Generating Multilingual Subjectivity Resources using English Language

Vandana Jha
Department of Computer Science and Engineering
University Visvesvaraya College of Engineering
Bangalore University, Bangalore, India

Shreedevi G. R.
Department of Computer Science and Engineering
University Visvesvaraya College of Engineering
Bangalore University, Bangalore, India

P. Deepa Shenoy
Department of Computer Science and Engineering
University Visvesvaraya College of Engineering
Bangalore University, Bangalore, India

Venugopal K. R.
Department of Computer Science and Engineering
University Visvesvaraya College of Engineering
Bangalore University, Bangalore, India

ABSTRACT

The text data can be of two types : facts and opinions. With the introduction of UTF-8 standards and development of Web 2.0, we are in abundance of opinionated text data available in many languages on the web. Subjectivity analysis aims at dividing those opinionated data into subjective and objective sentences and automatic extraction of subjective information from it. Many subjectivity resources as well as subjectivity analysis works are available in English language. In this paper, we examine different methods of generating subjectivity resources in Hindi language and other Indian languages using resources and tools available in English language. Two methods are proposed using word-level subjectivity annotations. These methods use English language OpinionFinder subjectivity lexicon and a small seed word list of Hindi language which can be expanded to generate subjectivity lexicon, respectively. Four methods are proposed using sentence-level subjectivity annotations. These methods use subjectivity annotated corpora and tools available in English language. Different evaluation strategies are used to validate the generated lexicon and corpora in Hindi language. The simulations conducted confirm that these methods are effective in rapidly creating subjectivity resources in Hindi language and other Indian languages.

General Terms

Text Mining, Subjectivity Analysis

Keywords

Data Mining, Text Mining, Subjectivity Analysis, Hindi Language, Natural Language Processing.

1. INTRODUCTION

World Wide Web contains two types of textual information: facts and opinions [26]. Facts refers to objective statements whereas opinions refers to subjective statements. The mental and emotional states (opinions, sentiments, attitudes) of people towards entities, events and their properties are called as private states and private states are represented by subjective statements. The development of Web 2.0 made it easier to post and access opinion on any topics in digital platforms. These opinions can be used in many ways, for example, to make various decisions and predictions on products or services usage. It is difficult to detect and extract these opinionated information from the objective facts. Hence subjectivity analysis is required which can perform the task automatically.

Subjectivity analysis is an important area of research in Natural Language Processing (NLP). Subjectivity analysis refers to the identification and extraction of subjective information from the available input [29]. Blogs, interaction forums, news, speech, tweets, product and movie reviews, question answering are some of the sources of subjective statements [36], [31]. Below are some examples where subjective and objective parts of the sentence is present together in a news article and subjective part is shown in bold:

1. Nuns **erupt in joy** as Mother Teresa is declared saint.
2. This year, due to high production of lentils, lentils prices have gone down by Rs. 10 in each category and this made public **happier than before**.

There are many applications and tools available for the subjectivity analysis in English language but only 26.3% of Internet users speak English[1]. Due to Internet and network penetration to remote and rural areas, world is connected more than ever before. Nowadays, Internet and mobile users are using their native language for the communication. Many websites, blogs, e-books, news portals are available in native languages. People can make decisions about the products or movies based on the reviews available in their native language. There is a huge non English content available on the web so it is required to perform subjectivity analysis in other languages as well. At the initial stage, the subjectivity analysis in other languages can be performed using already existing resources and tools available in English. Later on, identification and use of language specific clues can improve the results. This paper performs

[1] http://www.internetworldstats.com/stats7.htm, June 30, 2016.

this initial task and generate resources in Hindi (target) language which can be extended to other similar Indian languages. However it has some limitations. The translation methods are static, dependent on the machine translation or dictionary and can not be easily expanded/improved. The meaning of the translated data may not be satisfactory. The content may contain translation errors and the lexical syntax may not be good. It is lacking any additional language specific information. Despite these limitations, our results show that the difference in accuracy between translated English resources and language-specific resources is not very high as compared to the time and resources spent on creating it.

1.1 Motivation

Hindi is the most widely used language in Indian sub continent. Hindi and other Indian languages have their origin from either Sanskrit or Dravidian language. More than 310 million people use various Indian languages across the world consisting of 4.45% of the world population[2]. Lexical structure is not much different in the Indian languages and similarity in the usage of words, expressions, polarity, dialects does exist. Huge collection of Hindi and Indian languages data is available in the internet, websites and other web repositories in the areas of culture, sports, history, politics, climate, etc.. There is a need to mine these resources which will provide a path to perform research on other similar languages.

Some Subjectivity analysis related work is available in Hindi, Telugu, Tamil, Marathi and Bengali languages. But subjectivity analysis on Hindi and other Indian language is limited, mainly because of lack of tools and lexical resources. Recent introduction of unicode standard (utf-8) for non-English languages and high cost involved in the corpora creation delayed the research in these languages. A large collection of annotated corpora is essential for Natural Language Processing and sentiment analysis tasks. The existing available corpora is not sufficient to perform research work and large language-specific experiments. Hence, we propose an approach of generating multilingual subjectivity analysis resources in Hindi language using English as a source language.

1.2 Contribution

In this paper, we address the problem of generation of subjectivity resources at word-level and sentence-level in Indian languages. For word-level subjectivity resources, we propose two different methods using subjectivity annotated lexicon. The resources required for the first method are a subjectivity annotated lexicon in English language and a bi-lingual dictionary. The resources required for the second method are a selective small list of subjective word seeds in Hindi language, a bi-lingual dictionary and a raw corpus. For sentence-level subjectivity resources, we propose four different methods using tools and corpora available in English language. The resources required for these methods are a subjectivity annotated corpus in English language, a parallel corpus in Hindi language, machine translation engines to translate the English language documents into Hindi language documents and an automatic subjectivity annotation tool. We run simulations on Hindi language data to show the adaptability of the investigated methods.

1.3 English Language Resources

Following English language resources are used to perform subjectivity analysis in Hindi language:

OpinionFinder Subjectivity Lexicon. We use a subjectivity lexicon which is present in OpinionFinder [37], an English language subjectivity analysis system which classifies a sentence based on the existence (or non-existence) of words or phrases in a large lexicon. After avoiding multi word expressions, the lexicon contains 8221 entries of length 1 word. These words belong to either *strong* subjective type or *weak* subjective type, depending upon the effect its presence has on the context of full sentence. For example, one entry from the *OpinionFinder* lexicon is *type=strongsubj len=1 word1=abase pos1=verb stemmed1=y polannsrc=ph mpqapolarity=strongneg*, indicating that the word *abase* is a verb, strong clue of subjectivity and with a strongly negative polarity. As shown, each entry is taged with a polarity tag and has other information, such as length, stemming (stemmed=y or n), source, and hence forth.

OpinionFinder. OpinionFinder [37] is a subjectivity analysis system that processes English documents and illustrates the subjectivity of a new text based on the existence (or non-existence) of subjective words or phrases in a large lexicon. The classifier labels sentences in the document as subjective or objective based on a model trained on the MPQA Corpus.

MPQA. MPQA Opinion Corpus version 2.0 [36] contains 692 English-language documents, a total of 15802 sentences. It is a collection of 5 different sets of documents: MPQA original subset, XBank, OpQA (Opinion Question Answering) subset, ULA (Unified Linguistic Annotation) and ULA-LU (Language Understanding subcorpus). The Xbank, ULA, and ULA-LU data as well as some documents of the original 535-document release carry attitude and target annotations. The documents are from 187 different foreign and U.S. news sources on politics, sports, health, war, human rights etc. which are manually annotated for subjectivity. Here we use the sentence-level annotations associated with the data set [35].

hindencorp. hindencorp is a parallel corpus in Hindi and English language [7], having 132,300 sentences in Hindi and English. Corpus is collected from various web resources such as tides (DARPA-TIDES contest in 2002) which provided 50,000 sentences and preprocessed by IIIT Hyderabad for NLP contest, Daniel pipes news article commentaries in English and their Hindi translation, EMILLE corpus in English and its Hindi translation, and other smaller datasets from wikipedia and other resources. Here we use a subset of this corpus with 10478 sentences. This corpus is balanced in nature, covering several areas in sports, politics, education and others.

1.4 Organization

The paper is presented as follows: Section II provides a study of the related work. Section III describes two methods of generating subjectivity lexicon using word-level subjectivity annotations in Hindi language and evaluates their quality. Section IV describes four methods of generating subjectivity corpus using sentence-level subjectivity annotations in Hindi language and evaluates their quality. Conclusion is given in section V.

2. RELATED WORK

For subjectivity analysis in any language, corpora and lexicon play a very important role. Several methods and directions are proposed by many researchers by using tools and techniques available in English language. We divide the study of subjectivity analysis task into two groups: Non-Indian and Indian.

Non-Indian languages Research

[2] http://en.wikipedia.org/wiki /List_of_languages_by_number_of_native_speakers

In 1966, General Inquirer System was developed by IBM [32]. IBM called it as content analysis research problem in behaviour science and contains 11,789 words and each having at-least one instance. In 1998, Hatzivassiloglou and McKeown [14] proposed a method of predicting semantic orientations of adjectives, that is, in the usage of <adjectives> and <adjectives>, the adjectives must be of same polarity and this can be used for extracting the patterns. They achieved 82% accuracy. In 2002, Turney [33] extended the work of Hatzivassiloglou and McKeown [14] for other POS-tags and introduced five extraction patterns for polarity classification in review mining. They obtained 84% accuracy on automobile review data and 66% on movie reviews. In 2006, E. suli and Sebastiani developed SentiWordNet [13], [1] which has approximately 2 million words of four Part-of-Speech tags namely adjectives, adverbs, verbs and nouns. In SentiWordNet, each word is having positive, negative and objective scores with total equals to 1. WordNet and a ternary classifier were used to build SentiWordNet.

Using a selected small list of 60 words, an online dictionary and a small annotated corpus, Banea et. al. [3] created a subjective lexicon in Romanian language by bootstrapping method. LSA, a word level similarity is used to filter words. Kamps et. al. [22] developed a distance measure on WordNet to figure out the polarity of adjectives. With an approximation of 67.18% for English, they populated total 1608 words in all four categories. Paper [25] gave a method of determining/analysing judgement opinions in a four step process. The first step, identifying the opinion; the second step, identifying the valence; the third step, identifying the holder and last step, identifying the topic.

Rao and Ravichandran [30] proposed polarity detection problem and affirmed that a word can be classified as bipolar. They introduced semi-supervised label propagation in a graph. Each word is represented by a node with label positive or negative, which determines its polarity and to represent a relationship between two words, they were encoded by a weighted edge. They used English, French and Hindi languages but affirmed that the same methodology can be applied to any language which has WordNet. Paper [11], [5], [12] used Genetic Algorithms for association rule mining.

Indian languages Research

Very few works are available for Indian languages. Paper [8], steered a computational method for developing SentiWordNet(Bengali) exploiting English-Bengali bilingual dictionary and English Sentiment Lexicons. They successfully got 35,805 Bengali words by applying lexical-transfer technique at word level to each word in English SentiWordNet exploiting an English-Bengali Dictionary to get a Bengali SentiWordNet. Das and Bandopadhya [9], introduced four approaches to find the polarity of a word. An interactive game is proposed to find the polarity of words in first strategy. In second strategy, a bilingual dictionary is proposed for English and Indian Languages. In third strategy, word net expansion is proposed using antonym and synonym relations. In fourth approach, a pre-annotated corpus is employed for learning. Paper [10], proposed the method for tagging exploiting the Bengali words. Classification of words is done into six emotion classes according to three categories of intensities (low, general and high). Joshi et al. [21] used two lexical resources: English-Hindi WordNet Linking [23] and English SentiWordNet and created H-SWN(Hindi-SentiWordNet). They substituted words in English SentiWordNet with synonymous Hindi words to get H-SWN using WordNet linking.

Kim and Hovy [24] did the experimentation on sentiment analysis, however their work was restricted to synonyms. In paper [28], Hindi Subjective Lexicon and Hindi WordNet has been used for the identification of linguistics orientation of adjectives and adverbs. By employing a graph based method Bakliwal et al. [2] created subjectivity lexicon. Namita Mittal et al. [27] proposed an approach based on negation handling and discourse relation to determine the emotions from Hindi content. They proposed an annotated corpus for Hindi language and improved present Hindi SentiWordNet (HSWN) by adding more opinion words to it. Their proposed algorithm obtained approximately 80% accuracy on classification of reviews. Jha et al. [17] developed an opinion mining system in Hindi for Bollywood movie review data set. They obtained an overall accuracy of 87.1% for classifying positive and negative documents. Paper [15] performed sentence level subjectivity analysis. They achieved approximately 80% accuracy in classification on a parallel data set in English and Hindi having 71.4% agreement with human annotators. Jha et al. [19] proposed a sentiment aware dictionary in Hindi language for multi-domain data. Paper [16] proposed a stopword removal algorithm for Hindi Language which is based on a Deterministic Finite Automata (DFA). They achieved 99% accurate results. Paper [18] proposed a reputation system for evaluating trust among all good sellers of eBay website and able to rank the sellers efficiently.

3. USING WORD-LEVEL SUBJECTIVITY ANNOTATIONS

Sentiment and subjectivity analysis [25], [31], [38] starts with manually or semi-automatically constructed lexicons. In this paper, the first method is creation of a subjectivity lexicon by translation and given in subsection A. The second method is creation of a language specific subjectivity lexicon by expansion of the selected seed-list and given in subsection B. Both these methods are evaluated on the parallel dataset by constructing a rule-based classifier to classify the sentences into subjective and objective and can be extended to other Indian languages also.

3.1 Creating Subjectivity Lexicon by translation

Translation of an existing English language lexicon by using a dictionary or translator is the most common approach for creating subjectivity lexicon. Here, we used English subjectivity lexicon from *OpinionFinder* and translated it using translator[3] as well as English-Hindi bilingual online dictionary[4] for constructing a subjectivity lexicon for Hindi language. Inflected words present in English subjectivity lexicon made the translation process, a challenging task. The words whose inflected form is available in either translator[3]] or in dictionary[4]], those words are translated as it is. Only for the words whose inflected form translation is not available, they are first lemmatized and then translated. *OpinionFinder* lexicon has 8221 entries but the translated lexicon has 6323 entries. This is because translation of few words are repeated due to different forms of the word and for few words, Hindi translation is not available so these words are discarded from the list. Table I shows a sample from Hindi lexicon along with their English original form.

3.2 Creating Subjectivity Lexicon by Expansion

First, we randomly selected 60 seed words from the translated *OpinionFinder* dictionary. Here, Noun, Verb, Adjective and Adverbs categories are considered and each type has 15 words. Thus, the primary seed list is balanced and helps in good coverage of

[3] https://translate.google.co.in/
[4] http://www.shabdkosh.com/

Table 1. : A Sample of Hindi Subjectivity Lexica

English Word	Associated attributes	Hindi Word
aberration	strongsubj, adj, negative	पतन
abidance	strongsubj, noun, positive	पालन
absence	weaksubj, noun, negative	नदारद
understand	strongsubj, verb, positive	जानना
exclusively	weaksubj, adj, neutral	केवल
loot	strongsubj, verb, negative	लूटना

each part of speech category. Table II shows an example of the seed words.

Algorithm 1: Subjectivity Lexicon Expansion

```
Data: Initial Seed Word List
Result: Expanded Seed Word List
begin
    Initialize:
    words = [];
    related_words = [];
    bi_gram_word_formed = "";
    final_bi_gram = [];
    Dictionary = Dict();
    Perform:
    Parse the seed_word and related_words
    for each line in a document do
        word.append(seed_word)
        related_word.append(related_word)
        Dictionary[seed_word] = related_word
    end
    for each word in words do
        for each related_word in related_words do
            bi_gram_word_formed += word + related_word
            final_bi_gram.append(bi_gram_word_formed)
        end
    end
    Calculate PMI score with the generated bi_grams and add the
    words with PMI score > 0
end
```

The procedure for expansion of seed list into a fully-developed language-specific lexicon is given in algorithm 1. All related words for each seed word are collected from publicly available Hindi Wordnet [6]. These related words are synonyms, antonyms or any other word present in the definition of the seed word. We calculated the Pointwise Mutual Information (PMI) score for filtering some of the related words whose score is zero that means these are not occurring in the corpus. For calculating PMI score, we have collected a large corpus of Hindi files. The statistics for this is given in table III. After each new word is added in the initial seed word list, the above explained process is repeated to collect more subjective words. At the end, we have a collection of 4320 Hindi language subjective words which acts as dictionary.

Table 2. : A Sample of the Seed Words

Adjective *(translation in English)*	भयंकर *(fearful)*, खराब *(horrible)*, बंजर *(barren)*, असभ्य *(barbaric)*, दिवालिया *(bankrupt)*, तुच्छ *(scant)*
Adverb *(translation in English)*	लाभप्रद *(gainfully)*, पूर्णतया *(absolutely)*, केवल*(exclusively)*, उत्साह*(zealously)*, परिहासपूर्वक *(facetiously)*, काफी *(considerably)*
Noun *(translation in English)*	संन्यास *(renunciation)*, कष्ट *(torment)*, भाग्य *(luck)*, शांति *(peacefulness)*, उदारता *(nobleness)*, इच्छा *(desirability)*
Verb *(translation in English)*	आरोप *(impeach)*, निश्चयपूर्वक *(allege)*, लूटना *(loot)*, जानना *(understand)*, बढ़ाना *(multiply)*, सूखना *(dwindle)*

Table 3. : Hindi Corpus Statistics

Type of File	Number	Size	Sentences	Words
Hindi Novels[5]	10 files	9 MB	~1430 sentences per file	~10 words per sentence
Hindi Movie Reviews [17]	200 files	1.6 MB	~50 sentences per file	~8 words per sentence
Hindi Wikipedia [4]	506 files	42.8 MB	~700 sentences per file	~10 words per sentence

3.3 Evaluation

In this section, we want to evaluate that out of these two methods of lexicon generation, which one is giving better results in subjectivity analysis. For evaluation purpose, we are using 501 sentences of *hindencorp* as test dataset. These sentences are different from the sentences used as training dataset in the corpus based methods for sentence-level subjectivity annotations.

Evaluation strategy is given in algorithm 2. The English (Hindi) input file is first parsed at the sentence level and then at word level. When there is a match between the parsed word and the word present in the English (Hindi translated) *OpinionFinder* dictionary then its *word_type* is checked. If it is strong subjective type then its *strong_subj_words_count* is maintained. Similarly *weak_subj_words_count* is also maintained. If one strong subjective word occurs then the sentence is labelled as subjective sentence. For weak subjective words, sentences are labelled as subjective if its occurrence is two.

For language specific lexicon, dividing the subjective words into strong and weak types are difficult so the occurrence of each subjective word turns the sentence into a subjective sentence. The comparative results obtained by both the methods on 501 Hindi sentences and on 501 English sentences using *OpinionFinder* dictionary is shown in table IV.

As shown from table IV, on a parallel test dataset, *OpinionFinder* lexicon gives 38 subjective sentences and if we consider these results are accurate, language-specific lexicon is giving better accuracy with accuracy percentage as 92.1% than translated lexicon accuracy which is 84.2%.

[5] http://www.hindinovels.net/

Algorithm 2: Subjective or Objective Classification

```
Data: Input Data file and OpinionFinder dictionary
Result: Sentences labelled as Subjective or Objective and their
        count
begin
    Initialize:
    strong_subj_words_count = 0;
    weak_subj_words_count = 0;
    strong_subjective = [];
    weak_subjective = [];
    objective = True;
    Perform:
    Parse each sentence from the input file
    for each word in sentence do
        if word in dictionary then
            if wordtype in dictionary is strongsubj then
                strong_subj_words_count += 1
                if strong_subj_words_count > 0 then
                    objective = false
                end
            else
                if wordtype in dictionary is weaksubj then
                    weak_subj_words_count += 1
                    if weak_subj_words_count > 1 then
                        objective = false
                    end
                end
            end
        end
    end
    return objective
end
```

Table 4. : Result of Subjectivity Analysis using Lexicon

Lexicon	Subj Sentences	Obj Sentences
OpinionFinder Lexicon	38	463
Translated Lexicon	32	469
Language-Specific Lexicon	35	466

4. USING SENTENCE-LEVEL SUBJECTIVITY ANNOTATIONS

In this section, we investigate different ways of generating subjectivity corpora in Hindi language using sentence-level subjectivity corpora available in English language. We proposed two scenarios here, one by using manually annotated corpora and another by using automatically annotated corpora.

In the first scenario, we assume that a corpus is available in English language which is manually annotated for subjectivity and we can use the machine translation of this corpora in the required language and project the subjective and objective labels to it. We have implemented this in method 1 using manually annotated corpora of *MPQA* Opinion Corpus version 2.0 [36].

In the second scenario, we assume that an automatic subjectivity analysis tool is available in English language which can be used with manually or automatically generated parallel text to create a subjectivity annotated corpus in the required language. We have implemented this in method 2, 3 and 4 using *OpinionFinder* subjectivity analysis system.

4.1 Creating Subjectivity Corpora using Manually Annotated Corpora

Method 1: Machine translation. The input for this method is manually annotated corpora of *MPQA* Opinion Corpus version 2.0 [36]. This corpora is translated into Hindi language using translator^3] and subjectivity labels are projected from English language to Hindi language corpora. In this way, we have created a large corpus of 904 files (454 objective files and 450 subjective files) with 72,320 sentences, which can be used to train subjectivity classifier in Hindi language.

4.2 Creating Subjectivity Corpora using Automatically Annotated Corpora

Method 2: Manually translated parallel text. The input for this method is *hindencorp* parallel data with 10,478 sentences. English language sentences are passed to high-coverage *OpinionFinder* classifier which gives automatically annotated subjective/objective labels. These labels from *OpinionFinder* classifier is projected to Hindi language parallel text sentences and we have created 10,478 subjectivity annotated sentences in Hindi language which can train subjectivity classifier in Hindi language.

Method 3: Machine translation of English language training data. The input for this method is raw corpora of *MPQA* Opinion Corpus version 2.0 [36]. This corpus is passed to *OpinionFinder* classifier for automatic annotations. These annotations are used to separate the subjective and objective sentences from the corpus. The extracted subjective and objective sentences in English are translated into Hindi using translator^3]. The resulting Hindi data are used to train subjectivity classifier in Hindi language.

Method 4: Machine translation of Hindi language training data. The input for this method is raw corpora in Hindi language [17]. This corpus is first translated into English language using translator^3] and then passed to *OpinionFinder* classifier for automatic annotations. These annotations are projected back into Hindi language and in this way, we get Hindi data for training subjectivity classifier.

4.3 Evaluation

For evaluation purpose, we used same 501 sentences of *hindencorp* as test dataset. The corpus generated by method 1, 2, 3 and 4 are the different training sets and we are trying to find out, which method is giving better results using Support Vector Machine (SVM) classifier.

SVM classifier [20], [34] is based on the concept of decision planes that define decision boundaries using machine learning approach. Here, we provide document-label pair and the decision plane separates the document into different classes with labels. It is a perfect example of linear classifier. Here, the goal is to separate the document into subjective and objective classes by hyperplane with maximum-margin.

We train the classifier with training dataset generated with methods 1, 2, 3 and 4 and apply it to the common test dataset of 501 sentences and obtained the results as shown in table V.

The results show that, using manually annotated corpora for creation of subjectivity corpora is not performing as good as using automatic annotated corpora. This might be possible because human use different cues to express subjectivity and a classifier can

Table 5. : Result of Subjectivity Analysis using Corpora

Methods		All	Subjective	Objective
Method 1	Precision	71%	18%	83%
	Recall	71%	78%	23%
	F1	71%	30%	37%
Method 2	Precision	76%	86%	83%
	Recall	76%	7%	100%
	F1	**76%**	12%	**90%**
Method 3	Precision	73%	21%	85%
	Recall	73%	65%	45%
	F1	73%	31%	59%
Method 4	Precision	72%	19%	94%
	Recall	72%	98%	7%
	F1	72%	**32%**	13%

not be trained on these cues. On the other hand, automatic annotated corpora is robust to translation and gives better classification results. Among the three methods employed on automatic annotated corpora, using parallel text is giving better result because it is free from translation errors.

5. CONCLUSIONS

In this paper, we investigated different methods of generating subjectivity resources in Hindi language and can be extended to other Indian languages. Here, we are summarizing our findings.
We explored two methods for generating subjectivity lexicon and we have found that language-specific lexicon is giving better results for the subjectivity analysis than translating a fully developed *OpinionFinder subjectivity lexicon* in the required language. The difference in accuracy is almost 8%. This supports the time and effort required to grow the lexicon from a small selected seed list.
We explored four methods for generating subjectivity corpora and found that using automatic annotated corpora is giving better results with available parallel text in Hindi and English language. This is mainly because translation errors can be avoided in this method.

6. REFERENCES

[1] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[2] Akshat Bakliwal, Piyush Arora, and Vasudeva Varma. Hindi subjective lexicon: A lexical resource for hindi polarity classification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, 2012.

[3] Carmen Banea, Janyce M Wiebe, and Rada Mihalcea. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. 2008.

[4] Alberto Barrón-Cedeño, Andreas Eiselt, and Paolo Rosso. A comparison of models over wikipedia articles revisions. *ICON*, 2009, 2009.

[5] Veena H Bhat, Prasanth G Rao, R V Abhilash, P Deepa Shenoy, K R Venugopal, and L M Patnaik. A data mining approach for data generation and analysis for digital forensic application. *IACSIT International Journal of Engineering and Technology*, 2(3):314–319, 2010.

[6] Pushpak Bhattacharyya. Indowordnet. In *In Proc. of LREC-10*. Citeseer, 2010.

[7] Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

[8] Amitava Das and Sivaji Bandyopadhyay. Sentiwordnet for bangla. *Knowledge Sharing Event-4: Task*, 2, 2010.

[9] Amitava Das and Sivaji Bandyopadhyay. Sentiwordnet for indian languages. *Asian Federation for Natural Language Processing, China*, pages 56–63, 2010.

[10] Dipankar Das and Sivaji Bandyopadhyay. Labeling emotion in bengali blog corpus–a fine grained tagging at sentence level. In *Proceedings of the 8th Workshop on Asian Language Resources*, page 47, 2010.

[11] P Deepa Shenoy, K G Srinivasa, K R Venugopal, and Lalit M Patnaik. Evolutionary approach for mining association rules on dynamic databases. In *Advances in knowledge discovery and data mining*, pages 325–336. Springer, 2003.

[12] P Deepa Shenoy, K G Srinivasa, K R Venugopal, and Lalit M Patnaik. Dynamic association rule mining using genetic algorithms. *Intelligent Data Analysis*, 9(5):439–453, 2005.

[13] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

[14] Vasileios Hatzivassiloglou and Kathleen R McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

[15] Vandana Jha, N Manjunath, P Deepa Shenoy, and K R Venugopal. Hsas: Hindi subjectivity analysis system. In *2015 Annual IEEE India Conference (INDICON)*, pages 1–6. IEEE, 2015.

[16] Vandana Jha, N Manjunath, P Deepa Shenoy, and K R Venugopal. Hsra: Hindi stopword removal algorithm. In *Microelectronics, Computing and Communications (MicroCom), 2016 International Conference on*, pages 1–5. IEEE, 2016.

[17] Vandana Jha, N Manjunath, P Deepa Shenoy, K R Venugopal, and L M Patnaik. Homs: Hindi opinion mining system. In *Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on*, pages 366–371. IEEE, 2015.

[18] Vandana Jha, R Savitha, P Deepa Shenoy, and K R Venugopal. Reputation system: Evaluating reputation among all good sellers. In *Proceedings of NAACL-HLT*, pages 115–121, 2016.

[19] Vandana Jha, R Savitha, Sudhashri S Hebbar, P Deepa Shenoy, and K R Venugopal. Hmdsad: Hindi multi-domain sentiment aware dictionary. In *2015 International Conference on Computing and Network Communications (CoCoNet)*, pages 241–247. IEEE, 2015.

[20] Thorsten Joachims. *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[21] Aditya Joshi, AR Balamurali, and Pushpak Bhattacharyya. A fall-back strategy for sentiment analysis in hindi: a case study. *Proceedings of the 8th ICON*, 2010.

[22] Jaap Kamps, Maarten Marx, Robert J Mokken, and Maarten De Rijke. Using wordnet to measure semantic orientations of adjectives. In *LREC*, volume 4, pages 1115–1118. Citeseer, 2004.

[23] Arun Karthikeyan Karra, Prabhakar Pande, Rohan Railkar, Aditya Sharma, and Pushpak Bhattacharyya. Hindi english wordnet linkage. 2009.

[24] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

[25] Soo-Min Kim and Eduard Hovy. Identifying and analyzing judgment opinions. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 200–207. Association for Computational Linguistics, 2006.

[26] Bing Liu. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666, 2010.

[27] Namita Mittal, Basant Agarwal, Garvit Chouhan, Nitin Bania, and Prateek Pareek. Sentiment analysis of hindi review based on negation and discourse relation. In *Sixth International Joint Conference on Natural Language Processing*, page 45, 2013.

[28] Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande, and Pushpak Bhattacharyya. An experience in building the indo wordnet-a wordnet for hindi. In *First International Conference on Global WordNet, Mysore, India*, 2002.

[29] Bo Pang and Lillian Lee. 4.1.2 subjectivity detection and opinion identification. *Opinion mining and sentiment analysis*, 2008.

[30] Delip Rao and Deepak Ravichandran. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 675–682. Association for Computational Linguistics, 2009.

[31] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112. Association for Computational Linguistics, 2003.

[32] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer: A computer approach to content analysis. 1966.

[33] Peter D Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[34] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[35] Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497. Springer, 2005.

[36] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

[37] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. Opinionfinder: A system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations*, pages 34–35. Association for Computational Linguistics, 2005.

[38] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics, 2003.

7. AUTHOR'S PROFILE

Vandana Jha obtained her Bachelor of Engineering in Computer Science and Engineering from Maharshi Dayanand University, Gurgaon, India in 2003. She received her Masters of Technology specialized in the field of Computer Science and Engineering from Kuvempu University, Karnataka, India in 2009. Currently she is working as Research Scholar in the Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore, India. Her research interests include Information Retrieval, Data Mining, Opinion Mining and Web Mining.

Shreedevi G. R. has obtained B.E in Computer Science and Engineering(CSE) and pursuing M.E in CSE from University Visvesvaraya college of Engineering. Her research interest are Data Mining, Opinion Mining and Subjectivity Analysis.

P. Deepa Shenoy is currently working as Professor in the Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore, India. She did her doctorate in the area of Data Mining from Bangalore University in the year 2005. Her areas of research include Data Mining, Soft Computing, Biometrics and Social Media Analysis. She has published more than 150 papers in refereed International Conferences and Journals.

K. R. Venugopal is currently the Principal, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. He obtained his Bachelor of Engineering from University Visvesvaraya College of Engineering. He received his Masters degree in Computer Science and Automation from Indian Institute of Science Bangalore. He was awarded Ph.D. in Economics from Bangalore University and Ph.D. in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored and edited 70 books on Computer Science and Economics, which include Petrodollar and the World Economy, C Aptitude, Mastering C, Microprocessor Programming, Mastering C++ and Digital Circuits and Systems etc.. He has filed 100 Patents. During his three decades of service at UVCE he has over 500 research papers to his credit. His research interests include Computer Networks, Wireless Sensor Networks, Parallel and Distributed Systems, Digital Signal Processing and Data Mining.