5-2018

Sentiment Analysis on Financial News and Microblogs

Chinmay Talekar
Purdue University

SENTIMENT ANALYSIS ON FINANCIAL NEWS AND MICROBLOGS

by

Chinmay Talekar

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Computer & Information Technology

West Lafayette, Indiana

May 2018

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Julia Rayz, Chair

Purdue Polytechnic Institute

Dr. Baijian Yang

Purdue Polytechnic Institute

Dr. John Springer

Purdue Polytechnic Institute

Approved by:

Dr. Eric Matson

Head of the Graduate Program

I dedicate my thesis work to my family and my friends. A special feeling of thanks to my parents, Vijay and Padmaja Talekar who have believed in me and inspired me to constantly keep working hard, stay strong in difficult times and never give up. Also, to my sister Apoorva, without whose guidance, I would not be a student at Purdue University

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Julia Rayz for her guidance and time invested in my me and my work starting from getting admitted to this program, CNIT 581 - Natural Language Technologies, independent study, weekly and lab meetings. Thanks to Dr. Yang and Dr. Springer for agreeing to be on my committee and providing guidance.

I would like to acknowledge and thank the Department of Computer and Information Technology for providing me with this opportunity to expand my knowledge, conduct my research and provide any assistance requested. I would also like to thank Prof. Ravai and Dr. Mohler for their continued support.

Finally, I would like to thank my friends and family, their support has made my research an enjoyable and rewarding experience.

TABLE OF CONTENTS

GLOSSARY

“Natural language processing is the study of mathematical and computational modeling of various aspects of language and a development of a wide range of systems.” (Joshi, 1991).

“Sentiment analysis, also known as opinion mining, is a Natural Language Processing task aimed at the automatic identification and analysis of people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes in text.”(Liu, 2012)

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Networks
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
SST	Stanford Sentiment Treebank
GloVe	Global Vector for Word Representation
NER	Named Entity Recognition
PV-DM	Distributed Memory Model of Paragraph Vectors
PV-DBOW	Paragraph Vector: Distributed bag of words
CBOW	Continuous Bag of Words

ABSTRACT

Author: Talekar, Chinmay, V. MS.
Institution: Purdue University
Degree Received: May 2018.
Title: Sentiment Analysis on Financial News and Microblogs.
Committee Chair: Dr. Julia Rayz.

Sentiment analysis is useful for multiple tasks including customer satisfaction metrics, identifying market trends for any industry or products, analyzing reviews from social media comments. This thesis highlights the importance of sentiment analysis, provides a summary of seminal works and different approaches towards sentiment analysis. It aims to address sentiment analysis on financial news and microblogs by classifying textual data from financial news and microblogs as positive or negative. Sentiment analysis is performed by making use of paragraph vectors and logistic regression in this thesis and it aims to compare it with previously performed approaches to performing analysis and help researchers in this field. This approach achieves state of the art results for the dataset used in this research. It also presents an insightful analysis of the results of this approach.

CHAPTER 1. INTRODUCTION

1.1 Background

Sentiment analysis is important because it helps us to identify the opinion of people on any given topic. Sentiment analysis is also called opinion mining (Daiyan, Tiwari, Kumar, & Alam, 2015). Dave et al. (2003) were the first to use the term opinion mining. Understanding the opinion has multiple applications, which include calculating customer satisfaction metrics, identifying market trends for any industry or product, improving customer satisfaction by making use insights from sentiment analysis, identifying support or favor in political campaigns, identifying competitors, etc. With the growing availability of online resources that are rich in opinion, there has been a rise in information technologies which can be used to seek out and understand the opinion of others (Pang & Lee, 2008).

Microblogging has become a popular tool with internet users with the rise of websites like Twitter, Facebook, and LinkedIn and more people are making use of microblogging websites to express their opinions online (Pak & Paroubek, 2010). As of 2016, the United States of America is the biggest country using Twitter with a user base of 67.54 million monthly active users ("Countries with most Twitter users 2016 | Statistic," 2016). Research carried out in the field of economics has shown that news items and market fluctuations are associated with each other and they can be measured by an increase, decrease or no change in the prices of stocks (Goonatilake & Herath, 2007). Also, people express their sentiment about stock through Twitter microblogs. This expression by people and news authors affects the stock market in a positive or negative manner, which is also called bearish or bullish behavior in stock market terminology (Van de Kauter, Breesch, & Hoste, 2015).

Considering the prominence of Twitter in the United States of America and the growth of the internet and impact of news and microblogs on the stock markets in the United States of America, it would be beneficial to carry out sentiment analysis on financial news and microblogs. Applications of this could be beneficial for researchers in financial domain who study financial markets to optimize the profits.

This chapter provides information about the background, significance, and scope of this research. It highlights the research question. It also states the assumptions, limitations, and delimitations.

1.2 Significance

Sentiment analysis on financial news and microblogs as described in the previous sections can help market investors to identify bearish or bullish sentiment in a stock. By identifying bearish or bullish sentiment, investors can determine which stock may be most profitable. Fidelity investments have highlighted the importance of investing in US stocks ("Reasons To Invest In Stocks - Fidelity," 2017). A research carried out shows mood of the people microblogging can impact the prices of the stock (Smailović, Grčar, Lavrač, & Žnidaršič, 2014). Additionally, another research showed that "changes in public mood reflect the value shift in Dow Jones Industrial Index three to four days later" (Bollen, Mao, & Zeng, 2011). Understanding the link between market dynamics and public sentiment it is important to address this problem (Cortis et al., 2017).

On the other hand, the companies can also identify bullish or bearish sentiment towards their stocks. This can help them decide on a date to launch a product to get a favorable reaction from the stock market or help to understand the expectations of the market. This research can be used to make independent decisions to help decide which stocks to buy, which can be a small component to develop an automatic income generation application. Furthermore, the solution to the problem of detection of sentiment in the financial news and microblogs can be applied to different domains. To conclude, how much information can be extracted from a microblog of news and where it can be used is an open-ended and challenging question. In this thesis, no correlation to temporal dimension of stocks going up or down is made, however, it is shown that it is possible to identify a sentiment of a financial text within a corpus used for sentiment detection competition.

1.3 Scope

While the scope of sentiment analysis is broad, the research question focuses on identifying sentiment only in financial news and microblogs. Financial news and microblogs consist of textual data. Furthermore, quantifying sentiment is a challenging task. In this work, the sentiment is a quantified value that can either be positive, neutral, or negative. The opinion of public and news authors is considered. The impact of news appearing on a front page of a prominent newspaper could be higher than news appearing on pages other than the first page of a newspaper with a smaller reader base. This research assumes the impact is similar irrespective of the prominence of the newspaper and on which page the news was published. Additionally, the data includes news and microblogs that are related to a fixed number of companies and does not derive any conclusions where the news or microblog does not contain names of any companies.

1.4 Research Question

The research area for this thesis is natural language processing (NLP). This research focuses on textual data in the financial domain. The primary research question addressed in this research is: to what degree does doc2vec or paragraph vectors (Le and Mikolov, 2014) can classify sentiment in the financial domain? A sentiment is treated as a quantified value which can be positive, neutral, or negative.

1.5 Assumptions

The assumptions of this study are:

1. The natural language toolkit used to process the news and financial blogs produces accurate results; the results will not be altered unless specified by the researcher.
2. The sentiment score used to quantify the sentiment as positive or negative (a variable in the dataset) is assumed to be an accurate quantification of the financial news or microblog.

1.6 Limitations

The limitations of this study are:

1. The research will use the financial news and microblogs present only in the dataset to conclude about the population and is limited by the size of the dataset.
2. Real-time financial news and microblogs and their impact on the stock market are not considered. The sentiment score in the dataset quantifies the impact on the stock in this research.

3. Unusual words which do not have meaning will not be considered in this research. For example, “I will smash that fav button every time this gets rt'd”. Impact of words like rt’d is not be considered.
4. Sarcasm and irony are not considered in this research.
5. Hashtags are converted to lowercase words and not considered for this research.

1.7 Delimitations

The delimitations of this study are:

1. Analyzing the bearish or bullish sentiment for any company is a broad topic which is not limited to analyzing the sentiment of financial news and microblogs. It can include components such as historical trends, the current state of the economy in which the company is based, financial strategies employed by companies, large investment by investors etc. This study will take only into consideration the effect of financial news and microblogs.
2. Financial news and microblogs in the English language are only considered for this research. This research does not take into consideration the effect of news and microblogs in other languages used in the world.
3. This research does not take into consideration the popularity of the person and the other measures such as indegree, retweets, and user mentions associated with the tweet as described earlier (Cha, Haddadi, Benevenuto, & Gummadi, 2010).
4. For the financial news in the dataset, the reader base of the newspaper worldwide and in the United States of America, the mode of distribution of the newspaper for the news is not taken into consideration for this research.

5. Images accompanied by microblogs and their relationship with the microblog is not considered. Microblogs which are uploaded by users can contain images, the relationship between the image and microblog can imply multiple meanings which can polarize the sentiment of any microblog. For example, “this company got great funding $XYZ” accompanied by an image of a dollar bill implies that the company XYZ does not have good funding.
6. Rumors in news and microblogs are not considered. Emoticons, which can be used in microblogs such as “:)”, “:p”, are not considered.

CHAPTER 2. LITERATURE REVIEW

"NLP is the study of mathematical and computational modeling of various aspects of language and a development of a wide range of systems." (Joshi, 1991). Some of the NLP tasks are speech recognition, summarization, word sense disambiguation, question answering using chatbots, named entity recognition, sentiment analysis etc. "Sentiment analysis, also known as opinion mining, is a NLP task aimed at the automatic identification and analysis of people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes in text." (Liu, 2012).

2.1 Typology of affective states

The typology of affective states defined by Scherer (1984) has five affective states (Harmon-Jones, Gable, & Price, 2013) that are emotion, mood, interpersonal stance, attitudes, and personality traits. According to Scherer, attitudes capture disposition towards objects or persons which can be liking, loving, hating, valuing, hating, and desiring. While sentiment analysis can be used to determine the affective states as described previously it is primarily used to detect attitudes. The simple task of determining sentiment could be classifying the sentence as positive or negative whereas complex tasks could be assigning sentiment scores on a negative or positive scale.

2.2 Approaches to sentiment analysis

This section covers four approaches towards sentiment analysis which are lexicon based, aspect-specific and targeted, Word2vec (shallow neural network) based, and deep learning based.

2.2.1 Lexicon based sentiment analysis

Lexicon based sentiment analysis is one approach to solve the task of sentiment analysis, in which lexicons which may be defined positive, negative, or neutral are used for sentiment analysis. Some of the readily available lexicon lists are General Inquirer (Stone, Dunphy, & Smith, 1966), LIWC (James Pennebaker, Roger Booth, & Martha Francis, 2007), MPQA Subjectivity cues lexicon (Riloff & Wiebe, 2003), Wordnet (Miller, 1995) and Bing and Liu opinion lexicon (Hu & Liu, 2004).

While there are strengths (Hatzivassiloglou & McKeown, 1997; Turney, 2002) in using lexicon based approaches, there are drawbacks (Hamilton, Clark, Leskovec, & Jurafsky, 2016) as well. Research carried out by making use of machine learning techniques using unigrams did not perform well in sentiment classification tasks (Pang, Lee, & Vaithyanathan, 2002). Additionally, research shows that unigram features using Wordnet with part-of-speech labels achieved better results than without part-of-speech labels (Salvetti, Reichenbach, & Lewis, 2006).

Potts's (2010) approach involved considering the likelihood and scaled likelihood rather than raw counts and examined whether logical negation was associated with negative reviews. In this work, he examined how likely each word will occur in each classification class rather than a raw count of the positive or negative word from a defined list and focused on handling negation and negative words. One example to understand this could be "excellent" and "good" both being positive but the likelihood of "excellent" in positive class is much higher than "good".

2.2.1.1 Semi-supervised learning of lexicons

The one assumption of lexicon-based approach is that the lexicon list exists. However, this may not be true in every domain and a generic list described in the previous section may not be useful to sentiment analysis in each domain. This may be one of the problems of lexicon-based approach. So, there is a need to develop a lexicon list for a domain which has limited labeled lexicons. There have been approaches to developing such dynamic lexicon lists by (Hatzivassiloglou & McKeown, 1997; Hovy & Kim, 2004; Turney, 2002).

Hatzivassiloglou & McKeown (1997) made use of conjoined words to form lexicons, they hypothesized that two words occurring with "and" in between have the same polarity whereas word occurring with "but" in between have negative polarity. They make use of search engine to find word occurring with "and" and "but" in between and generate word lexicons. Turney (2002) considered phrases rather than words and made use of a metric called as phrase mutual information and calculated the polarity of the word by using the mutual information between the given phrase and "excellent" minus the mutual information between the the given phrase and "poor". His approach took into consideration the co-occurrence of a phrase with "excellent" and "poor" by making use of point mutual information. His results showed accuracy range between 66% to 84% for reviews of banks, automobiles, movies, and travel locations. S.-M. Kim (2004) work showed that identifying manually identifying seed words and adding synonyms and antonyms into positive or negative sentiment lexicons can lead to encouraging results. His work involved combining sentiment at word and sentence levels.

Approach to bootstrap lexicons has been carried out in medical treatment domain (Xu, Morgan, Das, & Garber, 2009). Their work identifies treatment concepts from 100 randomly chosen abstracts in the medical treatment dictionary. Research has been carried out to induce domain specific lexicons from unlabeled data in multiple domains and compare them words used currently and those used one fifty years ago (Hamilton et al., 2016). Hamilton et al. (2016) showed that use of lexicon in positive or negative polarities has changed significantly in one fifty years, from this we can deduce that use of dynamic lexicons lists is a better alternative than making use of preconstructed lists as polarities of words might change over time.

2.2.1.2 Problems in lexicon based approaches

If sentiment lexicons are extracted and used for sentiment analysis it can lead to certain issues like the word ‘bull’ in the financial domain will have negative polarity as it is related to bullish. Similarly, in a sports domain, the word bull can have positive polarity. For example, in football “He ran like a bull, crushing his opponents”, the word bull will have positive polarity. Words like “bull” can be present in multiple domains and may have positive, negative, or neutral meanings in different domains. Furthermore, such words can be more difficult to handle if sentiment scores are assigned on a negative or positive scale.

Comparison carried out by Aue and Gamon (2005), considers data from movie reviews, book reviews, product support services web data, knowledge-based web survey data to develop a new approach to “customizing a sentiment classification system to a new target domain in the absence of large amounts of labeled data”. Their results suggested that a small amount of labeled target data used with unlabeled data, both from the same domain to learn about model parameters for a generative Naive Bayes classifier achieved the best classification accuracy.

Looking at lexicon based sentiment analysis approaches and semi-supervised learning of lexicons, it can be said that it is important to handle negation in sentences and finding subsets of words is used in domain-specific sentiment analysis. Additionally, it is difficult for a single lexicon list to handle the different meanings of the same word in different domains. The use of semi-supervised learning methods and lexicon lists created using seeding has been used to induce lexicons.

2.2.2 Aspect specific and targeted sentiment analysis

2.2.2.1 Aspect specific sentiment analysis

Aspect specific sentiment analysis was introduced to look at tasks other than classifying sentiment as positive, negative and neutral analysis (Hu & Liu, 2004). Various aspects can be analyzed for any sentence rather than classifying the sentence as positive, negative, or neutral. For example, "The Teaching Assistants were good, the course was okay and the timings of the class were the worst". Valuable information in the sentence about the teaching assistants, the course, and the timings is lost if classification into positive, negative, and neutral is performed. If sentences are analyzed in a way described above various aspects of the teaching assistants, the course, and the timings can be considered. Previous work in this domain was carried out on restaurant and hotel reviews. Aspects of food, décor, service, and value were extracted from restaurant reviews and aspects like rooms, location, dining, service, and value were extracted from hotel reviews (Blair-Goldensohn et al., 2008).

Aspect specific sentiment analysis is valuable as it provides us with the ability to aggregate each aspect and analyze its sentiment. Additionally, Blair-Goldesohn et al. (2008) have presented a complete model that takes input as reviews and gives final summary by the process of text extraction, sentiment classification, aspect extraction, and aggregation. Their work concluded aspect specific sentiment analysis revealed user-provided labels, which can be used in that domain to draw additional insights and there is a need for more research to be required to be carried out in semi-supervised learning methods to find the aspects in a dataset.

2.2.2.2 Targeted sentiment analysis

Welch and Mihalcea (2016) worked on targeted sentiment analysis in which they define as identifying the sentiment which the writer holds towards the targets or entities mentioned in the sentence. They worked on identifying the courses and instructors mentioned in the comments by students on Facebook and the sentiment of students towards these courses and instructors by the two-step process of entity extraction and entity centered sentiment analysis. Furthermore, the task of attributing the sentiment towards each entity is also handled in targeted sentiment analysis. Hu and Liu (2004) concluded in their work that lexicon features extracted from the dataset play a significant role in target specific sentiment analysis.

SentiHood dataset (Welch & Mihalcea, 2016) extracted from online question answering platform where discussions regarding urban neighborhoods took place was introduced to perform target-specific sentiment analysis (Saeidi, Bouchard, Liakata, & Riedel, 2016). It was a pioneer attempt to use generic question answering platform for fine-grained sentiment analysis. They made use of logistic regression and bi-directional Long short-term memory (Hochreiter & Schmidhuber, 1997).

Work similar to target specific sentiment analysis which addresses the sentiment towards attributes has been done for service in a restaurant (Sauper & Barzilay, 2013) and camera of a mobile phone (Chamlertwat, Bhattarakosol, Rungkasiri, & Haruechaiyasak, 2012).

2.2.2.3 Difference between aspect-specific and targeted sentiment analysis

While the aspect specific sentiment analysis (Hu & Liu, 2004) considers the extraction of aspects the entities associated with the sentiment are also extracted in targeted sentiment analysis.

The target-specific sentiment analysis helps us to analyze fine-grained sentiment about the entities which are present in the review or sentence.

2.2.3 Sentiment analysis using embeddings

2.2.3.1 Sentiment analysis using word embeddings

Word embedding is a way to represent words or phrases as a vector of real numbers. Each word has one dimension in the continuous vector space. Word embeddings have helped to boost the performance of sentiment analysis (Socher, Perelygin, et al., 2013) and syntactic parsing (Socher, Bauer, Manning, & others, 2013). The central idea behind word embeddings is that "a word is characterized by the company it keeps" (Firth, 1957) which was popularized by Firth (1957), the concept of word embedding comes from the branch of linguistics called as distributional semantics (Harris, 1954).

Sentiment analysis has also been performed by using Word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and Paragraph Vector (Le & Mikolov, 2014) and results have been promising (Xue, Fu, & Shaobin, 2014). Xue, Fu, and Shaobin (2014) carried out sentiment analysis on a microblogging website called Sina Weibo and created an emotional dictionary of sentiments using word similarity distances obtained from word2vec. If the similarity distance between positive word and input word was positive, then the input word is a positive-attitude word and vice versa. They presented a new Semantic Orientation Pointwise Similarity Distance (SO-SD) model to build a dictionary using Word2vec (text as input and produces the word vectors as output) and emphasized on the requirement to develop context-specific lexicons. Another word embedding technique which has been used is called as GloVe (Pennington, Socher, & Manning, 2014) also called the global vector for word representations. Pennington, Socher, and Manning (2014) state that “a new global log-bilinear regression model for the unsupervised learning of word representations that outperforms other models on word analogy, word similarity, and named entity recognition (NER) tasks.” Apart from applications in natural language processing word embeddings approaches have also been proposed in bioinformatics for biological sequences (Asgari & Mofrad, 2015). Additionally, Improved Word Vectors (Rezaeinia, Ghodsi, & Rahmani, 2017) which improves the accuracy of pre-trained word-embeddings such as word2vec and Glove is also proposed. In their work, they test their word vectors using deep learning models on different sentiment datasets.

2.2.3.2 Sentiment analysis using document embeddings

Notable work by creating a vector representation of entire sentence and using it to classify movie reviews from the IMDB dataset (Maas et al., 2011) was performed at Google (Le & Mikolov, 2014). Their work achieved excellent results over vector averaging, Bag-of-words, Bag-of-bigrams, and Weighted Bag-of-bigrams. In their work, they propose an unsupervised learning algorithm which learns feature vectors for a variable length of texts and mention that paragraph vector overcomes multiple shortcomings of a bag of word models. They mention two algorithms, distributed memory model of paragraph vectors (PV-DM) and Paragraph Vector: Distributed bag of words (PV-DBOW). The PV-DM makes use of the paragraph vector and the words present in the document and either average or concatenates them. The PV-DBOW, on the other hand, can be used predict the words from the paragraph vector, these words could be in random order from the paragraph. They also claim to paragraph vector is effective in capturing the semantics of the sentences. However, there are multiple considerations for their claim when PV-DBOW are used. Another advantage of paragraph vectors is that they can be learned where enough labelled data is not available, and the document has multiple sentences in it. Furthermore, they also describe an approach to information retrieval in the applications of paragraph vectors.

A composition in a distributed model of semantics is also proposed which includes a framework to evaluate phrase similarity task (Mitchell & Lapata, 2010). They describe vector composition by using additive and multiplicative functions. Another approach which involves the use of linear models for compositional distributional semantics was proposed by Zanzotto et al. (2010).

To summarize, paragraph and word vectors both involve the creation and making use fixed length feature vectors for word and paragraphs and words respectively. The dimensions of these fixed length vectors can be specified during their creation. Dimensions for previously created embeddings for words and paragraph range anywhere from 50 to 400 ("Google Code Archive - Long-term storage for Google Code Project Hosting.," 2013; Le & Mikolov, 2014; Pennington et al., 2014). The choice of dimensions of the embedding depends on the application. While higher dimensions capture more information, they increase the computation cost as well. Both variations of paragraph vector algorithms, PV-DM and PV-DBOW make use word vectors. The PV-DM additionally uses paragraph vector in conjunction with word vectors. Word2vec has two algorithms which are Continuous Bag of Words (CBOW) and Skip gram. PV-DM is analogous to CBOW while PV-DBOW is analogous to Skip gram model.

2.2.4 Sentiment analysis using deep learning

Deep learning has been effective to perform the computer vision tasks (Krizhevsky, Sutskever, & Hinton, 2012). Convolutional neural networks (CNN) were primarily developed for computer vision tasks to extract features by using convolving layers (Lecun, Bottou, Bengio, & Haffner, 1998). It has also shown strengths in speech recognition by making use of recurrent neural networks (Graves, Mohamed, & Hinton, 2013). Work done in NLP using deep learning involves making use of word vectors. To create these word vectors the words are projected from a sparse, 1-of-Vocabulary size encoding onto a lower dimensional vector space using a hidden layer. These word vectors are essentially used to extract features that encode semantic features of words in their dimensions.

CNN's have been effective to perform multiple tasks in the field of NLP (Collobert & Weston, 2008). Kim (2014) made use of CNN's by utilizing TensorFlow (a software library for machine learning). The model presented in the paper achieves improves on 4 out of the 7 state-of-the-art classification tasks and has since become a standard baseline for new text classification architectures. The model achieves accuracy on text classification ranging from two to six classes and concludes that with some tuning of hyperparameters a CNN can achieve remarkable results (Kim, 2014). Socher et al. (2013) introduced the method of parsing with compositional vector grammar with the recurrent neural network. However, for its implementation large amount of data is required.

Hong and Fang (2015) demonstrated state-of-art performance on IMDB large movie dataset using paragraph vectors (Le & Mikolov, 2014) and competitive performance on Stanford Sentiment Treebank (SST) dataset (Socher, Perelygin, et al., 2013) using deep recursive neural networks and Long Short-Term Memory (LSTM) for binary and fine-grained classification tasks. Although word vector representations are effective in sentiment analysis (Xue et al., 2014), the semantic meaning of the word is lost in the process. On the other hand, recurrent neural networks help to capture relations between words and LSTM's have the ability to remember sequences of words (Hochreiter & Schmidhuber, 1997). Their best performance using deep recursive neural network performance achieved accuracies of 46.9% on fine (5 – class) and 84.7% on binary classification tasks. In their analysis, they highlighted that the word and paragraph embeddings produced from a smaller dataset are of significantly lower quality than large datasets. LSTM's have also been used for target-specific sentiment analysis (Saeidi et al., 2016).

Zimbra, Ghaiassi and Lee (2016) presented a brand-related twitter sentiment analysis on tweets which were related to Starbucks (a coffee brand) and used Dynamic architecture for the artificial neural network to perform sentiment analysis on three class and five class classification tasks. They achieved 86.06% and 85.56% accuracies on three and five class tweet classification on the Starbucks data set used in their research. The difference between the accuracies for three and five class accuracies suggest considerable potential for the use of dynamic architecture for the artificial neural network in classification tasks where multiple classes are involved.

2.3 Sentiment analysis in the financial domain

Sentiment analysis in financial domain falls under the category of domain-specific sentiment analysis. Previous work which uses textual analysis makes use of lexicon-based lists to measure tone (Loughran & Mcdonald, 2011). From the lexicon-based sentiment analysis as described previously, we can deduce that lexicons are domain specific. Thus, it would not be advisable to use lexicons from a different domain to analyze the sentiment of a financial news or microblog.

The task five of SemEval 2017 was to carry out fine-grained sentiment analysis on the financial news to predict prices of stocks. For that, they released a dataset which consisted of tweets and microblogs. Multiple researchers performed this task all over the world.

The notable research was carried out to examine whether supervised machine learning classifier to predict the sentiment scores of financial tweets provided by SemEval2017 (Tabari, Seyeditabari, & Zadrozny, 2017). Their process involved preprocessing the data, selecting the features which involved handling negation based words and word pair sentiment feature selection and sentiment prediction using random forest algorithm, support vector machines, Naïve Bayes, and logistic regression. They concluded that use of additional features in the feature selection process impacted the accuracy by 6%, use of bigrams reduced the accuracy of random forest and logistic regression and when classification using word embeddings showed less accuracy than the methods described previously. Although their approach did not capture complex linguistic structure, sarcasm and poorly structured sentences in tweets were not handled. Additionally, idioms in the financial domain could be handled effectively.

Research to answer the question of whether word embeddings in combination with CNN's can be used to detect sentiment was carried out (Mansar, Gatti, Ferradans, Guerini, & Staiano, 2017). Their procedure involved sentence preprocessing, word embedding generation, generation of multiple convolutional layers involving convolutional layer, concat layer and regularization layer and a loss function. Their analysis and results showed that use of pre-computed word representations allows reducing overfitting, but some basic pre-processing can improve the performance of this method.

Ghosal et al. (2017) presented an ensemble of deep learning and feature-based approaches where they ensembled the outputs from CNN, LSTM, and vector averaging to achieve the final score. They have also used a lexicon-based approach where they count the positive and negative count of lexicons appearing in tweets and microblogs.

Another work which utilized domain specific lexicons was presented by Nasim (2017), they made use of lexicon list constructed by Loughran and Mcdonald (2011). Their approach involved calculating the word count of the positive and negative word in financial domain, and their model calculated the sentiment score as a total of positive and negative scores in the financial domain. Moore and Rayson (2017) made use of LSTM on the same task and their analysis showed an improvement of 4-6% by making use of bidirectional LSTM. Additionally, they concluded that performance by making use of unigrams was better than making use of bigrams and extended the scope for aspect specific sentiment analysis on this problem using LSTM.

2.4 Summary

This chapter provides you with a background understanding of sentiment analysis, it discusses the approaches to sentiment towards sentiment analysis such as lexicon-based sentiment analysis, aspect specific, targeted sentiment analysis, sentiment analysis using word embeddings (word2vec and Glove) and document embeddings and deep learning approaches towards sentiment analysis. It also discusses merits and de-merits of some of the approaches. Additionally, it also discusses how sentiment analysis has been applied in the financial domain on the dataset by notable researchers from the contributions in the SemEval 2017 competition. The contributions include ensemble machine learning approaches which include feature engineering on text, CNN, LSTM, and count of positive and negative words present in the sentences (a type of lexicon based approach) using defined algorithms.

CHAPTER 3. METHODOLOGY

This chapter defines the type of research, the framework, the population, the sampling approach, the sample size, the unit of measurement, the variables, and the analysis of the data. This chapter also provides additional information which requires further investigation and study. This research is quantitative research, the population in this research is the financial news and microblogs. This research methodology tries to identify the relation between financial news and its sentiment by overlooking other lurking variables. The chapter also provides information about data and pre-processing required for implementation, the parameter required for training the paragraph vector model and the detailed step involved in training, prediction, and evaluation using cosine similarity.

Figure 3.1 depicts all the steps involved in the methodology such as data processing, model creation, training, prediction, and evaluation. The dataset used is obtained from the SemEval 2017 Task 5. The dataset contains two subsets, one for the news headlines and other for the microblogs. The dataset goes through a series of pre-processing steps after which paragraph vectors are trained on all the news headlines and microblogs separately. This leads to the creation of two models which contain the document and word embeddings in each model. The document embeddings are used to train the classifier along with the training split and the labels are predicted using the classifier for the testing split of the dataset. The cosine similarity is then evaluated for the testing and the predicted labels.

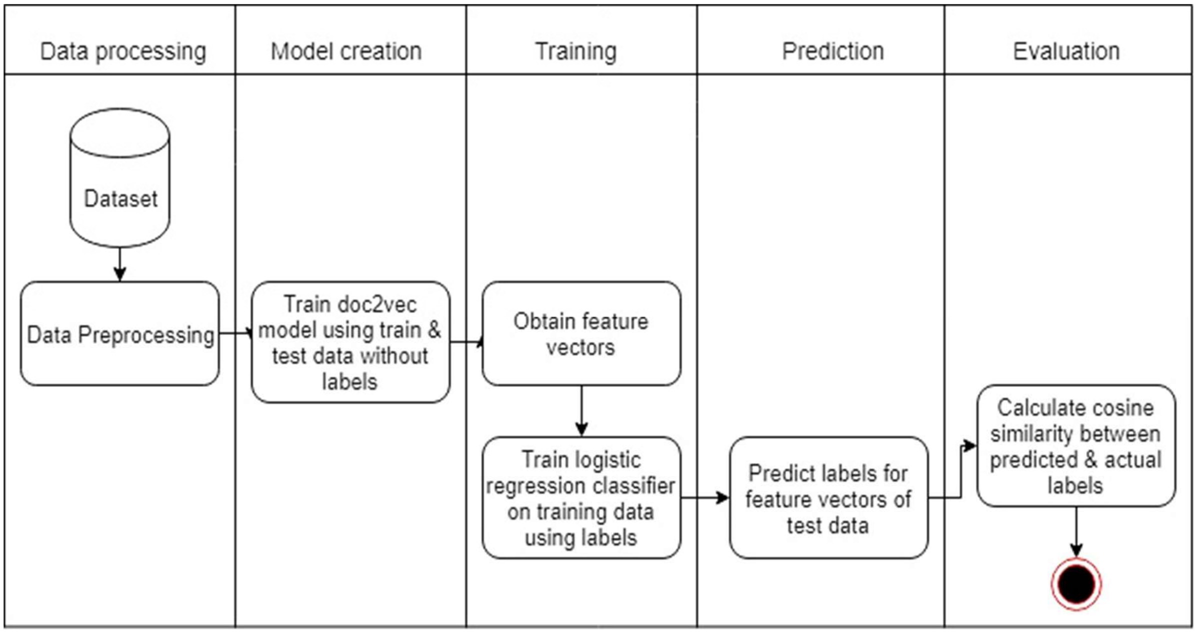

Figure 3.1 Steps in the implementation of the methodology

3.1 Dataset

The dataset ("Data and Tools < SemEval-2017 Task 5," 2016) used in this research are the datasets provided by SemEval 2017 competition for their Task 5 (Cortis et al., 2017). The task 5 in the competition involved two subtasks due to which the dataset contains financial news as well as microblogs. The subtask 1 involved analysis on microblogs which were collected from Twitter and StockTwits, the subtask 2 involved analysis on news headlines which were crawled from different sources on the Internet like Yahoo Finance. These two datasets were collected from random sampling and filtering of spam microblogs and news headlines. Cortis et al. (2017) stated that "The News Statements and Headlines have been collected from a pool of 20.000 RSS feeds in the period between August and November 2015 (e.g. AP News, Reuters, Handelsblatt, Bloomberg, and Forbes)." The dataset is resampled at different time unit levels to make the sample more random and an adequate representation of entire span of data. Additionally, this dataset is the gold standard. Four financial domain experts worked to annotate and consolidate the dataset as directed by the creators of the dataset (Cortis et al., 2017) for a total of 120 hours (30 hours by each expert). Three experts worked on annotations whereas one expert worked to consolidate the dataset.

For the financial microblogs, the dataset is annotated with the source, id, cashtag, sentiment, and spans (Figure 3.6). The source identifies the website with the microblog is posted which can wither be Twitter or StockTwits. The id provides a unique identifier to the microblog. The cashtag performs the task to identify the name the company with which the stock is associated. The sentiment is a floating-point value between -1 to 1, where 1 means positive and -1 means negative. The spans contain a part of the microblog expressing the sentiment about the microblog. The training set for microblog contains 1702 data points whereas the test set contains 799 microblogs.

Figure 3.2 shows the split of positive and negative labels present in the training data for Microblogs. It can be observed that the negative labels are approximately one third of the total microblogs.

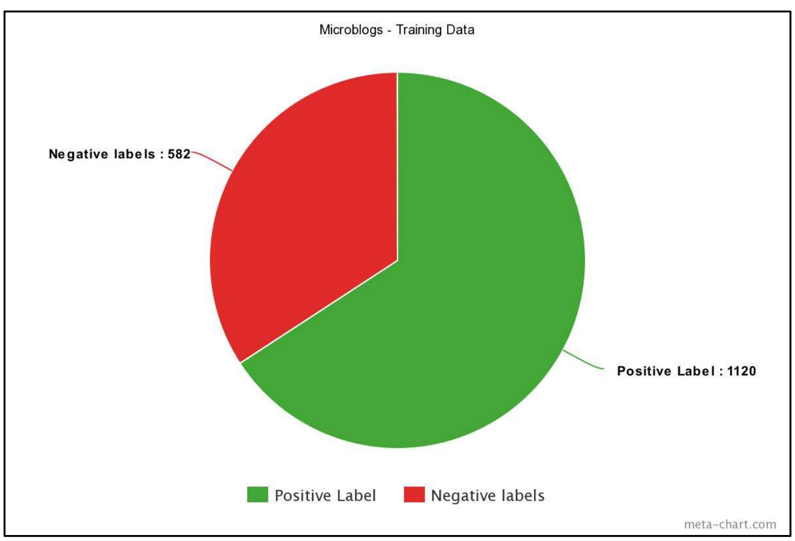

Figure 3.2 Pie chart showing labels for Microblogs training data

Figure 3.3 shows the split of positive and negative labels present in the testing data for Microblogs. It can be observed that the negative labels are approximately one third of the total microblogs and the split of positive and negative labels in the training and testing data is similar. The second observation is that the testing data is approximately one third the size of total data points.

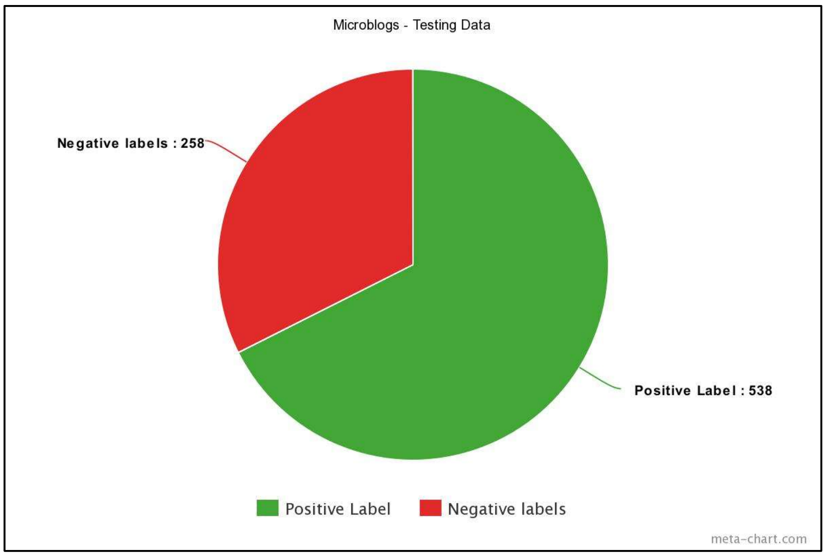

Figure 3.3 Pie chart showing labels for Microblogs testing data

For the news headline, each headline is annotated with the company, the company's uniqueid, the id of the news headline, a sentiment score ranging from -1 to 1. In this scenario a sentiment score of 1 means that the news headline is positive (bullish) and a sentiment score of -1 means that the news headline is negative (bearish). The training set contains 1142 annotated sentences whereas the testing set contains 491 annotated sentences.

Figure 3.4 shows the split of positive and negative labels present in the training data for News headlines. It can be observed that the negative labels are approximately greater than one third of the total news headlines by a small amount.

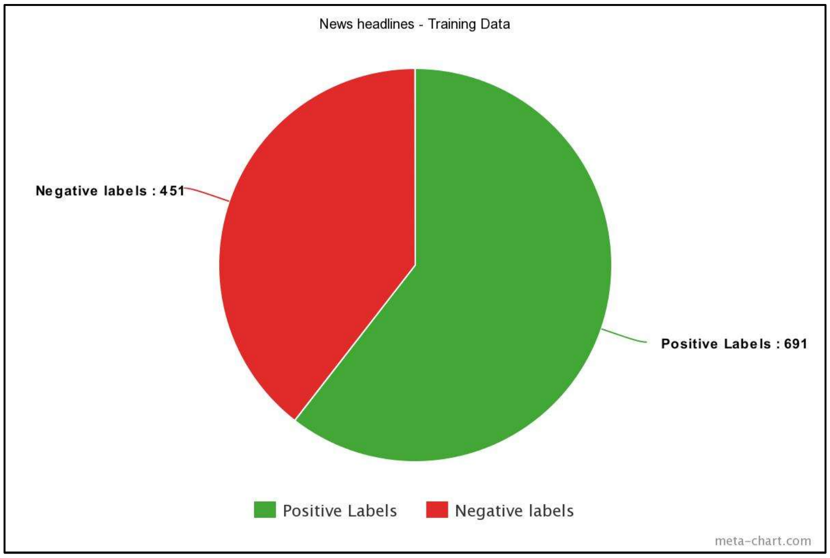

Figure 3.4 Pie chart showing labels for New headlines training data

Figure 3.5 shows the split of positive and negative labels present in the testing data for News headlines. It can be observed that the negative labels are approximately one third of the total news headlines and the split of positive and negative labels in the training and testing data is similar. The second observation is that the testing data is approximately one third the size of total data points.

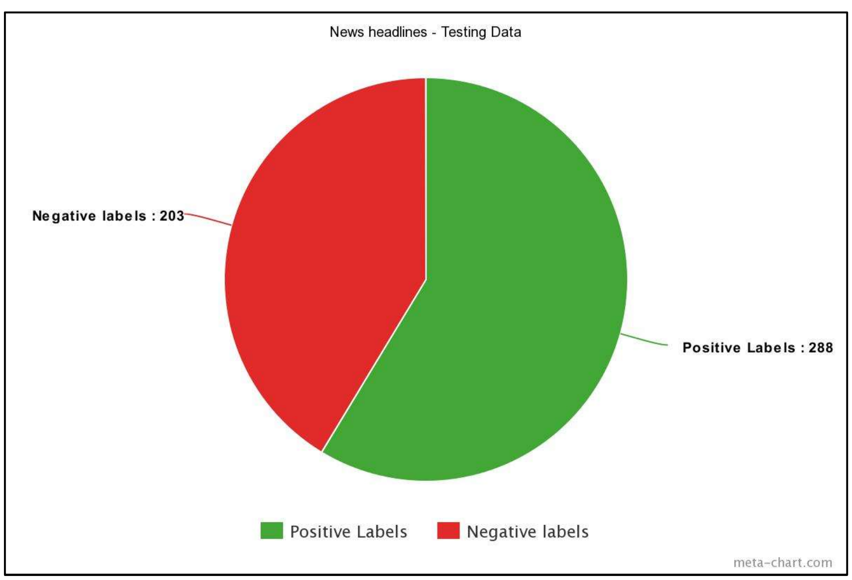

Figure 3.5 Pie chart showing labels for New headlines testing data

Figure 3.6 displays the information in one data point for a microblog in the microblog sub dataset. It contains the cash tag which is used to identify a stock, the sentiment score is considered as positive if the floating point score is 0 or greater and considered as negative is the floating point score is less than 0.


```
{
  "source": "twitter",
  "cashtag": "$DIA",
  "sentiment score": "0.460",
  "id": "719891468173844480",
  "spans": [
    "Looking for a strong bounce",
    "Lunchtime rally coming"
  ]
},
```


Figure 3.6 Screenshot of a data point from microblogs dataset

Figure 3.7 displays the information in one data point for a news headline in the headlines sub dataset. Unlike the cash tag which is used to identify a stock in the microblog dataset, company field is used to identify the company. The sentiment score is considered as positive if the floating point score is 0 or greater and considered as negative is the floating point score is less than 0. Additionally, instead of spans, the dataset contains the field title, and there is only a single title present unlike multiple spans for some microblogs.


```
{
  "UniqueID": "242_GSK",
  "id": 1633,
  "company": "GSK",
  "title": "GSK and Novartis complete deals to reshape both drugmakers",
  "sentiment score": 0.285
},
```


Figure 3.7 Screenshot of a data point from news headlines dataset

3.2 Population, sampling approach, and sample size

The population used in this study is the collection of all financial microblogs and news headlines from which the organizers collected the microblogs and news. Since the sample is selected randomly from the population, the sampling approach used in this research is simple random sampling. The sample size for financial microblogs subset is 2501 microblogs and for news headlines is 1833 headlines.

3.3 Variables

The variables in this research are the source, id, cashtag, sentiment, and span for the microblogs subset of the dataset. The variables for the news headlines subset are a company name, the headline, and sentiment score. In both the subsets, the independent variable is sentiment score which is predicted from the dependent variable of the span and the headline. The other variables of the source, id, cashtag for microblog and company name are not considered for this research.

It is important to understand that the news headline or microblog is one of the dependent variables which affects the sentiment. Some of the other variables which can be included are the popularity of the person who tweeted the microblog, the reader's base of the newspaper which published the headline, the time at which the news headlines or microblog was accessible to the public, the number of people who viewed the news headline or microblog, etc. All these dependent variables are not taken into consideration for this research.

3.4 Data processing

The first step will involve the preprocessing of the Microblog data which involves

1. To remove special characters
2. To remove URLs and usernames
3. To convert Microblogs to lowercase and merge hashtags with the sentence
4. Convert sentiment from floating point value to 1 or 0 (binary)
5. Merging of spans to form a single sentence

The second step will involve the preprocessing of Headlines data which involves

1. To remove special characters
2. To remove URLs and usernames
3. To convert Headlines to lowercase and merge hashtags with the sentence
4. Convert sentiment from floating point value to 1 or 0 (binary)

3.5 Methodology

The methodology section involves three primary tasks which are

1. Creation of the doc2vec model
2. Training of the classifier on training data
3. Prediction of labels using a classifier

3.5.1 Creation of doc2vec model

The doc2vec model is created using the PV-DM algorithm using gensim(*gensim*, 2011/2018a). The paragraph vectors and word vectors are created using the training and test provided by the organizers, this is required because each document (news headline or paragraph) needs to have a vector representation in the prediction task. The same method is used by Le and Mikolov (2014) and is demonstrated by Radim Rahurek in his tutorial on github (*gensim*, 2011/2018b).

PV-DM is used in this research as it is demonstrated to display better performance than other doc2vec algorithms (Le & Mikolov, 2014). The window size of the skip gram model is set to 10, it is the maximum distance between the predicted word and context words used for prediction within a document. Dimensionality of feature vectors is set to 100, Le and Mikolov have used dimensionality of 400, but 100 dimensions are used in this research to reduce the computation cost of the model. Number of epochs (number of iterations over the corpus) used to train this model are 50. The min_count parameter which is used to ignore all words with a total frequency lower than it is set to 1 since sentence labels appear only once.

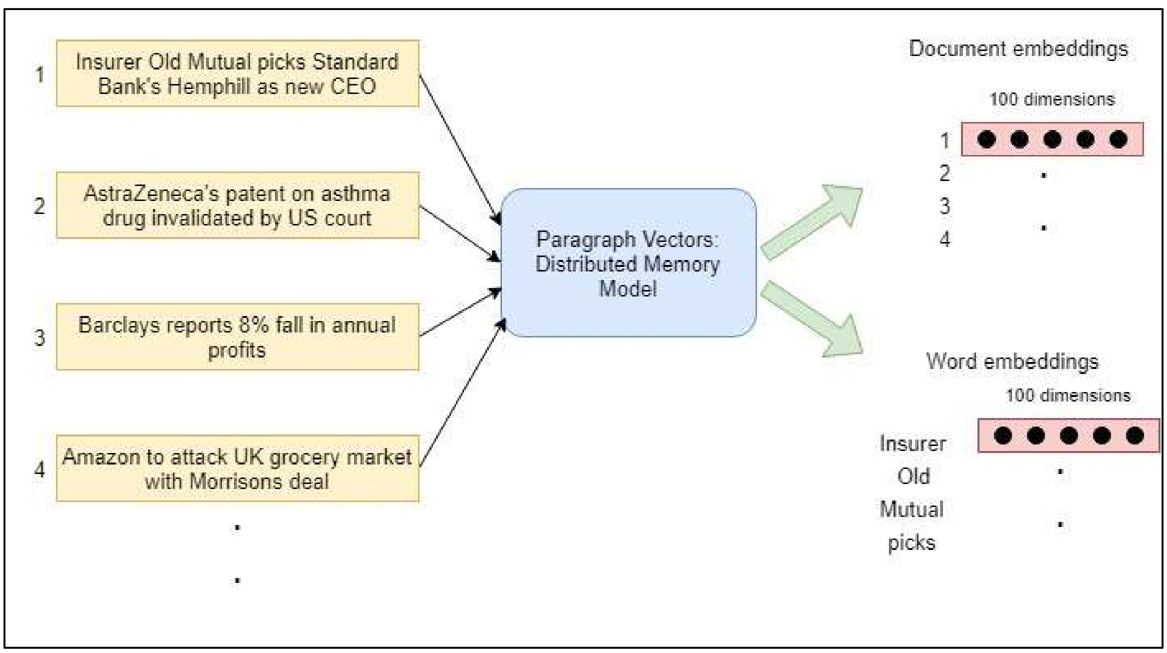

Figure 3.8 Creation of document and word embeddings for news headlines using PV-DM model

Figure 3.8 depicts the creation of word and document embeddings using the PV-DM model, the left hand side contains the news headlines along with the order in which they are fed to the model. The headlines and microblogs are fed in random order to the algorithm from the dataset. The model creates two embeddings matrices using the news headlines, document embeddings and word embeddings with 100 dimensions for each document and word.

3.5.2 Training of classifier

Le and Mikolov (2014) report that machine learning algorithms of k-means and logistic regression are the heart of many applications such as document retrieval, spam filtering, and web search. Additionally, for these applications text classification and clustering play an important role. Another experiment carried out to classify paragraph vectors using logistic regression on IMDB sentiment dataset showed the accuracy of 87 percent (Qiu, 2015/2018). Due to these reasons, logistic regression classifier is used in this research. The training of the logistic regression classifier takes place on testing data points in the microblogs and news headlines, this is the original testing and training split of the dataset. This has been described in the dataset section in detail

3.5.3 Prediction of label using classifier

The classifier created from step 3.5.2 is used to predict the sentiment label (0 or 1). During the step 3.5.1 feature vectors are created for the testing dataset of the two sub-datasets. They include 799 microblogs and 491 news headlines. These feature vectors are fed into the logistic regression classifier to predict the sentiment label for the test data.

After the labels are predicted a vector of the labels is created to compare it with the vector of testing data.

3.6 Evaluation

The results will be evaluated by making use of cosine similarity because it is useful in comparing the performance of this methodology to other approaches performed in the SemEval 2017 competition (Cortis et al., 2017). Cosine similarity is a measure which calculates the similarity between two vectors.

$$cosine(Test, Predicted) = \frac{\sum_{i=1}^{n} Test_i \times Predicted_i}{\sqrt{\sum_{i=1}^{n} Test_i^{\,2}} \times \sqrt{\sum_{i=1}^{n} Predicted_i^{\,2}}}$$

Equation 3.1 Calculation of cosine score

$$cosine\ weight = \frac{|Predicted|}{|Test|}$$

Equation 3.2 Calculation of cosine weight

$$cosine\ similarity = cosine\ weight \times cosine\ (Test, Predicted)$$

Equation 3.3 Calculation of cosine similarity

Cosine similarity for this research can be calculated as follows, Equation 3.2 calculates the cosine weight. The cosine weight is the ratio of predicted values to test values. Equation 3.3 uses the values from Equation 3.1 and Equation 3.2 to calculate the cosine similarity. The cosine weight value in Equation 3.2 is added to reward approaches which predict more values in the test dataset.

CHAPTER 4. RESULTS AND DISCUSSION

4.1 Results

The results of this research reported using cosine similarity measure for both the datasets, are higher than the best result in the SemEval 2017 competition, as can be seen in Tables 4.1 and 4.2. The values in Table 4.1 and Table 4.2 are incorporated from the data released by the organizers of the competition (Cortis et al., 2017).

4.1.1 Results on Microblogs dataset

The cosine similarity value is 0.819 which is higher than the reported value of 0.778 by 0.041

Table 4.1 Comparison of cosine similarity - Microblogs

Microblogs results			
Paragraph2vector using PV-DM (This thesis)			**0.819**
Rank	Team Name	Reference	Score
1	ECNU	(Jiang, Lan, & Wu, 2017)	0.778
	CodersGoneCrazy		0.760
	Zhiqiang		0.759
2	IITP	(Ghosal, Bhatnagar, Akhtar, Ekbal, & Bhattacharyya, 2017)	0.751
3	SSN_MLRG1	(Deborah, Rajendram, & Mirnalinee, 2017)	0.735

4	HHU	(Cabanski, Romberg, & Conrad, 2017)	0.730
5	IITPB	(Kumar et al., 2017)	0.726

4.1.2 Results on News headlines dataset

The cosine similarity value is 0.767 which is higher than the highest reported value of 0.745 by 0.022.

Table 4.2 Comparison of cosine similarity - New Headlines

News headlines results			
Paragraph2vector using PV-DM (This thesis)			**0.767**
Rank	Team Name	Reference	Score
1	Fortia-FBK	(Mansar et al., 2017)	0.745
2	RiTUAL-UH	(Kar, Maharjan, & Solorio, 2017)	0.744
3	TakeLab	(Rotim, Tutek, & Šnajder, 2017)	0.733
4	Lancaster A	(Moore & Rayson, 2017)	0.732
	CodersGoneCrazy		0.726
5	ECNU	(Jiang et al., 2017)	0.710
6	HHU	(Cabanski et al., 2017)	0.702

4.2 Discussion

Paragraph vectors were originally designed to handle large corpora and unlabeled data. However, the corpus used in this experiment is significantly smaller than the IMDB movie review dataset which contains 100000 movie reviews.

Table 4.3 Size of dataset versus IMDB movie review dataset

	Data points	Ratio to IMDB movie reviews dataset
New Headlines	1633	0.016
Microblogs	2501	0.025

After considering the ratio of data points as it can be inferred that the size of corpus on which the paragraph embeddings are trained is not an important factor to evaluate performance in binary classification of sentiments. Similarly, high dimension of embeddings captures more information. But in this experiment, the dimensions were scaled down from 400 to 100 to reduce the computation cost when compared to Le and Mikolov (2014). Another important consideration for comparison with Le and Mikolov (2014) is that the IMDB movie review dataset contains multiple sentences in one review. While this experiment contains merging of spans, the data points for which spans are merged is significantly smaller than the dataset. Additionally, the length of a single News headline and Microblog is less than a movie review in the IMDB movie review dataset.

Among the two datasets present in this thesis, the average length of news headlines is longer than microblogs. Figure 4.1 shows the Sequence Length vs the Frequency for News headlines present in the training data of the news headlines dataset. Further analysis of the same data reveals that the total number of words in the dataset is 10659 and the average number of words in the files is 9.35.

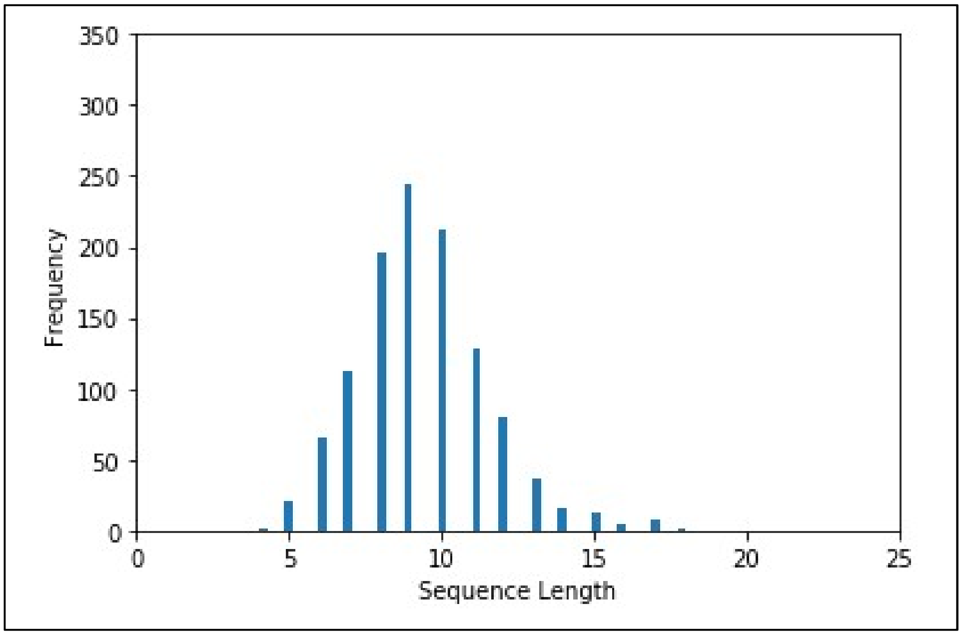

Figure 4.1 Sequence Length vs Frequency for News headlines dataset

Figure 4.2 shows the same data as 4.1 for the IMDB movie reviews dataset, the IMDB movie reviews training dataset has 25000 movie reviews for the total number of words in the reviews is 5844680. The average number of words in the reviews is 233.79. Comparing Figure 4.1 and Figure 4.2 it can be inferred that the average length of the document in the corpus does not affect performance metric of cosine similarity.

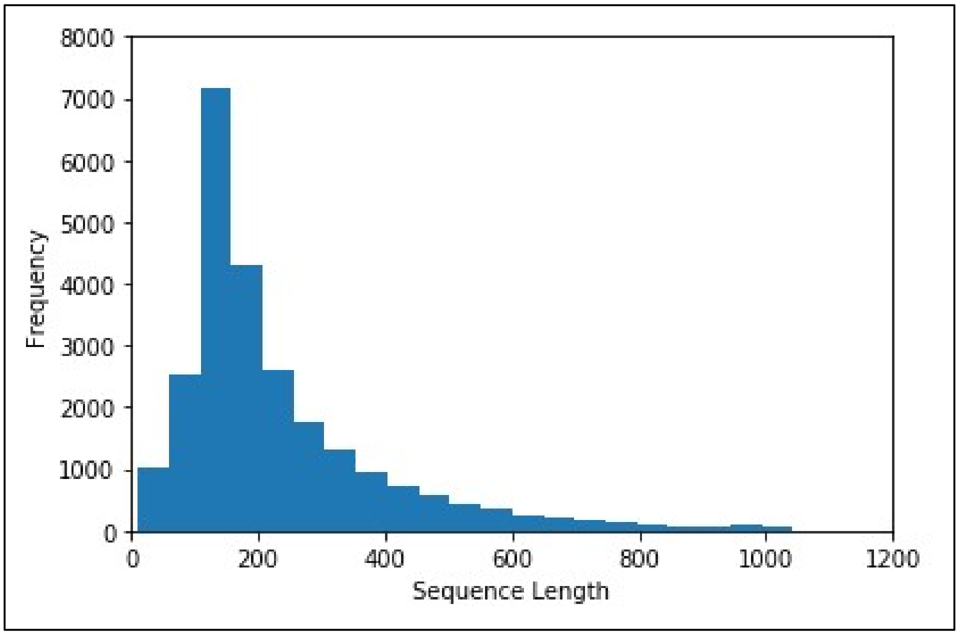

Figure 4.2 Sequence Length vs Frequency for IMDB movie reviews dataset

While some of the microblogs are rich in information, the data pre-processing and the short and informative nature of microblog renders some microblogs as one-word documents. Some one-word examples include

1. *"hold"*
2. *"long"*
3. *"wmt"*
4. *"fcx"*
5. *"oversold"*
6. *"buy"*
7. *"spy"*
8. *"uco"*
9. *"in"*
10. *"cat"*

Longer examples include

1. *"looks really interesting on drop grabbed some options and stock enough tme before earnings to grow the stock growth all sectors"*
2. *"looks like nice risk reward for starting to initiate position looks like nice risk reward for starting to initiate position"*

Some microblogs like

1. *"long over great find hope that wold be another anly or pzzi still holding second"*
2. *"don underestimate mark bounce off sma daily"*
3. *"airlines battle for cuba routes aal luv save jblu"*

For words like *"anly", "pzzi", "aal"*, it is difficult to infer meanings of such words and pre-trained word embeddings which do not include these words. The window size parameter of paragraph vector which is set to 10 in this experiment takes this into consideration when creating the paragraph vectors. Words like *"don"* misspelled for *"don't"*, *"luv"* for *"love"* and "*wold"* for "*would"* further question pre-trained embeddings and support the argument for window size which covers the entire news headline or microblog.

4.2.1 Analysis of similar words

Figure 4.3 shows ten most similar words to the words "*bullish"* and "*downtrend"* from paragraph vector model trained on microblogs dataset along with the cosine similarities. The function most_similar return top ten most similar microblogs to any word which was fed during the creation of the model.

```
model.most_similar('downtrend')

[('retest', 0.8862528800964355),
 ('itting', 0.8613978624343872),
 ('helping', 0.8425077795982361),
 ('line', 0.7428999543190002),
 ('average', 0.7249331474304199),
 ('bidders', 0.7095080018043518),
 ('broken', 0.7053303718566895),
 ('dropping', 0.6947202086448669),
 ('its', 0.6915990710258484),
 ('day', 0.6910512447357178)]
```

```
model.most_similar('bullish')

[('ief', 0.8448168039321899),
 ('awesomely', 0.8358787298202515),
 ('crossovers', 0.807542264615173),
 ('tlt', 0.7836184501647949),
 ('abc', 0.772089958190918),
 ('ewz', 0.768324077129364),
 ('producing', 0.7589824199676514),
 ('smh', 0.7480920553207397),
 ('decreased', 0.7479851245880127),
 ('xover', 0.7479449510574341)]
```

Figure 4.3 Ten most similar words to “bullish” and “downtrend” from the model trained on Microblogs

Figure 4.4 shows ten most similar words to the words “*falls”* and “*increase”* from paragraph vector model trained on news headlines dataset along with the cosine similarities.

```
model.most_similar('falls')

[('pullback', 0.7357574105262756),
 ('ashtead', 0.7130506038665771),
 ('devaluation', 0.711600124835968),
 ('petrol', 0.7097448110580444),
 ('wimpey', 0.6971498131752014),
 ('taylor', 0.6832544803619385),
 ('prices', 0.677039623260498),
 ('fresnillo', 0.6715071797370911),
 ('asda', 0.6630493402481079),
 ('higher', 0.6625311374664307)]
```

```
model.most_similar('increase')

[('death', 0.8344177007675171),
 ('suggests', 0.8297720551490784),
 ('onglyza', 0.7587854862213135),
 ('diabetes', 0.7050268650054932),
 ('fda', 0.6583082675933838),
 ('data', 0.6525143384933472),
 ('rate', 0.6500493288040161),
 ('savor', 0.6307816505432129),
 ('aims', 0.6213951110839844),
 ('reviews', 0.6160950660705566)]
```

Figure 4.4 Ten most similar words to “falls” and “increase” from the model trained on News headlines

By observing Figure 4.3 and Figure 4.4, we conclude that due to the small dataset, the model performance to identify similar words in Microblogs and News headlines is not accurate enough - which is understandable. If we have a relatively large dataset, there is a high probability that the model can identify similar words. However, some similar word results are promising. For example, the word "*falls*" is related with "*devaluation*" and "*pullback*", "*downtrend*" is related to "*dropping*" and "*broken*" and the word "*bullish*" is related to the word "*awesomely*".

4.2.2 Analysis of similar documents

4.2.2.1 Analysis of similar news headlines

On further analysis of the model to find most similar microblogs or new headlines, the results are promising. For the news headline: *"warren buffett berkshire hathaway quarterly profit jumps almost one third"* which is labelled as positive in the news headline training dataset, Table 4.4 lists most similar news headlines along with their labels and cosine similarities to the inferred vector of the headline. The function most_similar returns ten most similar documents and has been used to analyze documents. The cosine similarity obtained from the model is rounded to three decimal places. In gensim under doc2vec, there is a function named infer_vector which can be used to infer a vector of a document after bulk training. It is important to note that most similar vectors from inferred vectors can change or remain same after vectors are inferred from the model.

Original headline: *"warren buffett berkshire hathaway quarterly profit jumps almost one third"*

Table 4.4 Ten most similar news headlines to the inferred vector of "warren buffett berkshire hathaway quarterly profit jumps almost one third"

News headline	Label	Cosine similarity

uk winners losers aviva and friends life lead ftse gainers	1	0.759
bp shares tumble after bn fourth quarter loss	0	0.510
turnaround buys for bhp billiton plc and home retail group plc	1	0.502
astrazeneca digs into precision medicine with lung heart deals	1	0.500
ao world shares tumble on profit warning	0	0.496
astrazeneca juno latest to collaborate on immuno oncology drugs	1	0.460
passengers rise at easyjet and aer lingus	1	0.457
astrazeneca fda panel reviews savor study results for onglyza	1	0.434
astrazeneca share price company to carve out antibiotic unit into separate	1	0.406
warren buffett berkshire adds to favorites ibm wells fargo	1	0.400

From Table 4.4 we can observe that the model is effective in terms of mapping similar vectors together. Eight out of the ten most similar news headlines from table 4.4 are labelled positive which is the label of the original new headline. Additionally, the two negative headlines which are like the original headline have words which overlap to the original news headlines. The words quarter is like the word quarterly and the word profit is an exact overlap. The model learns to identify that "*quarterly*" and "*quarter*" mean the same thing. Furthermore, the two negatively labelled headlines have the two words "*shares tumble*" in them which explains why they could be mapped together in the vector space by the model.

The news headline *"astrazeneca fda panel reviews savor study results for onglyza"* does not contain any word or phrase which can lead to the identification of positive sentiment. The floating-point sentiment score provided in the dataset is 0.075. To perform binary classification, 0.075 is considered positive in this research. Comparing the value of 0.075 to 1.0 it can be inferred that model is effective in classifying sentiment in two classes. A news headline for the same company *"Berkshire Hathaway"* has also been identified close to the original news headline. The two negatively labelled headlines which are among the ten most like the original headline, *"bp shares tumble after bn fourth quarter loss"* and *"ao world shares tumble on profit warning"* have floating point sentiment scores of -0.927 and -0.815 respectively. These scores are close the value of -1.0 which is the negative end of the sentiment score. The occurrence of the phrase *"shares tumble"* in both the headlines and the word *"quarter"* present in the original headline can be an explanation for this result.

For the news headline: *"Kingfisher share price slides on cost to implement new strategy"* which is labelled as negative with a floating sentiment score of -0.786 in the news headline training dataset, Table 4.5 lists most similar news headlines along with their labels and cosine similarities to the inferred vector of the headline. The cosine similarity obtained from the model is rounded to three decimal places

Original headline: *"Kingfisher share price slides on cost to implement new strategy"*

Table 4.5 Ten most similar news headlines to the inferred vector of "Kingfisher share price slides on cost to implement new strategy"

News headline	Label	Cosine similarity
four ex barclays bankers sentenced for roles in libor rate rigging scandal	0	0.805
kingfisher takeover of mr bricolage could hit brick wall	0	0.651

barclays ceo staley says brexit slump caused by profit fears	0	0.637
rbs becomes shadow of its former self	0	0.588
valeant pearson says timing of return uncertain	0	0.584
prudential hit by withdrawals from investment arm	0	0.564
former barclays banker pleaded guilty to libor rigging offence	0	0.558
update bhp billiton profit dives to year low on commodities rout	0	0.530
berkshire seeks to boost its wells fargo stake above percent	1	0.515
oil giant shell to cut around jobs amid bg takeover	0	0.492

From Table 4.5 we can observe that the model is effective in terms of mapping similar vectors together. Nine out of the ten most similar news headlines from table 4.5 are labelled negative which is the label of the original new headline *"Kingfisher share price slides on cost to implement new strategy"*.

These results are promising, however for some new headlines like *"tesco sells half of stake in ecommerce site lazada to alibaba for"* which is labelled as positive with a floating-point sentiment score of 0.365 and *"peroni and grolsch put up for sale as ab inbev plans acquisition of sabmiller"* which is labelled as positive with floating-point sentiment score of 0.307. Table 4.6 and Table 4.7 lists ten most similar sentences to the two news headlines described earlier.

Table 4.6 Ten most similar news headlines to the inferred vector of "tesco sells half of stake in ecommerce site lazada to alibaba for"

News headline	Label	Cosine similarity

rbi surprises street sensex pares gains after hitting mount	1	0.332
diageo shares surge on report of possible takeover by lemann	1	0.316
the coca cola company and coca cola femsa to acquire ades soy based beverage business from unilever	1	0.294
easyjet leads britain ftse lower as global bond rout resumes	0	0.285
buffett berkshire builds deere stake dumps exxon	1	0.269
insight hires aviva david hillier for multi asset team	1	0.240
update exchanges barclays win dismissal of us high frequency trading case	1	0.235
bp boosts ftse towards four and half month high	1	0.200
spain caixabank launches new takeover bid for banco bpi	1	0.198
direct line rings up higher profit	1	0.194

If we look closer to headline *"tesco sells half of stake in ecommerce site lazada to alibaba for"* then it does not have the value for which Tesco sells half of the stake to Alibaba. On further examination of the original news headline from the dataset, the original headline is *"Tesco sells half of stake in ecommerce site Lazada to Alibaba for Â£90m"*. The data preprocessing to remove the punctuations and numbers removes the word *"Â£90m"*. The labelled sentiment is positive for the company and the associated company is *"Tesco PLC"*. The headline does not mention the for what amount the company is sold, and the model is limited to not knowing what a good amount is to sell the ecommerce website. Two key learnings from the analysis of this headline are

1. The data preprocessing in this research leads to loss of valuable information

2. The model still maps the vector of this news headline close to nine other headlines which are positive

Table 4.7 Ten most similar news headlines to the inferred vector of "peroni and grolsch put up for sale as ab inbev plans acquisition of sabmiller"

News headline	Label	Cosine similarity
standard life share price group gets approval to hike stake in india jv	1	0.529
stanchart capital raising would be major surprise investor Aberdeen	1	0.525
astrazeneca to buy zs pharma for billion	1	0.460
bhp names chairman board members to oversee samarco disaster	0	0.311
lawmakers worry ab inbev beer deal will hurt craft brewers	0	0.265
former aviva investors analyst mothahir miah banned and fined by fca	0	0.260
update sabmiller rejects informal offer from ab inbev as too low Bloomberg	0	0.251
crh adds cr laurence to acquisitions tally for bn	1	0.245
metals zinc soars pct fuelling metals surge after glencore cuts output	1	0.237
rbs bosses ordered to go out and meet small firms	1	0.235

Observing Table 4.7, the news headline *"peroni and grolsch put up for sale as ab inbev plans acquisition of sabmiller"* is labelled as positive and associated with company SABmiller. There are three companies involved in this headline, "*SABmiller*", "*AB Inbev*" and *"Peroni and Grolsch"*. The sentiment for the company *"Peroni and Grolsch"* can be negative since they are put up for sale. Similarly, for the company *"AB Inbev"* the sentiment can be positive since they are planning to acquire SABmiller. In the original dataset, the company associated with this headline is *"SABmiller"* and the headline mentions that it is planned to be acquired. On the contrary, the sentiment score for this headline is positive. Key learnings from the analysis of this headline are

1. Targeted sentiment analysis for specific companies can lead to improved understanding of sentiment for some news headlines in this dataset and the sentiment score associated with each headline can be possibly improved.
2. The model maps the vector of this news headline close to six other headlines which are positive and four headlines which are negative from which it can be inferred that the model does not learn the sentiment of this news headline as expected.

4.2.2.2 Analysis of similar microblogs

The analysis of microblogs leads to further observations.

1. Microblog: *"in for swing trade looks like want go up"*.
2. Original microblog: *"$BAC in for swing trade looks like want go up"*
3. Sentiment score: 0.433.

Table 4.8 Ten most similar microblogs to the inferred vector of "in for swing trade looks like want go up"

Microblog	Label	Cosine similarity
buy the dip	1	0.577
over million new internet users are being added globally every year	1	0.353

bto three puts	0	0.338
actually lost per share	0	0.252
add to fxp skf positions	1	0.228
what goes up	0	0.182
weak outlook	0	0.179
top holdings	1	0.177
looking for strong bounce lunchtime rally coming	1	0.174
earnings growth looks dim current multiple highr	0	0.167

Five out of the ten microblogs for the inferred vector of *"in for swing trade looks like want go up"* are positive. Also, the cosine similarity values are below 0.5 except for the first most similar microblog.

Analysis for

1. Microblog: *"add short"*
2. Original version: *"add short 119.2 $SPY"*
3. Sentiment score: -0.550.

Table 4.9 Ten most microblogs to the inferred vector of "add short"

Microblog	Label	Cosine similarity
guess everyone wants	1	0.759
multi top on mcd chart finally leads to fall in price stability is questioned in the stock but coming weeks will tell	0	0.754

beats the estimize eps consensus by and the estimize revenue consensus by	1	0.752
watchlist top stocks	1	0.749
premier retail dividend play	1	0.749
may have lower entry point	1	0.747
beat up ah	1	0.745
making higher highs higher lows as it consolidates over past several weeks	1	0.744
Downgrade	0	0.742
can one of the cheerleaders show me actual valuation work	0	0.741

Only ten of the three most similar microblogs has negative sentiment, which is the actual sentiment of the news headline. Compared to Table 4.8 the short length might be one of the reasons for this. However, the cosine similarities are comparatively higher than Table 4.8.

To further analyze this result, we compute top ten most similar microblogs to the two microblogs which are *"add short"* and "*in for swing trade looks like want go up"*. Table 4.10 lists top ten most similar documents to the microblog *"add short"* without inferring the vector. The microblog has already been used in the creation of the model to a vector representation of the microblog already exists.

Table 4.10 Ten most microblogs of "add short"

Microblog	Label	Cosine similarity

Overbought	0	0.990
see big down day in the near future	0	0.987
sold today	0	0.985
looks really toppy right here	0	0.984
short interest or not	0	0.976
no position	0	0.976
people are calling it bearish some heading for exits already	0	0.974
sanofi often on the wrong end of business	0	0.972
tsla recalls model vehicle shares volatile	0	0.970
fb is taking the shine off aapl and goog	0	0.970

All ten most similar microblogs from Table 4.10 return a microblog which is labelled as negative

Like 4.10, Table 4.11 lists top ten most similar documents to the microblog *"in for swing trade looks like want go up"* without inferring the vector.

Table 4.11 Ten most similar microblogs of "in for swing trade looks like want go up"

Microblog	Label	Cosine similarity
down premarket	0	0.212
sector stocks leading today	1	0.196
not feel terribly good	0	0.181
new squeeze plays	0	0.175

going in tomorrow evening	1	0.166
producing bullish macd xover	1	0.160
add to fxp skf positions	1	0.158
unusual call activity	1	0.158
five stocks that helped starboard value equity portfolio return in	1	0.153
gap up today they say the heat is coming	1	0.133

Seven out of ten microblogs return positive label which is the original label of *"in for swing trade looks like want to go up"*. From 4.10 and 4.11, we can say that the paragraph vector model returns the label correctly for the vector already present in the model. Also, when vectors are inferred for a microblog already present in the model, the labels are not accurate. The small length of the microblog may be one possible explanation for this.

CHAPTER 5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

Actionable insights can be drawn from this research towards sentiment analysis. Similar new headlines and microblogs were mapped together by the model, bigram features can be effective in identifying phrases such as *"shares tumble"*, *"going down"*. The data preprocessing in this research leads to loss of valuable information. Numbers could have possibly been included to gain further insights. Like numbers, special characters can also help to determine the sentiment in a specific country, one example for this could be $ or £. Use of named entity recognition during the preprocessing stages of headlines and microblogs can improve the performance of the classification by identification of a positive or negative sentiment for multiple

The relatively small size of this dataset did not affect the performance of this approach when compared to IMDB movie review dataset. The length of the document in the corpus did not affect performance metric of cosine similarity in this research. The cosine similarity was higher for the microblogs dataset as compared to the news headlines approach when the length of the microblogs was lower. However, when inferred vectors were calculated the news headlines paragraph vector model was able to identify the labels for the top ten most similar headlines more effectively as compared to the microblogs dataset.

Targeted sentiment analysis for specific companies can lead to improved understanding of sentiment for some news headlines in this dataset and the sentiment score associated with each headline can be possibly improved. This dataset does not associate multiple headlines with a single news head line or microblog, targeted sentiment analysis could be used to handle this. A single news headline or microblog can be used to calculate two or more sentiment values for different companies.

The most similar word embeddings were not as accurate compared to the document embeddings for both the models. One explanation of this can be due to the window size in training the paragraph vector model. The window size was approximately equal or greater than the average length of the headlines and microblogs.

5.2 Future work

Although the PV-DM algorithm of captures the context of the words present in the sentence. LSTM in combination with word embeddings could have captured the context in the financial news and microblogs. Additionally, the most readily available and used embeddings are of word2vec and Glove are not domain specific to financial news and microblogs. Training of word embeddings related to considerably large financial corpus and utilizing it for the previously mentioned approach would be interesting to experiment. Furthermore, in the paragraph vector algorithms, the word embeddings are trained on the corpus. Considering the size of this dataset it embeddings trained on large financial corpus could possibly increase the performance metrics of this experiment.

The researchers who participated in the SemEval 2017 competition used different approaches for data pre-processing of microblogs and headlines. It can possibly be performed prior to the model creation stage in this research to examine the results. The possible reason for this is due to the different nature of news and microblogs and the better prediction from inferring vectors from the news headlines model. Also, the top performance in the competition used an ensemble of machine learning and deep learning algorithms for prediction. This can be implemented to evaluate and compare the performance to approach presented in this thesis.

This research does not take lexicon-based approach into account, however, previously created wordlists (Loughran & Mcdonald, 2011) can be augmented by making use of word embeddings either from a pre-trained dataset or embeddings created using large financial corpora.

This research trained and evaluated the model by using 100 dimensions, experiments to determine the optimum number of dimensions for the embeddings could save the time required to perform similar experiments which involve big datasets. The other versions of the paragraph vector model creation algorithm such as PV-DBOW and PV-DM using concatenate can be used to create and compare the model on the same dataset. This research focused on a domain specific binary classification problem in the financial domain, the application of this research towards different datasets belonging to a domain other than finance can be performed.

LIST OF REFERENCES

Asgari, E., & Mofrad, M. R. K. (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLOS ONE*, *10*(11), e0141287. https://doi.org/10.1371/journal.pone.0141287

Blair-Goldensohn, S., Hannan, K., McDonald, R., Neylon, T., Reis, G. A., & Reynar, J. (2008). Building a sentiment summarizer for local service reviews. In *WWW workshop on NLP in the information explosion era* (Vol. 14, pp. 339–348). Retrieved from http://www.academia.edu/download/11179325/paper3.pdf

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Cabanski, T., Romberg, J., & Conrad, S. (2017). HHU at SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Data using Machine Learning Methods. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 832–836). Vancouver, Canada: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/S17-2141

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *Icwsm*, *10*(10–17), 30.

Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., & Haruechaiyasak, C. (2012). Discovering Consumer Insight from Twitter via Sentiment Analysis. *J. UCS*, *18*(8), 973–992.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160–167). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1390177

Cortis, K., Freitas, A., Daudert, T., Huerlimann, M., Zarrouk, M., Handschuh, S., & Davis, B. (2017). Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 519–535). Retrieved from http://www.aclweb.org/anthology/S17-2089

Countries with most Twitter users 2016 | Statistic. (2016). Retrieved September 21, 2017, from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Daiyan, M., Tiwari, D. S., Kumar, M., & Alam, M. A. (2015). A Literature Review on Opinion Mining and Sentiment Analysis. *International Journal of Emerging Technology and Advanced Engineering*, *5*(4). Retrieved from https://pdfs.semanticscholar.org/ad5d/ce5069ff5f1bf3fb8bd4bc28e12d126c8abc.pdf

Data and Tools < SemEval-2017 Task 5. (2016). Retrieved September 14, 2017, from http://alt.qcri.org/semeval2017/task5/index.php?id=data-and-tools

Deborah, A. S., Rajendram, S. M., & Mirnalinee, T. T. (2017). SSN_MLRG1 at SemEval-2017 Task 5: Fine-Grained Sentiment Analysis Using Multiple Kernel Gaussian Process Regression Model. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 823–826). Vancouver, Canada: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/S17-2139

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*. Retrieved from http://ci.nii.ac.jp/naid/10020680394/

gensim: Topic Modelling for Humans. (2018a). Python, RaRe Technologies. Retrieved from https://github.com/RaRe-Technologies/gensim (Original work published 2011) *gensim: Topic Modelling for Humans*. (2018b). Python, RaRe Technologies. Retrieved from https://github.com/RaRe-Technologies/gensim (Original work published 2011)

Ghosal, D., Bhatnagar, S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2017). IITP at SemEval-2017 task 5: an ensemble of deep learning and feature based models for financial sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 899–903). Retrieved from http://www.aclweb.org/anthology/S17-2154

Google Code Archive - Long-term storage for Google Code Project Hosting. (2013). Retrieved November 22, 2017, from https://code.google.com/archive/p/word2vec/

Goonatilake, R., & Herath, S. (2007). The volatility of stock market and news. *International Research Journal of Finance and Economics*, *3*, 53–65.

Graves, A., Mohamed, A. r, & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6645–6649). https://doi.org/10.1109/ICASSP.2013.6638947

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. *ArXiv Preprint ArXiv:1606.02820*. Retrieved from https://arxiv.org/abs/1606.02820

Harmon-Jones, E., Gable, P. A., & Price, T. F. (2013). Does Negative Affect Always Narrow and Positive Affect Always Broaden the Mind? Considering the Influence of Motivational Intensity on Cognitive Scope. *Current Directions in Psychological Science*, *22*(4), 301–307. https://doi.org/10.1177/0963721413481353

Harris, Z. S. (1954). Distributional Structure. WORD, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hatzivassiloglou, V., & McKeown, K. R. (1997). 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* (pp. 174–181). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/976909.979640

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hovy, E., & Kim, S.-M. (2004). Determining the Sentiment of Opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1220355.1220555

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168–177). New York, NY, USA: ACM. https://doi.org/10.1145/1014052.1014073

James Pennebaker, Roger Booth, & Martha Francis. (2007). *Linguistic inquiry and word count: LIWC [Computer software]*. Austin, TX. Retrieved from liwc.net

Jiang, M., Lan, M., & Wu, Y. (2017). ECNU at SemEval-2017 Task 5: An Ensemble of Regression Algorithms with Effective Features for Fine-Grained Sentiment Analysis in Financial Domain. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 888–893). Vancouver, Canada: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/S17-2152

Joshi, A. J. (1991). Natural Language Processing. *Science*, *253*(5025), 1242–1249.

Kar, S., Maharjan, S., & Solorio, T. (2017). RiTUAL-UH at SemEval-2017 Task 5: Sentiment Analysis on Financial Data Using Neural Networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 877–882). Vancouver, Canada: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/S17-2150

Kim, Y. (2014). Convolutional neural networks for sentence classification. *ArXiv Preprint ArXiv:1408.5882*. Retrieved from https://arxiv.org/abs/1408.5882

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Kumar, A., Sethi, A., Akhtar, M. S., Ekbal, A., Biemann, C., & Bhattacharyya, P. (2017). IITPB at SemEval-2017 Task 5: Sentiment Prediction in Financial Text. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 894–898).

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (pp. 1188–1196). Retrieved from http://www.jmlr.org/proceedings/papers/v32/le14.pdf

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1* (pp. 142–150). Association for Computational Linguistics.

Mansar, Y., Gatti, L., Ferradans, S., Guerini, M., & Staiano, J. (2017). Fortia-FBK at SemEval-2017 Task 5: Bullish or Bearish? Inferring Sentiment towards Brands from Financial News Headlines. *ArXiv Preprint ArXiv:1704.00939*. Retrieved from https://arxiv.org/abs/1704.00939

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119). Retrieved from http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34*(8), 1388–1429.

Moore, A., & Rayson, P. (2017). Lancaster A at SemEval-2017 Task 5: Evaluation metrics matter: predicting sentiment from financial news headlines. *ArXiv Preprint ArXiv:1705.00571*. Retrieved from https://arxiv.org/abs/1705.00571

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10). Retrieved from http://crowdsourcing-class.org/assignments/downloads/pak-paroubek.pdf

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10* (pp. 79–86). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1118693.1118704

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Qiu, L. (2018). *word2vec-sentiments: Tutorial for Sentiment Analysis using Doc2Vec in gensim (or "getting 87% accuracy in sentiment analysis in under 100 lines of code")*. Jupyter Notebook. Retrieved from https://github.com/linanqiu/word2vec-sentiments (Original work published 2015)

Reasons To Invest In Stocks - Fidelity. (2017). Retrieved September 21, 2017, from https://www.fidelity.com/viewpoints/retirement/why-you-need-stocks

Rezaeinia, S. M., Ghodsi, A., & Rahmani, R. (2017). Improving the Accuracy of Pre-trained Word Embeddings for Sentiment Analysis. *ArXiv:1711.08609 [Cs]*. Retrieved from http://arxiv.org/abs/1711.08609

Riloff, E., & Wiebe, J. (2003). Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (pp. 105–112). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1119355.1119369

Rotim, L., Tutek, M., & Šnajder, J. (2017). TakeLab at SemEval-2017 Task 5: Linear aggregation of word embeddings for fine-grained sentiment analysis of financial news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (pp. 866–871). Vancouver, Canada: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/S17-2148

Saeidi, M., Bouchard, G., Liakata, M., & Riedel, S. (2016). SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods. *ArXiv:1610.03771 [Cs]*. Retrieved from http://arxiv.org/abs/1610.03771

Salvetti, F., Reichenbach, C., & Lewis, S. (2006). Opinion Polarity Identification of Movie Reviews. In *Computing Attitude and Affect in Text: Theory and Applications* (pp. 303–316). Springer, Dordrecht. https://doi.org/10.1007/1-4020-4102-0_23

Sauper, C., & Barzilay, R. (2013). Automatic aggregation by joint modeling of aspects and values. *Journal of Artificial Intelligence Research*. Retrieved from https://www.jair.org/media/3647/live-3647-6805-jair.pdf

Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, *285*(Supplement C), 181–203. https://doi.org/10.1016/j.ins.2014.04.034

Socher, R., Bauer, J., Manning, C. D., & others. (2013). Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 455–465). Retrieved from https://aclanthology.info/pdf/P/P13/P13-1045.pdf

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642). Retrieved from http://www.aclweb.org/anthology/D13-1170

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The general inquirer: A computer approach to content analysis.* Oxford, England: M.I.T. Press.

Tabari, N., Seyeditabari, A., & Zadrozny, W. (2017). SentiHeros at SemEval-2017 Task 5: An application of Sentiment Analysis on Financial Tweets. Retrieved from http://www.aclweb.org/anthology/S17-2146

Turney, P. D. (2002). 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* (pp. 417–424). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073153

Van de Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, *42*(11), 4999–5010. https://doi.org/10.1016/j.eswa.2015.02.007

Welch, C., & Mihalcea, R. (2016). Targeted Sentiment to Understand Student Comments. In *COLING* (pp. 2471–2481). Retrieved from http://web.eecs.umich.edu/~mihalcea/papers/welch.coling16.pdf

Xu, R., Morgan, A., Das, A. K., & Garber, A. (2009). Investigation of Unsupervised Pattern Learning Techniques for Bootstrap Construction of a Medical Treatment Lexicon. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (pp. 63–70). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1572364.1572373

Xue, B., Fu, C., & Shaobin, Z. (2014). A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec. In *2014 IEEE International Congress on Big Data* (pp. 358–363). https://doi.org/10.1109/BigData.Congress.2014.59