in QAS. In: The 7th international conference on Multiple classifier systems. Prague, Czech Republic. Springer-Verlag, 82-92.

Chung, M. K. (2007) Correlation Coefficient. Encyclopedia of Measurement and Statistics. SAGE.

Cicchetti, D. V. & Feinstein, A. R. (1990) High agreement but low kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology, Vol. 43, 6, pp. 551-558.

Cohen, J. (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement, Vol. 20, pp. 37-46.

Cohen, J. (1968) Weighted kappa; nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, pp. 213-220.

Cohen, W. W. (1995) Learning to classify English text with ILP methods. In: RAEDT, L. D. (ed.) Advances in inductive logic programming. Amsterdam, ISO Press.

Collins, M. (1997) Three generative, lexicalised models for statistical parsing. In: The 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. Madrid, Spain. Association for Computational Linguistics, 16-23.

Collins, M. (2003) Head-Driven Statistical Models for Natural Language Parsing. Association for Computational Linguistics, Vol. 29, 4, pp. 589-637.

Consoli, D., Diamantini, C. & Potena, D. (2008) A Conceptual Framework for Web Opinion Mining. In: 5th Conference of the Italian Chapter of AIS Paris, France.

Croft, B., Metzler, D. & Strohman, T. (2009) Evaluating Search Engine. In: Search Engines: Information Retrieval in Practice. Cloth, Addison-Wesley. 552.

Das, D. & Martins, A. F. T. (2007) A Survey on Automatic Text Summarization. Literature Survey for the Language and Statistics II course. Carnegie Mellon University.

Das, S. R. & Chen, M. Y. (2007) Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. Management Science, Vol. 53, 9, pp. 1375-1388.

Dave, K., Lawrence, S. & Pennock, D. M. (2003) Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: The 12th international conference on World Wide Web. Budapest, Hungary. ACM, 519-528.

Demars, C. E. & Erwin, T. D. (2005) Neutral or Unsure: Is there a Difference? . In: The annual meeting of the American Psychological Association. Washington, DC. James Madison University, 1-12.

Devitt, A. & Ahmad, K. (2008) Sentiment analysis and the use of extrinsic datasets in evaluation. In: The Sixth International Language Resources and Evaluation (LREC'08). Marrakech, Morocco. European Language Resources Association (ELRA).

Dey, L. & Haque, S. K. M. (2008) Opinion mining from noisy text data. In: The second workshop on Analytics for noisy unstructured text data. Singapore. ACM, 83-90.

Ding, X., Liu, B. & Yu, P. S. (2008) A holistic lexicon-based approach to opinion mining. In: The international conference on Web search and web data mining. Palo Alto, California, USA. ACM, 231-240.

Egghe, L. (2004) A universal method of information retrieval evaluation: the "missing" link M and the universal IR surface. Information Processing and Management: an International Journal, Vol. 40, 1, pp. 21-30.

Ekman, P. (1985) Telling Lies: Clues to Deceit in the Marketplace. Politics, and Marriage. New York, Norton.

Esuli, A. (2008) Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms and Applications. Ph.D., UNIVERSITÀ DI PISA.

Esuli, A. & Sebastiani, F. (2005) Determining the semantic orientation of terms through gloss classification. In: The 14th ACM international conference on Information and knowledge management. Bremen, Germany. ACM, 617-624.

Esuli, A. & Sebastiani, F. (2006a) Determining Term Subjectivity and Term Orientation for Opinion Mining. In: The 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy. 193-200.

Esuli, A. & Sebastiani, F. (2006b) SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In: 5th Conference on Language Resources and Evaluation. Genova, IT.

Fei, Z., Huang, X. & Wu, L. (2006) Mining the relation between sentiment expression and target using dependency of words. In: 20th Pacific Asia Conf. on Language, Information and Computation (PACLIC20). Wuhan, China. 257-264.

Fellbaum, C. (1997) WordNet. An Electronic Lexical Database and Some of its Applications. Cambridge, MA, The MIT Press.

Fiscus, J. G. & Doddington, G. R. (2002) Topic detection and tracking evaluation overview. In: Topic detection and tracking. Kluwer Academic Publishers. 17-31.

Fleiss, J. L. & Cohen, J. (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement Vol. 33, pp. 613-619.

Framenet (n.d.-a) FrameNet Data [WWW]. Available from: https://framenet.icsi.berkeley.edu/fndrupal/ [Accessed March 23, 2012].

Framenet (n.d.-b) FrameNets In Other Languages [WWW]. Available from: https://framenet.icsi.berkeley.edu/fndrupal/framenets_in_other_languages [Accessed March 23, 2012].

Ganapathibhotla, M. & Liu, B. (2008) Mining opinions in comparative sentences. In: The 22nd International Conference on Computational Linguistics. Manchester, United Kingdom. Association for Computational Linguistics, 241-248.

Gantz, J., Boyd, A. & Dowling, S. (2009) Cutting the Clutter: Tackling Information Overload. IDC White Paper. Xerox Corporation.

Godbole, N., Srinivasaiah, M. & Skiena, S. (2007) Large-Scale Sentiment Analysis for News and Blogs. In: International Conference on Weblogs and Social Media. Boulder, Colorado, U.S.A.

Stanford University (2004) Sentiment Extraction and Classification of Movie Reviews.

Gómez-Soriano, J. M., Buscaldi, D., Asensi, E. B., Rosso, P. & Arnal, E. S. (2006) QUASAR: the question answering system of the Universidad Politécnica de Valencia. In: The 6th international conference on Cross-Language Evalution Forum: accessing Multilingual Information Repositories. Vienna, Austria. Springer-Verlag,

Gopal, R., Marsden, J. R. & Vanthienen, J. (2011) Information mining — Reflections on recent advancements and the road ahead in data, text, and media mining. Decision Support Systems Vol. 51, 4, pp. 1-5.

Gorman, K. (2009) On VARBRUL - Or, The Spirit of `74. Philadelphia, PA, USA, Institute for Research in Cognitive Science, University of Pennsylvania.

Greene, B. B. & Rubin, G. M. (1971) Automatic Grammatical Tagging of English. Department of Linguistics, Brown University.

Greene, S. C. (2007) Spin: Lexical Semantics, Transitivity, and the Identification of Implicit Sentiment. Ph.D., University of Maryland.

Grefenstette, G., Qu, Y., Shanahan, J. G. & Evans, D. A. (2004) Coupling niche browsers and affect analysis for an opinion mining application. In: Computer-Assisted Information Retrieval (Recherche d'Information Assistée par Ordinateur (RIAO)). University of Avignon, France.

Grobelnik, M. & Mladenic, D. (2004) Text-Mining Tutorial. In: Learning Methods for Text Understanding and Mining. Grenoble, France.

Guo, H., Zhu, H., Guo, Z., Zhang, X. & Su, Z. (2010) OpinionIt: a text mining system for cross-lingual opinion analysis. In: The 19th ACM international conference on Information and knowledge management. Toronto, ON, Canada. ACM, 1199-1208.

Gupta, V. & Lehal, G. S. (2009) A Survey of Text Mining Techniques and Applications. Journal of Emerging Technologies in Web Intelligence, Vol. 1, 1, pp. 60-76.

Gyamfi, Y., Wiebe, J., Mihalcea, R. & Akkaya, C. (2009) Integrating knowledge for subjectivity sense labeling. In: Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Boulder, Colorado. Association for Computational Linguistics, 10-18.

Hamouda, A., Marei, M. & Rohaim, M. (2012) Building Machine Learning Based Senti-Word Lexicon For Sentiment Analysis. The Journal of Communications and Computer Engineering, Vol. 2, 1, pp. 199-203.

Harley, T. A. (2007) Understanding the Structure of Sentences. 3rd edition. Psychology Press.

Hatzivassiloglou, V. & Mckeown, K. R. (1997) Predicting the Semantic Orientation of Adjectives. In: The eighth conference on European chapter of the Association for Computational Linguistics. Madrid, Spain. Association for Computational Linguistics, 174-181

Hatzivassiloglou, V. & Wiebe, J. M. (2000) Effects of adjective orientation and gradability on sentence subjectivity. In: The 18th conference on Computational linguistics Saarbrücken, Germany. Association for Computational Linguistics, 299-305

Hayek, F. A. (1945) The Use of Knowledge in Society. The American Economic Review, Vol. 35, 4, pp. 519-530.

Higashinaka, R., Prasad, R. & Walker, M. A. (2006) Learning to generate naturalistic utterances using reviews in spoken dialogue systems. In: The 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Sydney, Australia. Association for Computational Linguistics, 265-272.

Hindle, D. (1983) Deterministic parsing of syntactic non-fluencies. In: The 21st annual meeting on Association for Computational Linguistics. Cambridge, Massachusetts. Association for Computational Linguistics,

Hong, S. J. & Weiss, S. M. (1999) Advances in Predictive Model Generation for Data Mining In: 1st International Workshop Machine Learning and Data Mining in Pattern Recognition.

Hu, M. & Liu, B. (2004) Mining and summarizing customer reviews. In: The tenth ACM SIGKDD international conference on Knowledge discovery and data mining. Seattle, WA, USA. ACM, 168-177.

Hunston, S. & Sinclair, J. (2000) A local grammar of evaluation. In: HUNSTON, S. & THOMPSON, G. (eds.) Evaluation in Text: authorial stance and the construction of discourse. Oxford, England, Oxford University Press.

Jacoby, J., Speller, D. E. & Kohn, C. A. (1974) Brand Choice Behavior as a Function of Information Load. Journal of Marketing Research, Vol. 11, 1, pp. 63-69.

Jijkoun, V., Rijke, M. D. & Weerkamp, W. (2010) Generating focused topic-specific sentiment lexicons. In: The 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden. Association for Computational Linguistics, 585-594.

Jin, F., Huang, M. & Zhu, X. (2009) A query-specific opinion summarization system. In: IEEE International Conference on Cognitive Informatics. Hong Kong, China. 428-433.

Jin, W. & Ho, H. H. (2009) A novel lexicalized HMM-based learning framework for web opinion mining. In: The 26th Annual International Conference on Machine Learning. Montreal, Quebec, Canada. ACM,

Jindal, N. & Liu, B. (2006a) Annotated DataSet. Department of Computer Sicence, University of Illinois at Chicago

Jindal, N. & Liu, B. (2006b) Identifying comparative sentences in text documents. In: The 29th annual international ACM SIGIR conference on Research and development in information retrieval. Seattle, Washington, USA. ACM, 244-251.

Jindal, N. & Liu, B. (2006c) Mining comparative sentences and relations. In: The 21st national conference on Artificial intelligence - Volume 2. Boston, Massachusetts. AAAI Press, 1331-1336.

Jindal, N. & Liu, B. (2008) Opinion spam and analysis. In: The International Conference on Web Search and Web Data Mining. 219-230.

Johns, R. (2010) Likert items and scales. SURVEY QUESTION BANK: Methods Fact Sheet. University of Strathclyde.

Kanayama, H. & Nasukawa, T. (2006) Fully automatic lexicon expansion for domain-oriented sentiment analysis. In: The 2006 Conference on Empirical Methods in Natural Language Processing. Sydney, Australia. Association for Computational Linguistics, 355-363.

Kennedy, A. & Inkpen, D. (2006) Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, Vol. 22, 2, pp. 110-125.

Kessler, J. S., Eckert, M., Clark, L. & Nicolov, N. (2010) The 2010 ICWSM JDPA Sentment Corpus for the Automotive Domain. 4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge.

Kessler, J. S. & Nicolov, N. (2009) Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations. In: 3rd International AAAI Conference on Weblogs and Social Media (ICWSM 2009). San Jose, California.,

Kim, S.-M. & Hovy, E. (2004a) Determining the sentiment of opinions. In: Proceedings of the 20th international conference on Computational Linguistics. Geneva, Switzerland. Association for Computational Linguistics,

Kim, S.-M. & Hovy, E. (2006a) Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In: Workshop on Sentiment and Subjectivity in Text. Sydney, Australia. Association for Computational Linguistics, Kim, S. M. & Hovy, E. (2004b) Determining the sentiment of opinions. In: The 20th international conference on Computational Linguistics. Geneva, Switzerland. Association for Computational Linguistics, 1367.

Kim, S. M. & Hovy, E. (2006b) Automatic identification of pro and con reasons in online reviews. In: The COLING/ACL Main Conference Poster Sessions. Sydney, Australia. 483-490.

Kim, S. M. & Hovy, E. (2007) Crystal: Analyzing Predictive Opinions on the Web. In: Empirical Methods on Natural Language Processing-Computational Natural Language Learning (EMNLP-CoNLL). ACL, 1056-1064.

Kipper, K., Korhonen, A., Ryant, N. & Palmer, M. (2006) Extending VerbNet with Novel Verb Classes. In: 5th international conference on Language Resources and Evaluation.

Kosala, R. & Blockeel, H. (2000a) Web Mining Research: A Survey. ACM SIGKDD, Vol. 2, 1,

Kosala, R. & Blockeel, H. (2000b) Web Mining Research: A Survey. ACM SIGKDD Explorations Newsletter Vol. 2, 1, pp. 1-15

Kothari, C. R. (2008) Research Methodology: Methods And Techniques. 2nd edition. Boston, New Age International (P) Ltd.

Ku, L. W. & Chen, H. H. (2007) Mining opinions from the Web: Beyond relevance retrieval. Journal of the American Society for Information Science and Technology, Vol. 58, 12, pp. 1838-1850.

Ku, L. W., Liu, I. C., Lee, C. Y., Chen, K. H. & Chen, H. H. (2008) Sentence-Level Opinion Analysis by CopeOpi in NTCIR-7. In: NTCIR-7 Workshop Meeting. Tokyo, Japan. 260-267.

Langville, A. N. & Meyer, C. D. (2006) Information Retrieval and Web Search.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.

Leeds_University (n.d) Automatic Mapping Among Lexico-Grammatical Annotation Models (AMALGAM) [WWW]. Available from: http://www.scs.leeds.ac.uk/ccalas/amalgam/amalgover.html [Accessed March 18, 2012].

Lehnert, W., Cardie, C., Fisher, D., Mccarthy, J., Riloff, E. & Soderland, S. (1992) University of Massachusetts: Description of the CIRCUS System as usedfor MUC-4. In: The Fourth Message UnderstandingConference. 282-288.

University of Massachusetts (1991) The CIRCUS System as Used in MUC-3.

Levin, B. (1993) English Verb Classes and Alternations: A Preliminary Investigation. Chicago, University of Chicago Press.

Li, C. H., Yang, J. C. & Park, S. C. (2012) Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet. Expert Systems with Applications: An International Journal Vol. 39, 1, pp. 765-772.

Li, N. & Wu, D. D. (2010) Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Systems, Vol. 48, 2, pp. 354-368.

Lihui, C. & Lian, C. W. (2005) Using Web structure and summarisation techniques for Web content mining. Information Processing &amp; Management, Vol. 41, 5, pp. 1225-1242.

Lin, W. H., Wilson, T., Wiebe, J. & Hauptmann, A. (2006) Which side are you on?: identifying perspectives at the document and sentence levels. In: The Tenth

Conference on Computational Natural Language Learning. New York City, New York. Association for Computational Linguistics, 109-116.

Lingpipe (n.d.) Word Sense Tutorial [WWW]. Available from: http://alias-i.com/lingpipe/demos/tutorial/wordSense/read-me.html [Accessed September 01, 2012].

Liscombe, J., Riccardi, G. & Hakkani-Tâ¨Ur, D. (2005) Using Context to Improve Emotion Detection in Spoken Dialog Systems. In: Interspeech. Lisbon, Portugal. 1845-1848.

Liu, B. (2010) Sentiment Analysis and Subjectivity. In: INDURKHYA, N. & DAMERAU, F. J. (eds.) Handbook of Natural Language Processing. 2009 or 2010 edition. Boca Raton, Florida - USA, Chapman & Hall / CRC Press : Taylor & Francis Group.

Liu, B. (2011) Opinion Mining: Abstraction and Techniques. Data Science Summer Institute. UIUC.

Lloret, E., Balahur, A., Gómez, J. M., Montoyo, A. & Palomar, M. (2012) Towards a unified framework for opinion retrieval, mining and summarization. Journal of Intelligent Information Systems, pp. 1-37.

Lloret, E., Romá-Ferri, M. T. & Palomar, M. (2011) COMPENDIUM: A Text Summarization System for Generating Abstracts of Research Papers. Natural Language Processing and Information Systems, Lecture Notes in Computer Science, pp. 3-14.

Lu, B. (2010) Identifying opinion holders and targets with dependency parser in Chinese news texts. In: The NAACL HLT 2010 Student Research Workshop. Los Angeles, California. Association for Computational Linguistics, 46-51.

Lu, Y., Castellanos, M., Dayal, U. & Zhai, C. (2011) Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach. In: The 20th international conference on World wide web. Hyderabad, India. 347--35.

Luhn, H. P. (1958) The automatic creation of literature abstracts. IBM Journal of Research and Development, Vol. 2, 2, pp. 159-165.

Luo, Q., Chen, E. & Xiong, H. (2011) A semantic term weighting scheme for text categorization. Expert Systems with Applications: An International Journal Vol. 38, 10, pp. 12708-12716.

Lutz, C. & White, G. M. (1986) The Anthropology of Emotions. Annual Review of Anthropology, Vol. 15, pp. 405-436.

Manning, C. D., Raghavan, P. & Schütze, H. (2009) An Introduction to Information Retrieval. Cambridge, England, Cambridge University Press.

Marneffe, M. C. D., Maccartney, B. & Manning, C. D. (2006) Generating Typed Dependency Parses from Phrase Structure Trees. In: The 5th International Conference on Language Resources and Evaluation. Genoa, Italy., 449-454.

Maynard, D., and Greenwood, M. A. (2014). *Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis.* In Proceedings of LREC 2014.

Mcenery, T. & Wilson, A. (1993) Corpora and Translation: Uses and Future Prospects. Lancaster, UK, University Centre for Computer Corpus Research on Language, Lancaster University

Mcknight, W. (2005) Text Data Mining in Business Intelligence : Building Business Intelligence. Information management. New York, Tony Carrini

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B. & Grishman, R. (2004) Annotating Noun Argument Structure for NomBank. In:

4th International Conference On Language Resources And Evaluation. Lisbon, Portugal.

Michael, G. (2004) Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In: The 20th international conference on Computational Linguistics. Geneva, Switzerland. Association for Computational Linguistics, 841.

Mihalcea, R., Banea, C. & Wiebe, J. (2007) Learning Multilingual Subjective Language via Cross-Lingual Projections. In: The 45th Annual Meeting of the Association of Computational Linguistics. 976-983.

Miller, G. A. (1995) WordNet: a lexical database for English. Communications of the ACM, Vol. 38, 11, pp. 39-41.

Mishne, G. & Rijke, M. D. (2006a) Capturing global mood levels using blog posts. AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs. AAAI Press.

Mishne, G. & Rijke, M. D. (2006b) MoodViews: Tools for Blog Mood Analysis. In: AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006). AAAI Press, 153-154.

Miyoshi, T. & Nakagami, Y. (2007) Sentiment Classification of Customer Reviews on Electric Products In: The IEEE International Conference on Systems, Man and Cybernetics. Montréal, Canada.

Moore, A. (2002) The Structure of English Language [WWW]. Available from: http://www.universalteacher.org.uk/lang/engstruct.htm [Accessed December 13, 2013].

Morinaga, S., Yamanishi, K., Tateishi, K. & Fukushima, T. (2002) Mining product reputations on the Web. In: The eighth ACM SIGKDD international conference on Knowledge discovery and data mining. Edmonton, Alberta, Canada. ACM, 341-349.

Mukherjee, S. & Bhattacharyya, P. (2012) Feature specific sentiment analysis for product reviews. In: The 13th international conference on Computational Linguistics and Intelligent Text Processing - Volume Part I. New Delhi, India. Springer-Verlag,

Na, J. C., Sui, H., Khoo, C., Chan, S. & Zhou, Y. (2004) Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. In: MCILWAINE (ed.) Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference. Wurzburg, Germany. Ergon Verlag, 49-54.

Na, J. C., Thet, T. T. & Khoo, C. S. G. (2010) Comparing sentiment expression in movie reviews from four online genres. Online Information Review, Vol. 34, 2, pp. 317-338.

Navigli, R. (2009) Word sense disambiguation: A survey. ACM Computing Surveys, Vol. 41, 2, pp. 1-69.

Neviarouskaya, A., Prendinger, H. & Ishizuka, M. (2007) Textual Affect Sensing for Sociable and Expressive Online Communication. In: The 2nd international conference on Affective Computing and Intelligent Interaction. Lisbon, Portugal. Springer-Verlag, 218 - 229

Neviarouskaya, A., Prendinger, H. & Ishizuka, M. (2009) SentiFul: Generating a reliable lexicon for sentiment analysis. In: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on. 1-6.

Neviarouskaya, A., Prendinger, H. & Ishizuka, M. (2011) SentiFul: A Lexicon for Sentiment Analysis. IEEE Transactions on Affective Computing, Vol. 2, 1, pp. 22-36.

Nist (2007a) Automatic Content Extraction (ACE) Evaluation [WWW]. Available from: http://www.itl.nist.gov/iad/mig/tests/ace/ [Accessed June 20, 2012].

Nist (2007b) Topic Detection and Tracking Evaluation [WWW]. Available from: http://www.itl.nist.gov/iad/mig/tests/tdt/ [Accessed July 25, 2012].

Niu, Y., Zhu, X., Li, J. & Hirst, G. (2005) Analysis of polarity information in medical text. The American Medical Informatics Association 2005 Annual Symposium,

Nivre, J. (2005) Dependency Grammar and Dependency Parsing. Växjö University.

Osherenko, A. & Andr, E. (2007) Lexical Affect Sensing: Are Affect Dictionaries Necessary to Analyze Affect? In: The 2nd international conference on Affective Computing and Intelligent Interaction. Lisbon, Portugal. Springer-Verlag, 230-241.

Palmer, M. & Marcus, M. (n.d.) Proposition Bank [WWW]. Available from: http://verbs.colorado.edu/~mpalmer/projects/ace.html [Accessed May 02, 2012].

Pang, B. & Lee, L. (2005) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: The 43rd Annual Meeting on Association for Computational Linguistics. Michigan, USA. Association for Computational Linguistics, 115-124

Pang, B. & Lee, L. (2008) Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, Vol. 2, 1, pp. 1-135.

Pang, B., Lee, L. & Vaithyanathan, S. (2002) Thumbs up?: Sentiment Classification Using Machine Learning Techniques. In: The ACL-02 conference on Empirical methods in natural language processing Philadelphia, PA, USA. Association for Computational Linguistics, 79-86.

Penn_Treebank (1992) The Penn Treebank Project [WWW]. Available from: http://www.cis.upenn.edu/~treebank/ [Accessed April 21, 2012].

Perrin, P. & Petry, F. E. (2003) Extraction and representation of contextual information for knowledge discovery in texts. Information Sciences—Informatics and Computer Science: An International Journal, Vol. 151, pp. 125-152.

Plutchik, R. (2002) Emotions and Life. Washington, D.C, American Psychological Association.

Polanyi, M. (1966) The Tacit Dimension. London, UK, Routledge & Kegan Paul.

Popescu, A. M. & Etzioni, O. (2005) Extracting product features and opinions from reviews. In: The conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada. Association for Computational Linguistics, 339-346.

Powers, D. M. W. (2011) Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, Vol. 2, 1, pp. 37-63.

Princeton.Edu (n.d.) WordNet A lexical database of English. [WWW]. Available from: http://wordnet.princeton.edu/wordnet/ [Accessed September 01, 2012].

Proofread_Bot (n.d.) Phrases and Parts of Speech Tags - Penn Treebank Tags [WWW]. Available from: http://proofreadbot.com/phrases-and-parts-speech-tags-penn-treebank-tags [Accessed January 10, 2014].

Qiu, G., Liu, B., Bu, J. & Chen, C. (2009) Expanding domain sentiment lexicon through double propagation. In: The 21st international jont conference on

Artifical intelligence. Pasadena, California, USA. Morgan Kaufmann Publishers Inc., 1199-1204.

Qiu, G., Liu, B., Bu, J. & Chen, C. (2011) Opinion word expansion and target extraction through double propagation. Computational Linguistics Vol. 37, 1, pp. 9-27.

Quan, C. & Ren, F. (2010) A blog emotion corpus for emotional expression analysis in Chinese. Computer Speech and Language, Vol. 24, 4, pp. 726-749.

Rauber, A. & Merkl, D. (1999) Automatic Labeling of Self-Organizing Maps: Making a Treasure-Map Reveal Its Secrets. In: The Third Pacific-Asia Conference on Methodologies for Knowledge Discovery and Data Mining. Springer-Verlag,

Rill, S., Adolph, S., Drescher, J., Reinel, D., Scheidt, J., Schütz, O., Wogenstein, F., Zicari, R. V. & Korfiatis, N. (2012a) A Phrase-Based Opinion List for the German Language. In: 1st Workshop on Practice and Theory of Opinion Mining and Sentiment Analysis (PATHOS). Vienna, Austria.

Rill, S., Scheidt, J., Drescher, J., Schutz, O., Reinel, D. & Wogenstein, F. (2012b) A generic approach to generate opinion lists of phrases for opinion mining applications. In: The First International Workshop on Issues of Sentiment Discovery and Opinion Mining. Beijing, China. ACM,

Riloff, E. & Wiebe, J. (2003) Learning extraction patterns for subjective expressions. In: The 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics, 105-112

Riloff, E., Wiebe, J. & Phillips, W. (2005) Exploiting subjectivity classification to improve information extraction. In: The 20th national conference on Artificial intelligence - Volume 3. Pittsburgh, Pennsylvania. AAAI Press, 1106-1111.

Saggion, H. & Lapalme, G. (2002) Generating indicative-informative summaries with sumUM. Computational Linguistics, Vol. 28, 4, pp. 497-526.

Sarmento, L., Carvalho, P., Silva, M. J. & Oliveira, E. D. (2009) Automatic creation of a reference corpus for political opinion mining in user-generated content. In: The 1st international CIKM workshop on Topic-sentiment analysis for mass opinion. Hong Kong, China. 29-36.

Scherer, K. R. (1984) Emotion as a Multicomponent Process: A model and some cross-cultural data. In: SHAVER, P. (ed.) Review of Personality and Social Psych. 37-63.

Schlenker, P. (2006) Sentence Structure I: Syntactic Trees [WWW]. Available from: [Accessed January 16, 2013].

Schuller, B. R., Batliner, A., Steidl, S. & Seppi, D. (2011) Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, Vol. 53, 9â€“10, pp. 1062-1087.

Seco, N. A. L. (2005) Computational Models of Similarity in Lexical Ontologies. Masters in Computer Science, University College Dublin.

Shaikh, M. A. M., Prendinger, H. & Ishizuka, M. (2008) Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. Applied Artificial Intelligence, Vol. 22, 6, pp. 558-601.

Shalizi, C. (2009) Lecture 1: Similarity Searching and Information Retrieval, Data Mining. Statistics. Pittsburgh, Department of Statistics, Carnegie Mellon University.

Sheth, B. & Maes, P. (1993) Evolving Agents For Personakzed Information Filtering. In: The Ninth IEEE Conference on Artificial Intelligence for Applications. Orlando, Florida. 245-252.

Snyder, B. & Barzilay, R. (2007) Multiple aspect ranking using the good grief algorithm. In: The Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL). Rochester, New York. Association for Computational Linguistics, 300-307.

Somasundaran, S. (2010) Discourse- Level Relations for Opnionion Analysis. Ph.D., University of Pittsburgh

Stelling, L. E. (2011) The effects of grammatical proscription on morphosyntactic change: Auxiliary variation in Franco-American French. dentités linguistiques, langues identitaires : à la croisée du prescriptivisme et du patriotisme, Vol. 1,

Stoyanov, V. & Cardie, C. (2006) Toward opinion summarization: linking the sources. In: The Workshop on Sentiment and Subjectivity in Text. Sydney, Australia. Association for Computational Linguistics, 9-14.

Stoyanov, V. & Cardie, C. (2008) Topic identification for fine-grained opinion analysis. In: The 22nd International Conference on Computational Linguistics Manchester, United Kingdom. Association for Computational Linguistics, 817-824.

Strapparava, C. & Valitutti, A. (2004) WordNet-Affect: An affective extension of WordNet. In: The 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal. 1083-1086.

Stumme, G., Hotho, A. & Berendt, B. (2006) Semantic Web Mining: State of the art and future directions. Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 4, 2, pp. 124-143.

Subrahmanian, V. S. & Reforgiato, D. (2008) AVA: Adjective-Verb-Adverb Combinations for Sentiment Analysis. IEEE Intelligent Systems, Vol. 23, 4, pp. 43-50.

Suchanek, F. M., Kasneci, G. & Weikum, G. (2007) YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: The 16th international conference on World Wide Web. Banff, Alberta, Canada. . ACM, 697-706.

Sykora, M., Jackson, T., O'brien, A. & Elayan, S. (2013a) EMOTIVE Ontology: Extracting Fine-Grained Emotions from Terse, Informal Messages. In: 7th IADIS Intelligent Systems Agents and Data Mining (ISA-DM) conference. Prague, Czech Republic.

Sykora, M., Jackson, T. W., O'brien, A. & Elayan, S. (2013b) National security and social media monitoring: A presentation of the EMOTIVE and related systems. In: IEEE European Intelligence and Security Informatics Conference. Uppsala, Sweden.

Tadano, R., Shimada, K. & Endo, T. (2009) Effective Construction and Expansion of a Sentiment Corpus Using an Existing Corpus and Evaluative Criteria Estimation. In: Conference of the Pacific Association for Computational Linguistics: PACLING 2009. Sapporo, Japan.

Takamura, H., Inui, T. & Okumura, M. (2007) Extracting Semantic Orientations of Phrases from Dictionary. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. Association for Computational Linguistics, 292-299.

Tan, L. K.-W., Na, J.-C., Theng, Y.-L. & Chang, K. (2011a) Sentence-level sentiment polarity classification using a linguistic approach. In: The 13th international conference on Asia-pacific digital libraries: for cultural heritage, knowledge dissemination, and future creation. Beijing, China. Springer-Verlag, 77-87.

Tan, L. K. W., Na, J.-C. & Theng, Y.-L. (2011b) Phrase-Level Sentiment Polarity Classification Using Rule-. Based Typed Dependencies. In: The ACM SIGIR 3rd Workshop on Social Web Search and Mining SIGIR-SWSM 2011. Beijing, China.

Tan, L. W., Na, J. C., Theng, Y. L. & Chang, K. (2011c) Phrase-Level Sentiment Polarity Classification Using Rule-Based Typed Dependencies and Additional Complex Phrases Consideration. Journal of Computer Science and Technology, Vol. 27, 3, pp. 650-666.

The_Ice_Project (n.d) International Corpus of English (ICE) [WWW]. Available from: http://ice-corpora.net/ice/ [Accessed July 20, 2012].

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in twitter events. Journal of the American Society for Information Science and Technology, 62(2), 406-418.

Thelwall, M., Buckley, K. & Paltoglou, G. (2012) Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, Vol. 63, 1, pp. 163-173.

Thet, T. T., Na, J. C. & Khoo, C. S. G. (2010) Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science, Vol. 36, 6, pp. 823-848.

Thompson, W. D. & Walter, S. D. (1988) A reappraisal of the kappa coefficient. Journal of Clinical Epidemiology, Vol. 41, 10, pp. 949-958.

Töllinen, A., Järvinen, J. & Karjaluoto, H. (2012) Social Media Monitoring in the industrial Business to Business Sector. World Journal of Social Sciences, Vol. 2, 4, pp. 65-76.

Toprak, C., Jakob, N. & Gurevych, I. (2010) Sentence and expression level annotation of opinions in user-generated discourse. In: The 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden. Association for Computational Linguistics, 575-584.

Torres-Moreno, J.-M., St-Onge, P.-L., Gagnon, M. & Bellot, P. (2009a) Automatic Summarization System coupled with a Question-Answering System (QAAS). NLP News Computation and Language,

Torres-Moreno, J. M., St-Onge, P. L., Gagnon, M., El-Bèze, M. & Bellot, P. (2009b) Automatic Summarization System coupled with a Question-Answering System (QAAS). The Computing Research Repository, Vol. 905,

Turmo, J., Ageno, A. & Catal, N. (2006) Adaptive information extraction. ACM Computing Surveys Vol. 38, 2, p. Articale No. 4.

Turney, P. D. (2002) Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: The 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania. Association for Computational Linguistics, 417-424

Uebersax, J. (1987) Diversity of decision-making models and the measurement of interrater agreement. Psychological Bulletin, Vol. 101, pp. 140-146.

Ur-Rahman, N. & Harding, J. A. (2011) Textual data mining for industrial knowledge management and text classification: A business oriented approach. Expert Systems with Applications, Vol. 39, 5, pp. 4729-4739.

Verbnet (n.d) A Class-Based Verb Lexicon [WWW]. Available from: http://verbs.colorado.edu/~mpalmer/projects/verbnet.html [Accessed April 20, 2012].

Viera, A. J. & Garrett, J. M. (2005) Understanding Interobserver Agreement:The Kappa Statistic. Family Medicine, Vol. 37, 5, pp. 360-363.

Voorhees, E. M. (2001) The Philosophy of Information Retrieval Evaluation. In: Second Workshop of the Cross-Language Evaluation Forum. Darmstadt, Germany 355-370.

Waegel, D. (2006) The Development of Text-Mining Tools and Algorithms. Distinguished Honors in Computer Science, Ursinus College.

Wajeed, M. A. & Adilakshmi, T. (2009) Text Classification using Machine learning. A Journal of Theoretical and Applied Information Technology, Vol. 7, 2, pp. 119-123.

Webster, D. E. (2010) Realising context-oriented information filtering. Ph. D., The University of Hull, Scarborough Campus.

Westerski, A. (2008) Sentiment Analysis : Introduction and the State of the Art overview. Madrid, Spain, Universidad Politecnica de Madrid, Spain.

Wiebe, J. & Riloff, E. (2005) Creating subjective and objective sentence classifiers from unannotated texts. In: The 6th international conference on Computational Linguistics and Intelligent Text Processing. Mexico City, Mexico. Springer-Verlag, 486-497.

Wiebe, J., Wilson, T. & Cardie, C. (2005) Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation, Vol. 1, 2, pp. 164-210.

Wilks, Y. & Stevenson, M. (1998) The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. Journal of Natural Language Engineering, Vol. 4, 2, pp. 135-144.

Wilson, T. (2008) Fine-Grained Subjectivity Analysis. PhD, University of Pittsburgh.

Wilson, T. & Wiebe, J. (2003) Annotating Opinions in the World Press. In: The 4th ACL SIGdial Workshop on Discourse and Dialogue (SIGdial-03). 13-22.

Wilson, T., Wiebe, J. & Hoffmann, P. (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: The conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada. Association for Computational Linguistics, 347-354.

Wilson, T., Wiebe, J. & Hoffmann, P. (2009) Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics Vol. 35, 3, pp. 399-433.

Wilson, T., Wiebe, J. & Hwa, R. (2004) Just how mad are you? finding strong and weak opinion clauses. In: The 19th national conference on Artifical intelligence. San Jose, California. AAAI Press, 761-767

Witten, I. H., Bray, Z., Mahoui, M. & Teahan, B. (1999) Text Mining: A New Frontier for Lossless Compression. In: The Conference on Data Compression. IEEE Computer Society,

Wu, G. S., Wu, X. Y. & Wei, J. J. (2012) Sentiment Analysis of Comparative Sentences for Chinese Document. Applied Mechanics and Materials, Vol. 157 - 158, pp. 1079-1082.

Wu, Y., Zhang, Q., Huang, X. & Wu, L. (2009) Phrase dependency parsing for opinion mining. In: The 2009 Conference on Empirical Methods in Natural Language

Processing: Volume 3 - Volume 3. Singapore. Association for Computational Linguistics, 1533-1541.

Www.Stanford.Edu (n.d) Stanford Dependencies [WWW]. Available from: http://nlp.stanford.edu/software/stanford-dependencies.shtml [Accessed Feburary 22, 2012].

Xu, G., Zhang, Y. & Li, L. (2010) Web Mining and Social Networking: Techniques and Applications : Volume 6 of Web Information Systems Engineering and Internet Technologies Book Series. London, Springer.

Xu, R., Xu, J. & Wang, X. (2011) Instance level transfer learning for cross lingual opinion analysis. In: The 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Portland, Oregon. Association for Computational Linguistics, 182-188.

Yadav, V., Elchuri, H. & Palanisamy, P. (2013) Serendio: Simple and Practical lexicon based approach to Sentiment Analysis. In: SemEval-2013 : Semantic Evaluation Exercises International Workshop on Semantic Evaluation. Atlanta, Georgia.

Yang, X. P. & Liu, X. R. (2008) Personalized multi-document summarization in information retrieval. In: International Conference on Machine Learning and Cybernetics. 4108 - 4112

Yi, J., Nasukawa, T., Bunescu, R. & Niblack, W. (2003) Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: Third IEEE International Conference on Data Mining, 2003. ICDM 2003. . San Jose, CA, USA 427 - 434

Zhang, L. & Liu, B. (2011) Identifying noun product features that imply opinions. In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2. Portland, Oregon. Association for Computational Linguistics, 575-580.

Zhu, F. & Zhang, X. M. (2010) Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. Journal of Marketing, Vol. 74, 2, pp. 133-148.

-

Appendix A

Existing Corpora

MPQA Opinion Corpus

MPQA Opinion Corpus is mainly based on word and phrase as a unit. It is all manually annotated for the private states (“a general term that covers opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments. (Wiebe et al., 2005)). They generated ‘private states frames’, ‘objective speech event frames’ and ‘agent frames’. They defined each private state frame as ‘text anchor’: a span of text which represents the opinion, ‘source’: an entity that expressed opinion, ‘target’: topic of opinion. It assigns the ordinal value to each opinion word as ‘intensity’ over a scale of low, medium, high, or extreme. There is another property about ‘expression of intensity’ over a scale of neutral, low, medium, high, or extreme. It assigns a property of ‘attitude type’ which can have a value of positive, negative, other, or none. An example of all these as quoted by Wiebe et al. (2005), “The report is full of absurdities,” Xirao-Nima said. [“US Human Rights Report Defies Truth,” 2002-02-11, By Xiao Xin, Beijing China Daily, Beijing, China]

Objective speech event frame:

Text anchor: the entire sentence

Source: <writer>

Implicit: true

Direct subjective frame:

Text anchor: said

Source: <writer,Xirao-Nima>

Intensity: high

Expression intensity: neutral

Target: report

Attitude type: negative

Expressive subjective element frame:

Text anchor: full of absurdities

Source: <writer, Xirao-Nima>

Intensity: high

Attitude type: negative

A tabular representation of MPQA Subjectivity Lexicon is given by Potts (2011).

	Strength	**Length**	**Word**	**Part-of-speech**	**Stemmed**	**Polarity**
1.	type=weaksubj	len=1	word1=abandoned	pos1=adj	stemmed1=n	priorpolarity=negative
2.	type=weaksubj	len=1	word1=abandonment	pos1=noun	stemmed1=n	priorpolarity=negative
3.	type=weaksubj	len=1	word1=abandon	pos1=verb	stemmed1=y	priorpolarity=negative
4.	type=strongsubj	len=1	word1=abase	pos1=verb	stemmed1=y	priorpolarity=negative
5.	type=strongsubj	len=1	word1=abasement	pos1=anypos	stemmed1=y	priorpolarity=negative
6.	type=strongsubj	len=1	word1=abash	pos1=verb	stemmed1=y	priorpolarity=negative
7.	type=weaksubj	len=1	word1=abate	pos1=verb	stemmed1=y	priorpolarity=negative
8.	type=weaksubj	len=1	word1=abdicate	pos1=verb	stemmed1=y	priorpolarity=negative
9.	type=strongsubj	len=1	word1=aberration	pos1=adj	stemmed1=n	priorpolarity=negative
10.	type=strongsubj	len=1	word1=aberration	pos1=noun	stemmed1=n	priorpolarity=negative
...						
8221.	type=strongsubj	len=1	word1=zest	pos1=noun	stemmed1=n	priorpolarity=positive

Table 1: A fragment of the MPQA subjectivity lexicon Potts (2011).

In the representation in Table 1 information like Part of Speech and stemmed or not is also included. However the syntactic analysis is missing in MPQA corpus. More emphasis is given on the expression of opinion and less on content of opinion (Bloom, 2011). Although the scale for ordinal representation is extended in later versions for example 'attitude type' which had a value of {positive, negative, other, or none} now has range of possible values from {agree-neg, agree-pos, arguing-neg, arguing-pos, intention-neg, intention-pos, other-attitude, sentiment-neg, sentiment-pos, speculation}.

It also has added the properties like 'polarity' having range of values from {negative, positive, both, neutral, uncertain-negative, uncertain-positive, uncertain-both, uncertain-neutral}. However it has not assigned any numeric values to the corpus which makes it difficult to use this corpus as training set and make aggregation and averaging rules.

JDPA Sentiment Corpus

The J.D. Power and Associates (JDPA) Corpus consists of user-generated content (blog posts) containing opinions about automobiles and cameras (Kessler et al., 2010). All these documentshave been manually annotated for named, nominal, and pronominal mentions of the entities (Kessler et al., 2010). The annotators have used help from web searches by using a variety of car-related search terms by retrieving the results from certain blog-host sites. The annotated mentions in the Corpus are single or multi-word expressions which refer to a particular real world or abstract entity. The mentions are annotated to indicate sets of mentions which constitute co-reference groups referring to the same entity (Brown, 2011a). Five relationships are annotated between these entities: PartOf, FeatureOf, Produces, InstanceOf, and MemberOf. The main highlight of JDPA Corpus is the extraction of relations between entities even over many sentences (Brown, 2011b). Sentiment expression in JDPA is captured by calculating contextual polarity and contextual modifiers like negators (modifier that inverts the polarity of a sentiment expression: for example; noise have been suppressed, avoids any reduction and not a good car. Here suspended, avoids and not act to invert the polarity), neutralizers (modifiers that do not commit the speaker to the truth of the target sentiment expression: for example; if the interior is poor and I tried to get used to it and like it. Here it is targeting poor and tried to get used to and like neutralize it), committers (modifier that shift speaker's certainty toward a sentiment expression: for example; "sure this will drive well". Here sure is giving confirmation of polarity) and intensifiers (modifier that shift the intensity of a sentiment expression: for example; considerable benefits. Here considerable strengthens the polarity of benefits).

Although JDPA corpus is quite refined for performing relation extraction, it mainly relies on five main relations as described above and is manually annotated.

SentiWordNet

SentiWordNet is a lexical resource developed to support sentiment classification and opinion mining systems. It is freely distributed for non-commercial use, and licenses are available for commercial applications. SentiWordNet was based on WordNet (Miller, 1995) synsets which generally comprise of terms with similar meanings. SentiWordNet is manly motivated by the assumption that "different senses of the same term may have different opinion-related properties," (Esuli, 2008). Esuli and Sebastiani (2006) developed a method employing eight ternary classifiers and quantitatively analysing the glosses associated with synsets (Esuli and Sebastiani, 2006b) . They assign a triplet of numerical score {Positive, Negative, Objective} describing how strongly the each term enjoy each of the three properties. The scores range from 0.0 to 1.0 and sum up to 1.0. In the final version of SentiWordNet, SentiWordNet 3.0 all these scores are automatically generated by a two step process. in first step they have used semi-supervised learning method by using two small seed lists of seven positive and seven negative words (Baccianella et al., 2010). The second step using an iterative random walk algorithm (Baccianella et al., 2010) assign positive and negative scores on basis of definiens-definiendum binary relationship. The score for objective value is assigned so as to make the three values sum up to one. If the sum of positive and negative value is greater than 1, they have normalised the two values to sum up to 1.

The reliability of SentiWordNet was questioned by researchers (Neviarouskaya et al., 2009; Neviarouskaya et al., 2011). During the experiments these highlighted issues were tested and tried. It was observed that the issues identified by Neviarouskaya et al. (2009, 2011) are rectified in SentiWordNet. However, while experimenting with SentiWordNet for this research many of the unidentified issues were discovered and mentioned in Chapter 2. Another limitation of SentiWordNet is that it is solely based on words. It does not even give any relationships amongst words like WordNet. Although, it is built upon WordNet synsets and WordNet is built upon the conceptual and semantic relationships as discussed in the section about WordNet. Another limitation is, it provides no help in terms of word sense disambiguation. Based on WordNet synsets, SentiWordNet generally gives a list of (more than one) values (triplet of numerical scores) across each word. Whereas, only one of these values is acceptable in the queried scenario and resolving this is very complicated.

WordNet

WordNet is a semantic lexicon database for the English language developed at the Cognitive Science Laboratory of Princeton University (Suchanek et al., 2007; Princeton.Edu, n.d.). Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. WordNet synsets are interlinked have different (conceptual-semantic and lexical) relations (i.e. is-a, part-whole, transitive) with each other. WordNet is planned to model the human glossary and psycholinguistic findings have also been taken into account in its design phase (Suchanek et al., 2007). WordNet keeps track of the context of situations in which words are being used, which provides help in defining semantically similar words as synonyms. WordNet also provides the taxonomic relations between words (i.e. Super, sub and sibling relationship of words). WordNet is the most widely used lexicon by the community of language processing (Seco, 2005).

There are many other corpora and lexical resources available in the domain of opinion analysis and linguistic analysis, some of them are discussed in chapter 2 and out of them MPQA Corpus, JDPA Corpus, SentiWordNet and WordNet are further reviewed. However, this review and the experiments performed have brought forward following requirements and guidelines for a corpus in area of opinion analysis.

A brief over view of Existing resources and corpora

There are a diverse set of research and projects in the field of corpus development and refinement which is started about three decades ago. Some of interesting work in the evolution of corpus development and refinement are discussed further.

The earliest work in the field of linguistic corpus was in 1979 when David Sankoff and Henrietta developed an automated way Varbrul (a program). It was a statistical package using multivariate analysis for linguistic analysis and was basically developed to analyse phonological data (Bhattacharyya et al., 2009b; Gorman, 2009). Anthony Kroch and Don Hindle (1981, 1982) also used Varbrul for analysis of syntactic data (Hindle, 1983; Bhattacharyya et al., 2009c). Later, Don Hindle and Susan Pintzuk contributed to Varbrul by writing programs to manipulate coding strings (Stelling, 2011). In 1991, Anthony Kroch and Ann Taylor started a pilot project to develop a syntactically annotated corpus of Middle English (Bhattacharyya et al., 2009c). It was a funded project of National Science Foundation which resulted into a publication in 1994 as the first phase of the Penn-Helsinki Parsed Corpus of Middle English (PPCME1) (Bhattacharyya et al., 2009c). PPCMEI was the first version of corpus to describe the structure of the sentence in Middle English. The next phase of this project PPCME2, began in 1995. It was the time when Eric Brill (a graduate student at University of Pennsylvania already had written a tagger famously known as Brill tagger (Brill, 1992). The limitation of this tagger was that it needed a training set of manually label example sentences. This training set help tagger to develop a lexicon and a set of rules to assign tags. This tagger made it possible to include POS in PPCME2.

An interesting approach to obtain a corpus was followed by Pang et al (2002), they chose a collection of movie review data which is already been tagged explicitlyas assigning stars by reviewers (Pang et al., 2002). It was an effort to analyses opinion at document level, however it was extensively argued that document level opinion mining cannot guarantee that same opinion was maintained throughout the document (Greene, 2007). Later in 2003, Wilson and Wiebe proposed annotating scheme for annotation of expressions of opinions, beliefs, emotions, sentiment and speculation. It ended up into the most important corpus available for sentiment analysis in English: MPQA (Wilson and Wiebe, 2003). It is manually annotated at word and phrase level and later carefully revised. Later in 2005, Wiebe and Riloff proposed a rule based method to automatically generate a corpus with subjective and objective sentences classified using only un-annotated texts for training (Wiebe and Riloff, 2005). MPQA is mainly focused on the problem of subjectivity and MPQA 1.0 annotation scheme focus on identifying different ways of communicating opinions, less emphasis is given to the content of those opinions (Bloom, 2011). In MPQA 2.0 attitude and target annotations were highlighted.

In 1997, FrameNet Project was started at the International Computer Science Institute in Berkeley, California. It is a large set of manually annotated sentences, labelled according to their semantic roles (Framenet, n.d.-a). However FrameNet has limited number of roles and words in its annotated corpus and researchers have to use other techniques like clustering, in order to predict the frame for an unseen word (Kim and Hovy, 2006a).

There are a number of other corpora which follow the trend like Proposition Bank (PropBank) (2005) (Palmer and Marcus, n.d.). It was funded by ACE. The motive was to create a corpus of text which is annotated with information about basic semantic propositions. It was an extension in Penn TreeBank (Palmer and Marcus, n.d.), the predicate-argument relations were added to the syntactic trees of the Penn Treebank (Penn_Treebank, 1992). VerbNet is another project (2006), it maps PropBank verb types to their corresponding Levin classes (Levin, 1993) (in fact also added 57 new classes from Korhonen and Briscoe's (2004) (Kipper et al., 2006) as extension to Levin’s classes). It is a lexical resource that incorporates both semantic and syntactic information about its contents (Verbnet, n.d). Similar to VerbNet and PropBank, NomBank provided annotation scheme of noun arguments in Penn Treebank II (PTB) (Meyers et al., 2004). However, each of them have their own limited utility based on their point of emphasis e.g., verbs, nouns and propositions.

Devitt and Ahmed (2008) proposed a way to use extrinsic sources in order to build a corpus of sentiment bearing news. They proposed to match news with stock market index, when stock index was rising the news will be positive and vice versa (Devitt and Ahmad, 2008). This type of research requires an extensive analysis based o cognitive theories instead of quantitative aspects of the task which requires broad range of external resources and time.

The automatic creation of corpus is quite extensively researched since past couple of years as it is an open fact that it is a tedious job, needing a lot of resources in terms of time, money and experts. Even, after all this extensive work the chances of inconsistencies, inaccuracies and imprecision are very high. Therefore, automatic generation of corpus is considered to be the result. Although, it also needs a lot of human interaction and manual annotation for training dataset, yet it is considered to be a small portion as compared to corpus itself. The J.D. Power and Associates (JDPA) Corpus (Kessler et al., 2010) contains the user generated blogs posts about automobiles. Posts have been manually annotated for mentions, co-reference, meronymy, sentiment expressions and modifiers of sentiment expressions (Kessler et al., 2010). The main highlight of JDPA Corpus is the extraction of relations between entities even over many sentences (Brown, 2011b). A similar effort was done in form of Automatic Content Extraction Corpus (ACE) (Nist, 2007a). It was an ongoing project, which started back in 1999 and the latest version of the corpus available is ACE 2008. The limitation of ACEis its emphasis on relation extraction within the same sentence (Brown, 2011b). Another project under National Institute of Standards and Technology (NIST) is Topic Detection and Tracking (TDT) (Nist, 2007b). TDT generates a TDT Corpus. First version of corpus contained 26K news stories from Reuters and CNN and researchers annotated them based on 25 pre-defined events. Later, each version of TDT contains more data and more topics (Fiscus and Doddington, 2002).

Refinement and XMLifying the existing corpora (Bhattacharyya et al., 2009a; Bhattacharyya et al., 2009c), multilingual corpus (Framenet, n.d.-b) or even sometimes if corpus is not available in the language the translation of corpus (Mcenery and Wilson, 1993). There are other ongoing projects in corpora generation and improvement like The International Corpus of English (ICE). It was a project started in 1990 with the primary aim of collecting material for comparative studies of English worldwide. Twenty-four research teams around the world are preparing electronic corpora of their own national or regional variety of English (The_Ice_Project, n.d). There are different ways of automation of corpus implemented so far. Some are ontology based (making systems intelligent), some are getting external information, some others are heuristic and rule based, some are extensions of other corpora and some require manually annotated training data. All of them have some issues over each other but basic thing is all text analysers is grammatical analysis or parsing of each sentence (Leeds_University, n.d). There is a need to have a framework to automatically generation of a corpus based on the basic information of parsing and tokenising to extend semantic and subjectivity analysis. It is required to have extensively defined rules and heuristics along with the use of some basic dictionaries to generate such corpus. Such a corpus can be self-populating like the web crawlers which gives a good size to it.

Appendix B

List of part-of-speech (POS) tags used in the Penn Treebank Project taken from Bies et al. (1995) and Proofread_Bot (n.d)

Clause Level

S - simple declarative clause, i.e. one that is not introduced by a (possible empty) subordinating conjunction or a *wh*-word and that does not exhibit subject-verb inversion.

SBAR - Clause introduced by a (possibly empty) subordinating conjunction.

SBARQ - Direct question introduced by a *wh*-word or a *wh*-phrase. Indirect questions and relative clauses should be bracketed as SBAR, not SBARQ.

SINV - Inverted declarative sentence, i.e. one in which the subject follows the tensed verb or modal.

SQ - Inverted yes/no question, or main clause of a *wh*-question, following the *wh*-phrase in SBARQ.

Phrase Level

ADJP - Adjective Phrase.

ADVP - Adverb Phrase.

CONJP - Conjunction Phrase.

FRAG - Fragment.

INTJ - Interjection. Corresponds approximately to the part-of-speech tag UH.

LST - List marker. Includes surrounding punctuation.

NAC - Not a Constituent; used to show the scope of certain prenominal modifiers within an NP.

NP - Noun Phrase.

NX - Used within certain complex NPs to mark the head of the NP. Corresponds very roughly to N-bar level but used quite differently.

PP - Prepositional Phrase.

PRN - Parenthetical.

PRT - Particle. Category for words that should be tagged RP.

QP - Quantifier Phrase (i.e. complex measure/amount phrase); used within NP.

RRC - Reduced Relative Clause.

UCP - Unlike Coordinated Phrase.

VP - Vereb Phrase.

WHADJP - *Wh*-adjective Phrase. Adjectival phrase containing a *wh*-adverb, as in *how hot*.

WHAVP - *Wh*-adverb Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing a *wh*-adverb such as *how* or *why*.

WHNP - *Wh*-noun Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing some *wh*-word, e.g. *who, which book, whose daughter, none of which,* or *how many leopards*.

WHPP - *Wh*-prepositional Phrase. Prepositional phrase containing a *wh*-noun phrase (such as *of which* or *by whose authority*) that either introduces a PP gap or is contained by a WHNP.

X - Unknown, uncertain, or unbracketable. X is often used for bracketing typos and in bracketing *the...the*-constructions.

Word level

CC - Coordinating conjunction

CD - Cardinal number

DT - Determiner

EX - Existential there

FW - Foreign word

IN - Preposition or subordinating conjunction

JJ - Adjective

JJR - Adjective, comparative

JJS - Adjective, superlative

LS - List item marker

MD - Modal

NN - Noun, singular or mass

NNS - Noun, plural

NNP - Proper noun, singular

NNPS - Proper noun, plural

PDT - Predeterminer

POS - Possessive ending

PRP - Personal pronoun

PRP$ - Possessive pronoun (prolog version PRP-S)

RB - Adverb

RBR - Adverb, comparative

RBS - Adverb, superlative

RP - Particle

SYM - Symbol

TO - to

UH - Interjection

VB - Verb, base form

VBD - Verb, past tense

VBG - Verb, gerund or present participle

VBN - Verb, past participle

VBP - Verb, non-3rd person singular present

VBZ - Verb, 3rd person singular present

WDT - Wh-determiner

WP - Wh-pronoun

WP$ - Possessive wh-pronoun (prolog version WP-S)

WRB - Wh-adverb

Appendix C

50 Sentences Dataset

Number **Sentence**

1 It was truly hard to shoot on the beach, on the Walking Street, especially in low light.

2 The Fx500 superiority in quality, performance and value all add up to make anyone drool over!

3 So Sony, goodbye to you; your ugly Cyber shots are unreasonably expensive.

4 It took some fiddling to change the settings and get the shot I was looking for.

5 I absolutely loved my Canon S3 and S5 cameras.

6 First impression was that this model is larger and heavier and did not fit my favourite camera cases anymore, than previous models.

7 Image quality was not as good as expected.

8 Out of the box, I noticed the FZ28 is amazing lightweight! .

9 This is definitely a great camera for the value.

10 He found the interface to be quite complicated, whereas I found it simple and intuitive.

11 It is the best featured of the series on most counts, particularly its higher resolution of 12 MP and dual picture stabilization, mechanical and digital.

12 They do not tend to be flashy, they rarely have unique features, and they just are not very interesting.

13 Besides its decent performance, the L11 offers very nice image quality for a budget camera.

14 Its lens produced only minimal distortion, with telephoto shots coming out nearly distortion free and wide angle shots manifesting only minor barrel distortion around their edges.

15 Getting the best camera and with all the features I want was my greatest problem.

16 With the large LCD screen, I suddenly realized that her greatest worry was rendered moot.

17 There are many experts who think that the whole restructuring strategy is misbegotten.

18 But if you are in the restructuring business, you can t let these stray thoughts get in the way of your restructuring.

19 In October, they warned the Bush administration of a possible bankruptcy filing and started restructuring.

20 The Ford Interceptor Concept is jaw dropping, but not necessarily in a good way.

21 All the underpinnings sound fine to us, and in all likelihood could end up in a future Ford sedan.

22 While the Audi A4 and BMW 3 Series earn good ratings in frontal offset tests, both are rated marginal for side impact protection and poor for protection in rear crashes.

23 Most of the debates on this board come down to people refusing to admit other people might like something different than them.

24 The styling is handsome but inoffensive, the spec sheet is solid but unremarkable, and Suzuki is not much of a household name unless you dig motorcycles.

25 This Audi makes a great case for the traditional wagon.

26 It means the new 2010 Ghost now begins at $645,000, meaning more access for more buyers.

27 In addition, the E46 M3 Comp Pack also came with a quicker steering ratio.

28 Before continuing one sentence further, am I aware that I sound like the worlds most spoiled rotten brat?

29 The personal contract purchase payments I was offered on the MX5 were $75 a month lower than my local dealer could manage.

30 But even so, all models fall into the top tax band and the ML280 CDI and 320 CDI offer the best fuel economy, at an average of 34 mpg.

31 The reason is simple, it is priced well and has a very useful design and most people are going to take pleasure in the attention.

32 In other words, Toyota has another winner in its stable, and this one is going to delight the all new Scion dealer network because this little rig is a blank canvas waiting for some creative options.

33 Car looked sharp and had a pretty nice interior, with the best feature of all.

34 You can drive in style and feel like it too.

35 I am however embarrassed to admit I forgot his name.

36 The picture quality is good, but the battery life is short.

37 It would be sad for Mr Gonzalez to abandon them to appease his foes.

38 There is no question that the service at this restaurant is excellent.

39 Another nice thing is that the unit has both optical and coax digital audio outputs, though the latter was not mentioned in the literature I had scanned before buying.

40 This movie has a lot of creepiness to it and it has a lot of parts that made me jump.

41 It is also kind of a sad movie as well but a well done horror movie.

42 With the possible exception of a couple of British cars, no interior even comes close.

43 No, I am not talking about cheapo pocket digital cameras that everyone carries these days.

44 I have been so obsessed with my new toy fresh from New York City that I have neglected to blog, eat, or sleep.

45 The Olympus E1 and E400 suck because it has low continuous shooting less than 3 and just 3 AF points.

46 Engine performance and handling are excellent.

47 The overall fit and finish of the RL means I can be comfortable driving this car anywhere.

48 The screen for navigation and the rear camera is poor and difficult to see in the daytime.

49 The third row seating has minimal leg room and is hard to get into.

50 I am not happy that it does not have an MP3 player in front.