Applying Content Specific Information to Enhance SentiWordNet Based Sentiment Classification

Author

MUHAMMAD LATIF

2011-NUST-MS PhD-CSE(E)-22

Supervised By

DR. USMAN QAMAR

DEPARTMENT OF COMPUTER ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY (NUST)
H12, ISLAMABAD

OCT, 2014

Applying Content Specific Information to Enhance SentiWordNet Based Sentiment Classification

Author

MUHAMMAD LATIF

2011-NUST-MS PhD-CSE(E)-22

A thesis submitted in partial fulfillment of the requirements for the degree of

MSComputer Software Engineering

Thesis Supervisor:

DR. USMAN QAMAR

Thesis Supervisor's Signature:_______________________________________

Declaration

I certify that this research work titled "***Applying Content Specific Information to Enhance SentiWordNet Based Sentiment Classification***" is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student:___________________

MUHAMMAD LATIF

2011-NUST-MS PhD-CSE(E)-22

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student:___________________
MUHAMMAD LATIF
2011-NUST-MS PhD-CSE(E)-22

Signature of Supervisor:______________________
DR. USMAN QAMAR

Copyright Statement

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made onlyin accordance with instructions given by the author and lodged in the Library of NUST College of E&ME. Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST College of E&ME, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the College of E&ME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST College of E&ME, Rawalpindi.

Acknowledgements

I am thankful to my Creator **ALLAH**Subhana-Watala to have guided me throughout this work at every step and for every new thought which You setup in my mind to improve it. Indeed I could have done nothing without Your priceless help and guidance. Whosoever helped me throughout the course of my thesis, whether my parents or any other individual was Your will, so indeed none be worthy of praise but You.

I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout in every department of my life.

I would also like to express special thanks to my supervisor DR.USMAN QAMAR for his help throughout my thesis and also for Advance Software Engineering and Software Quality Engineering which he has taught me. I can safely say that I haven't learned any other engineering subject in such depth than the ones which he has taught.Each time I got stuck in something, he came up with the solution. Without his help I wouldn’t have been able to complete my thesis. I appreciate his patience and guidance throughout the whole thesis.

I would also like to thank DR. SHOAB A KHAN, DR. ALI HASSAN and DR. REHAN HAFIZfor being on my thesis guidance and evaluation committee.I would also like to pay special thanks to Mr.AbdualWahabMuzzafar for his tremendous support and cooperation.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Dedicated to my exceptional parents and my wife whose tremendous support and cooperation led me to this wonderful accomplishment

ABSTRACT

Sentiment classification concerned with the automated techniques that predict the polar orientation of the text. It is an important and sub-research area of the opinion mining and text mining, with applications and benefits on different areas including customer recommender and feedback analysis, business intelligence, information retrieval and social well beings services.

English language lexical resource SentiWordNet have the highest no of lexicons where each synset (sets of synonyms) is labeled with subjective and objective numerical scores for sentiment information. It is specifically designed to assist opinion mining tasks. By using such readily available resource more effective sentiment analysis methods can easily build with the help of this sentiment biased information.

This research specifically used the SentiWordNet to put a solution for automatic sentiment classification problem on multi domain sentiment dataset of product reviews and polarity dataset of movie reviews. At first, sentiment features were collected from subjective terms of SentiWordNet and used in machine learning based sentiment classification. Due to limitation of subjective terms in SentiWordNet, text with null or few sentiment features could reflect ambiguous or null sentiments. We proposed a new dimension of content specific features i.e. syntactic noun and verb phrases along unigrams features, used to reinforce the performance of sentiment feature based classifier on the underlying reviews. Different scenarios in features combinations were executed to find the best representative features also with F-Score based feature selection to reduce dimensionality.

The obtained results are compared to other documented methods discussed in the literature. It was highlighted that obtained results of sentiment features along content specific features outer perform the results of similar approaches used on same data set of reviews. It indicates that content specific verb and noun phrases features could become a new dimension for sentiment classification.

Key words: Opinion Mining, Sentiment Classification, Lexical Resources, SentiWordNet, Sentiment Features, Content Specific Features.

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

Business communities are keen to utilize sentiment analysis and opinion mining for the purpose of business intelligence and identifying consumer behavior. There is huge data available over the internet in the form of reviews, blogs discussions, emails, feedbacks and tweets. This data creates an opportunity to improve corporate decision making.

The ability to utilize text mining to innovate products and services using automated methods from available databases became relevant to the success of organizations. The goal of sentiment analysis is to detect subjective information from text. Sentiment classification and opinion mining services could be utilized for the detailed analysis of identifying consumer behavior, feedback analysis and fraud prevention.

1.1. Analytical Treatment of Text and Sentiment Information

Text is the secondary mean of communication and transferring information after oral communication however more important than oral communication because it provides asynchronous communication and does not require the participant to be present at the same time. Also, it is pertinent that information is commonly stored in text format that bears explicit orientationalmost 85% of information maintained by the organizations is in text format [33]. One of the most significant forms of information available in organization's computers today is textual data. Furthermore, it can be said that even the internet represents a greater part of information in a textual format bearing human legibility so as to ensure comprehension.

Therefore, in this context it can be inferred that how text mining (i.e. data mining on text) gained significance while articulating its importance. . Text mining exploits knowledge discovery methods while applying them on unstructured data bearing textual orientation and leveraging other research areas such as natural language processing, artificial intelligence and tackles the complexities of information extraction from unstructured textual format. Text mining techniques have been applied in a number of knowledge discovery scenarios, such as automatic categorization of documents, trend analysis and spam detection.

An important analytical treatment of text concerns the ability to detect and extract opinions, or sentiment information. Detecting the sentiment of customers towards a new product based on feedback available in text format could be an important element affecting decision making and the product's future direction. Sentiment Analysis and Opinion Mining, generally refers to detection and extraction of information from text. It does so, by invoking automated statistical methods maintaining conformity and accordance. It has many potential applications like building more efficient recommender systems, financial analysis, product engineering and market research.

Sentiment analysis has been largely divided into two categories; sentiment classification and semantic orientation. In semantic-orientation, the polarity of a given text is known through sentiment bearing lexicons either with the use of rules based or unsupervised approach. In this technique some use corpus to find sentiment bearing words and phrases and some prefer available dictionary or lexicon resources. On the other hand, sentiment classification classifies the text into positive, negative or neutral classes. Different individual and combination of features were considered in literature to perform the sentiment classification. Different machine learning algorithms e.g. Support Vector Machine, Naïve Bayes and Max Entropy were also employed in isolation and in combination.

English language lexicon resource SentiWordNet [16] is specifically designed to assist sentiment analysis tasks. Previously, used in semantic orientation approach, such that the terms bias sentiment information was used to determine the sentence polarity and subsequently the whole document. However, in sentiment classification, sentiment features were derived from the 7% subjective terms available in SentiWordNet. It was noted that text usually contained few or null sentiment feature, and ambiguous or null polarity characterized by the classifier e.g. a negative review taken from kitchen domain:

> *"I've had my share of dutch ovens in my time, and I have to say that this is the foulest one yet. I thought I smelled a good deal when I got it, but boy was I mistaken"*

In the above example only one subjective term "good" was found as sentiment feature which was insufficient to determine the polar orientation of review. Ultimately inadequate performance observed when only sentiment feature was considered.

Combinations of different features with sentiment features were experimented in literature to enhance performance of sentiment classification. Still performance was not remarkable and better representation of feature was required to reinforce the SentiWordNet based sentiment classification.

This is a stimulating question that can be used to measure the effectiveness of the *SentiWordNet* for detecting its limitation on sentiment classification, and how its effectiveness can be increased with the help of other features in this approach.

1.2. Background

Within opinion mining research, sentiment classification pertains to its applicability of automatic methods that are required for predicting sentiment *orientation* existing in the text documents. These predictions are given according to pre-defined values for polarity of sentiments.

For illustration, sentiments pertaining to film reviews can be classified as positive, ("thumbs-up") or negative ("thumbs-down"); author sentiment on text i.e. articles belonging to a particular subject, like [48]. A phrase that co-occurred more with "Excellent" can be expected as positive phrase and the phrase more co-occurred with "Poor" expected as negative. This point wise mutual information- PMI could be used to rank sentiment strength and orientation of reviews.

The problem pertaining to the evaluation of sentiment orientation for the purpose of classification has received considerable research effort and consideration. Numerous approaches are developed for that matter as surveyed and presented in [39]. One of the influential experimentations published in the literature are reported in [37], where commonly known bag-of-words approach a machine learning based method is used for classification of film reviews.

It was observed experientially that the results achieved by utilizing text classification methods relying completely on the bag-of-words methodology seen in [37] persisted below that of conventional topic based text classification, suggesting the extraction of patterns that capture sentiment information in text requires additional linguistic analysis, and has fuelled research effort in the field of sentiment and opinion mining. In [10] experimentation reveals that the employment of higher order n-grams relying on pair of words alongside three word combinations can harvest improved classification results, provided the training data set is adequately large.

Another approach signifying the use of linguistic parts of speech information as mainline features is seen in [42] and [50]. The relationship between sentiment orientation and recognition of subjective and objective segments (i.e. sentences) within a document is presented in [38], with considerable improvements upon the baseline bag-of-words approach.

Ensemble of different features including part of speech and word relation were also explored to improve sentiment classification performance by [53] and [1] compares effect of two word bigrams and bi-tagged phrases (sentiment rich bigrams) along unigrams on sentiment classification performance.

English language lexical resource SentiWordNet [16] specifically designed to assist opinion mining tasks. In SentiWordNet, each term is attributed with positive and negative sentiment information in numerical values. Sentiment features generated from SentiWordNet as a new dimension along with traditional topic based text classification features used to build sentiment classification [15].

Finally, some studies highlighted the deficiencies of SentiWordNet such as limited number of subjective terms and objective words are used in [27]with revised score along with subjective word of SentiWordNet to enhance classification performance.

1.3. Research Objective

To perform opinion mining, a readily available SentiWordNet database could become a valuable resource since it provides sentiment information of the English language terms. It can be noted as well that SentiWordNet maintains the potential for its application to documents on different domains. SentiWordNet offers prospective benefits to opinion mining but only have less than 10% subjective terms. Thus the mainline objective of this research is:

1.4. The Intellectual Challenge

To handle the research objective proposed above, firstly to the design the sets pertaining to features extracted in combination with SentiWordNet that capture comprehensive sentiment information to an utmost extent from text documents. The feature design was based on a comprehensive evaluation of the SentiWordNet database, in identifying and comprehending the limitations of SentiWordNet. So, the main challenge of this dissertation is:

1. New feature dimension consists of content specific syntactic pattern phrases such as verb and noun phrases.
2. Find the best representative feature combination from content specific and sentiment features to enhance sentiment classification performance.
3. Reduce dimensionality and computation by features selection to further increase the performance.

Finally, the ability with which the proposed model is implemented in a text mining package to compare the performance with related studies.

1.5. Research Methodology

In the light of the research objective and intellectual challenges as articulated in the previous sections, the methodology of this research can be enlisted as follows:

- Outline approaches proposed in the literature for performing sentiment classification and especially with SentiWordNet.
- Collection of publically available standardized review data sets.
- Preprocessing performed on the datasets using NLP techniques.
- Features extraction performed from the annotated documents also with the help of SentiWordNet and dimensionality reduction with features selection to find the most discriminative features.
- Train and test a machine learning classifier based on sentiment and content specific features for sentiment classification.
- Obtain the results then analyze and compare with outlined studies in the literature
- Concludes this dissertation and future opportunities are explored.

1.6. Resources

a. Human Resources

- Supervisor and guidance and evaluation committee, for review and guidance.
- Access to supplementary members of NUST research base as required, for resolving more technical queries and questions while sharing ideas.

b. Technical Resources

- Laptop of contemporary specification for setting up to execute experiment.
- Approachability to resources for research in books and periodicals online and remote access to required material should be used at whatever time possible.
- Natural language processing using GATE [11]
- Data Mining Application.: LIBSVM [8]
- Programming language.(C#, Python, XML)
- SentiWordNet Database [16]

1.7. Scope and Limitations

The main focus of this research is evaluating sentiment features from SentiWordNet and designing the content specific information to reduce the limitation of SentiWordNet for sentiment classification and comparing it to other approaches in the literature. To this end, selecting classification algorithms and their required algorithmic parameters are a pre-requisite required for the data mining aspect of the experiment. Whereas potentially better results are likely to be achieved by utilizing a diverse choice of parameter and classifier than the ones presented here. As an alternative, emphasis of the experiment, the choice of features set and their combination will be compared with previous results in the literature.

1.8. Organization of this dissertation

The following chapters of this dissertation are organized as shown in figure 1:

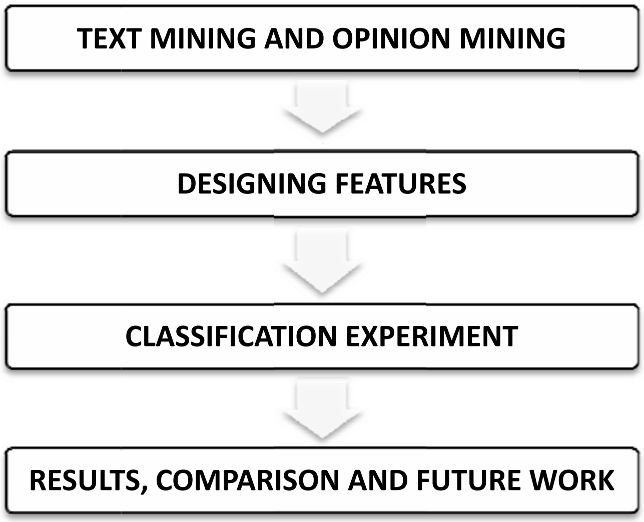

TEXT MINING AND OPINION MINING
DESIGNING FEATURES
CLASSIFICATION EXPERIMENT
RESULTS, COMPARISON AND FUTURE WORK

Figure 1:Organization of Dissertation Chapters

Chapter 2 introduces the area of text and opinion mining, exploring their supplementary challenges and how they can be related to the general domain of data mining. Opinion Mining is the mainline focus of this dissertation. The SentiWordNet lexical resource is presented and potential applications of such a resource are deliberated.

Chapter 3 the first portion of this dissertation’s experimentation is presented: our proposed model for extracting sentiment information from text documents as features for sentiment classification using SentiWordNet based on a comprehensive evaluation of this lexical resource.

Chapter 4 describes the experimental setup for classification, their selection and evaluation while considering performance followed by our experiment setup is laid out, and discussed.

Finally **Chapter 5** presents the experiment results. The obtained results are examined in more details and deliberated in light of related research in the literature. Thanconcludes of this dissertation is presented along with opportunistic and futuristic research directions.

2. TEXT MINING, OPINION MINING

This chapter reviews research literature in the fields of text mining, opinion mining and classification algorithms. It deliberates the motivations for carrying out knowledge discovery on textual data sources while illustrating how the domain of text mining is meticulously related to the generic domain of data mining in general, but with its own supplementary research concerns stemming-out from the need to comprehend and process the intricacies and gradations of unstructured textual data.

The significance of text mining machination to the area of knowledge management is explored as well. Next, the area of Opinion mining with its scope in text is deliberated. Opinion mining is a challenging and relatively new field, concerned about the detection of subjective content from the text, used in multiple real world applications. It is the mainline subject of this dissertation's experimentation, thus a comprehensive survey on the sophisticated approaches to carry out opinion mining is presented, and the research is placed in the contextual setting of the objectives as articulated by this research.

2.1. Text Mining

Text is a rich and usual means of keeping and transporting information, with the Internet being the most notable examples. The situation is alike inside organizations, where a great diversity of textual data encapsulated in emails, memos, wikis, portals and corporate documentations are now completely authored and made accessible in digital format, with approximate estimates representing that 85% of business data is recorded in the shape of unstructured textual documents [33].

The accessibility of information in textual format advocates an opportunity for ameliorating business decision making by relying comprehensively on textual data sources, and the great volumes are thus prospective targets for automated approaches of identifying new knowledge. This opportunity activated the development of the growing area of text based knowledge discovery, or in other words text mining [18]. The fundamental inference of using textual data sources for carrying out knowledge discovery is that data is *unstructured* in nature. Text documents on the contrary are far more malleable and richer in their power of expressivity, but such benefits are inhibited by the increased complexity inherent to the imprecision, vagueness and fuzziness existent in any natural.

For this reason, the discipline of text mining or knowledge discovery in text leverages contributions from diverse research domains in computer science, like artificial intelligence, computational linguistics, information extraction and machine learning. The characterization of what is reflected within the jurisdiction of text mining diverges and sometimes intersects with that of different disciplines that are also related to the computational handling of textual data, like information extraction and natural language processing [31].

In [52] a task focused on interpretation of text mining is anticipated, encapsulating the subsequent aspects:

- **Information Retrieval** or attaining a subset of documents from a corpus centered on user defined search criteria, as perceived on internet search engines besides document searching competences of text knowledge sources.

- **Information Extraction** which compacts with extracting precise information from textual documents, like extracting the time and date an event happened from newspaper documents, or numerically quantified values for a particular characteristic – the price of a particular asset for example. Text summarization methodologies and techniques that intend at provision of a summarized representation of the encapsulated information in a specific document would also come in this category.

- **Text Data Mining** the applicability of data mining methodologies and techniques on text data sources, like classification, clustering and empirical data analysis for the resolutions of extracting new and handy information.

In [18]*Knowledge Discovery in and from Text* is defined as the solicitation of knowledge discovery techniques and methods to text data, closely resembling that of textual data mining articulated above. From the aforementioned argument, the descriptions of text mining in writings can be generally classified in two types: firstly, text mining can be demarcated as all activities encapsulating the handling of text for logical and analytical perspectives, together with extraction and retrieval methodologies; secondly, it can be seen solely as textual data mining along with the purposes of the explanation of knowledge discovery as articulated in [17], while leveraging text as the foundation of data for the detection of new yet unidentified knowledge.

It is pertinent to mention that however in somewhat cases, textual data mining is meticulously related to different research domains encapsulating the computational handling of text, which is not scarce to see the use of text mining approaches and techniques to other linked areas and vice versa: detecting new configurations in text may lure in for assistance of information retrieval. Information retrieval alongside knowledge extraction are expedient methodologies for textual data mining, as will be explained in the succeeding sections deliberating text mining applications alongside its techniques.

2.1.1. Areas of Text Mining

Exploratory Text Analysis

The examination of huge textual data sets is a handy method for achieving an insight from data not commonly possible by labor-intensive inspection. In [35] a system for the interactive analysis of patterns in text is presented, with a particular case study on support tickets where documents can be examined by their relationship to classification categories, urgency and client feedback. Descriptive text mining techniques are applied to improve customer relationship management. Another innovative use of text sets is finding trends in documents formation agreeing to topic, keywords or timeline.

Visualization techniques for exploring documents clustered into topics, and graphical representations pertaining to relationships amongst entities such as companies and executives are presented in [19]. An approach for organizing documents into utilizing clustering methods while relying on a legal documentations data set are articulated in [12].

Information Extraction

Information extraction concerns the identification of relevant information present in text documents\ that can be extracted to a much structured database for advance use, or utilized as supplementary metadata in the examination of textual sources. Automated procedures could be employed for example, to extract company, industry and executive names from news sources to build a searchable database of company particulars.

Systems that carry out information extraction have been applied in law enforcement to aid in the analysis of seized documents, where entities like name, bank accounts and addresses were extracted into a database for the purposes of visual analysis and reporting [52]; In [23], product attributes and characteristics were extracted from textual sources to enrich the content of a decision support system based on transactional data, and had further uses in competitive intelligence and recommendation systems; In [14] another example of attribute extraction from financial news is presented, for the purposes of exploration of documents.

Automatic Text Classification

Text classification encapsulates the application of classification approaches and techniques in textual data for the forecasting of a class for a particular document. One usual use of text based classification is in automatic text categorization in accordance with the topics, as perceived in the categorization of news sources in [28]; an alike example can be grasped in [21] for minimizing manual involvement and fast-tracking while routing the call to the relevant support teams. Supervised classification techniques for text are also employed at the core of many approaches for filtering unwanted content like email spam [30],[34] and [40]. The classification of text for forensic purposes like author proof of identity has also been explored in [13].

Text Mining and Knowledge Management

As presented in the aforementioned examples, text mining expertise open up opportunities for the formation of new knowledge sources from text while enriching data sources with pulling and extracting data from text specific documents and optimizing information retrieval from data sources for decision support. The intensified collections of text data in digital format available in contemporary age's companies put forward those techniques that are useful tools for knowledge management systems.

Applications of text mining technology as an assisting machinery for knowledge management initiatives are outlined in [32], such as automated classification of documents into classification categories, the organization of knowledge stores, and text summarization approaches to alleviate information surplus. Another technique is visualized in [29] where a system pertaining to knowledge management was developed for automatic organization of knowledge hierarchies to expedite the extraction of relevant knowledge.

2.1.2. Processing on Text Data

To realize the benefits of text mining and its applications, approaches are required to address the intricacies and uncertainties of natural language. In addition, a structured processing of text that recapitulates pertinent information pertaining to documents is a compulsory prerequisite for many text mining tasks such as classification and clustering.

Natural Language

Due to its lack of structural configuration, textual data usually experiences a training stage that attempts to recapitulate key mechanisms of natural language for its employment for the text mining task. The treatment applied to the source data will prescribe the model's attributes and information that can be taken out. The table 1 below extracted from [46] articulates the key concerns usually found in processing natural language for data mining, with text preparation encapsulating all but the preceding task.

Issue	**Objectives**
Stop lists	Elimination of terms arising with high frequency and possibly of little significance.
Stemming or Lemmatization	Reducing words to a normalized arrangement, or stem.
Noisy data	Correction of spelling mistakes, word restrictions and unconventional forms.
Tagging	Adding syntactic categories to terms.
Word Sense Disambiguation	Determining meaning of ambiguous terms that best applicable to background of text.
Collocations	Identifying terms defined by numerous words.
Tokenization	Determine policy for combining units of text related information.
Text Representation	Conversion of textual document into a model that preeminently groups pertinent features for text mining.

Table 1: Key issues in Text Mining [46]

One of the foremost concerns in recapitulating pertinent information from text is the establishment of *stop lists*. These lists indicate what terms from the document collection are extremely probable to appear, and bring minimal information when attempting to detect patterns.

Most common words in the English language such as “the”, “and”, “of” are typical candidates for stop lists. Carefulness must be taken nonetheless to develop a list with the precise intentions of the data mining job in mind, as stop words may become pertinent on diverse scenarios.

The mainline goal of *lemmatization and stemming* is to minimize the number of distinctions of a term by converting similar existences to a canonical type, or lemma, or by minimizing words to their inflectional roots, or stem. This will considerably reduce the number of attributes to be analyzed in the text collection, reducing noisy signals and the aspect of the data set. Stemming is exemplified as the conversion of singular and plural orientations alongside present tense and past tense transformation to a single form.

As with any data collected in uninhibited environments, data clean up activity is required to exterminate *redundant data* required to be taken into consideration. In text, data cleansing is an important issue that adopts the form of spelling error and varying spellings amendments, undertaking term restrictions and abbreviations, shedding markup language tags, and transforming text to a lowercase or uppercase where suitable.

There are instances in natural language where the same term may yield different meanings, on the basis of their use within a sentence, the area the document belongs to. Cogitate for example the two virtually distinct senses of the word “book” as narrated in the sentences below:

- *“This is an excellent book.”*
- *“You can book your hotel from international website.”*

Identifying the meaning a precise term speak of in a sentence is recognized as *word sense disambiguation (WSD)*, and is an dynamic research topic in machine translation and natural language processing. This problem has received great attention in the domain of machine translation from novice and early stages, where its intrinsic difficulties were noticed broadly, to obtain reasonable results in word sense disambiguation an increased amount of information is needed about the context such as its role in the sentence and discourse features - and on the existence of external knowledge originating points where sources of sense can be queried.

Tokenization refers to the process of segmenting a textual input into its atomic units. The approach to tokenize a document relies comprehensively on the mining objectives; one common approach is to use individual words as tokens, and spaces and punctuation marks as dividers. However, punctuation marks quite often are part of the examination, as seen in [13] and might instead be used as tokens. *Word collocations* are those terms that are described by multiple words and should be stated to as a component for examination. Collocations can be established by statistical resemblance when analyzing text sets, or plagiarized via information extraction approaches and dictionaries.

Finally, relying on data mining objectives, it is important to determine the grammatical class a term fits into. This can be achieved by attributing tags representing the part of speech that are actually used by a word inside the textual sentence or sentences. A *part of speech tagger* is an application that performs this task. The Brill is part of speech tagger [7] which is one of the most commonly used algorithm based on building tagging rules from annotated documents. Other approaches to part of speech tagging have been proposed using maximum entropy techniques [47] and in developing statistical markov models [6].

The aforementioned discussion articulates the mainline perspectives of natural language processing that are part of a text data mining implementation. In the subsequent section text illustrations for mining text documents are discussed and techniques for text classification are surveyed at a greater level of comprehension.

2.1.3. Document Representation Techniques

Bag of Words

The most usual depiction of documents for textual classification relies comprehensively on the word vector depictions, also stated as *bag of words*, articulated in the previous section. A contemporary example using binary occurrence word vectors and SVM support vector machines applied to spam filtering is presented in [30]. Other applications of this representation to text classification have been surveyed in the literature and can be found in [52]. In [37] linguistic information from parts of speech and adjectives and position extracted from text are employed as features to text categorization of reviews sources.

Bag of Phrases

Observations about bag of words representations are that information available from original text documents are destroyed such that syntactic order information is discarded. The goal of using phrases based representation [43] is to preserve some of the information left out of the bag of words such as high order n-grams and syntactic phrases. A bag of phrases representation has the huge potential increase in the number of features. Plagiarism detection usually n-gram-based searching method is used.

Feature weighting[45]

Bernoulli document model/ Feature Presence:
A document is represented by a feature vector with binary elements taking value 1 if the corresponding word is present in the document and 0 if the word is not present.

Multinomial document model/ Feature Frequency:
A document is represented by a feature vector with integer elements whose value is the frequency of that word in the document.

Term frequency – inverse document frequency (TF-IDF*):*
The numerical statistic of the feature or word is used to represent a document that is intended to shows the relative importance of the word to a document with collection or corpus.

Dimensionality and Feature Selection

As explained earlier, word vector representation of textual documents creates data sets with an increased number of attributes that are liable to concerns related to the profanity of dimensionality. One approach to mitigate this problem is to use linguistics pre-processing like word lists alongside stemming to minimize the number of terms before instigating text classification. Another approach is to employ feature selection mechanisms grasped in data mining to automatically exterminate less pertinent features while slightly affecting classification performance. Studies grasped in [41], [20] report well on using statistical feature selection methods for minimizing the dimension of the feature space for text classification problems.

F-Score - Feature Selection [9]

F-score is a filter type feature selection method used to find the discrimination between two real number sets. The feature is likely to more discriminative with larger the F-score.

Given training vectors x_k; k = 1, 2,…, m and the number of positive and negative instances are n+ and n− , respectively, then the F-score of the i^{th} feature is defined as:

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1} \sum_{k=1}^{n_+} (x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 - \frac{1}{n_- - 1} \sum_{k=1}^{n_-} (x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}$$

Where $\boldsymbol{\bar{x}}, \bar{x}_i^{(+)}, \bar{x}_i^{(-)}$are the averages of the i^{th} feature of the whole, positive, and negative data sets, respectively; $\bar{x}_{k,i}^{(+)}$is the i^{th} feature of the k^{th} positive instance, and $\bar{x}_{k,i}^{(-)}$ is the i^{th} feature of the k^{th} negative instance. The numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets.

2.1.4. Algorithms for Classification

The mainline goal of the section is to present and deliberate predictive algorithms that carry out the data mining assignment of classification they are also defined as *classifiers*. Of specific concern to this research document is the topic of text classification relevant to text mining. This section recapitulates strategies and concerns interfaced in topic based classification of text, and its association to data mining algorithms. In present day research, data centric machine learning methods prevail as the main standard for performing text categorization [44].

Algorithms for classification are concisely surveyed, with emphasis on approaches usually applied in the literature and with greater relevance to text classification. Finally, success criteria metrics to classification are discussed.

Supervised Learning Algorithms

A supervised learning algorithm attempts to predict future values of a particular variable on the basis of information contained on a previously present data set employed for training. The data set encapsulates cases of the variables that we wish to forecast, and it is supposed that future values preserve a certain resemblance to already observed values, which can be “learned” by a supervised learning algorithm. This dependency on the available data as being representative for predictions is worth stressing: that if future values possess no similarity to previously seen data, forecasted results will not be reliable [2] and [52]. Thus, the design of good supervised learning algorithms has a dependency on the data available for training.

Naïve Bayes

The Naïve Bayes classifier employs a probabilistic methodology for forecasting the class of a particular data point. The conditional probability is the starting point of Bayes theorem, such that, for *x* as a data point and *C* as the class:

$$P(C/x) = \frac{P(x/C) \cdot P(C)}{P(x)}$$

Furthermore, for a particular data point $x = \{x_1, x_2, \ldots x_j\}$ by making the assumption, that in a given class the occurring probability of each its attributes is independent, the probability estimate of *x* as follows [24]:

$$P(C/x) = P(C) \cdot \prod P(x_j/C)$$

Training of classifier i.e. Naïve Bayesrequires initially for each attributes occurring for the predicted classes, the conditional probabilities are calculated on the basis of a training data set. It provides the better results with easy estimation of conditional probabilistic interpretation. However, the model's mainline weakness lies on the presumption of individuality of occurrence of characteristic attributes.

Large Margin Classifiers: Support Vector Machines

Support Vector Machines (SVM) are a class of algorithms for classification that belong to parametric methods – that is, identifying an acceptable function that divides the solution space so as to separate the training data points according to the class labels being predicted, in the light of an assumption that future forecast and prediction adopts the similar pattern. Example of a simple case is considered where a resulting hyper plane as the linear function divides the solution space into two classes. The figure 2 illustrates a sample of hyper planes dividing the points into 2 classes:

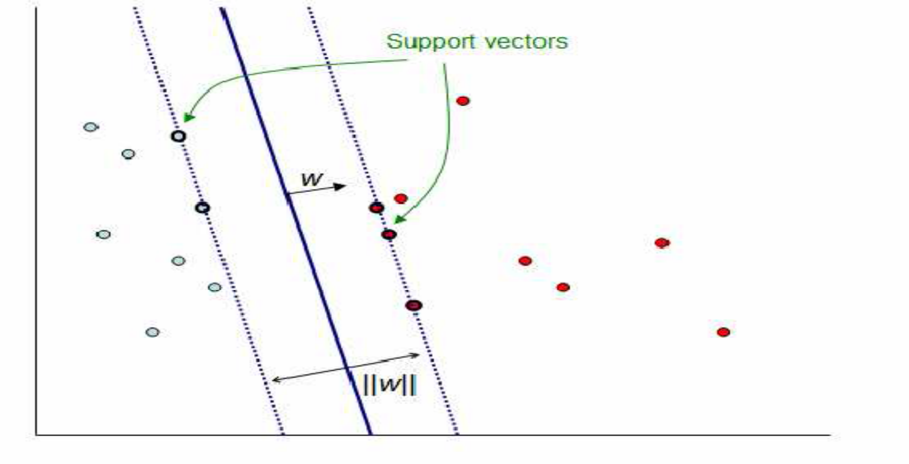


Figure 2: Hyper plane Separating Two Classes

The example shows above, there is unlimited number of hyper planes potentially dividing or separating the two classes. The possible best one has to be chosen, such that a hyper plane with the largest distance would be to choose between two classes from any points, thus SVM also called the maximum margin classifier.

Support Vector Machine algorithm has the objective to find such hyper plane with the maximum margin and first presented in [5]. The SVM optimization problem is defined as

$$\min_{\mathrm{w,b,\varepsilon}} \frac{1}{2}\mathrm{W}^{\mathrm{T}}\mathrm{W} + \mathrm{C}\sum_{\mathrm{i=1}}^{\mathrm{l}} \varepsilon_{\mathrm{i}}$$

s.t. $\mathrm{y_i}(\mathrm{W^T}\emptyset(\mathrm{x_i}) + \mathrm{b}) \geq 1 - \varepsilon_{\mathrm{i}}, \;\; \varepsilon_{\mathrm{i}} \geq 0$

Where b is the bias, W is the weights vector, and ø (.) maps input space nonlinearly into high-dimensional feature space. C > 0 is the penalty parameter of the error term.

Another important feature of Support Vector Machines in dealing with scenarios not linearly distinguishable is its capability to plot the problem space into alternative, possibly more suitable space by means of a *kernel function*, where the points allow for better separation. Numerous kernel functions have been employed to support vector machine classification that has been surveyed in [2].

2.1.5. Evaluating Classifier Performance

Identifying how fine a classifier will make forecasts and prediction on an unseen data is the most crucial aspects of any supervised learning task. A series of approaches and methods for evaluating the classification performance are explored in this section.

Cross-validation

If the performance of classifier is verified against data employed for the training, it can be anticipated that the predictions and forecasts will be expectantly biased, as these are data points previously "seen" by the classifier [24]. Therefore, a better choice is to verify the classification outcomes on a distinct data set, not used for training, also data set can be partitioned into two segments: the *training* segment is a subset employed to train the classifier; while the *validation* segment is used to valuate predictions employing the trained algorithm, and quantify classification results.

An additional extension of this method takes into consideration the fact that while testing only on a specific subset of data, there exists a chance that algorithm will perform unusually good or bad merely by chance, on the basis of the particular data points for every single subsets. To alleviate this issue, the training and testing cycles could be repeated using diverse subsets extracted from the data set, this process is known as *cross-validation*. Idealistically cross-validation is to partition the data set into multifarious subsets, or *folds*, to be used as the test set, though the rest of the data set shall be applied for algorithm training. Therefore A 10-fold cross validation will create 10 cycles for training and testing, hence valuating the algorithm's performance when different partitions of data are employed for training. For each cycle, algorithm performance is measured using sufficient metrics, and inspected individually, or aggregated over all folds.

Performance Metrics

To better illustrate the possible types of classification error, outcomes are frequently shown in terms of correct and incorrect classifications for each class, in a *confusion matrix* [51] as shown in table 2 for a classification problem pertaining to two classes (positive class and negative class).

Predicted Value	**Real Value**	
	Positive	**Negative**
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Table 2: Confusion Matrix for Binary-Class Classification Problem

To warrant the classifier is indeed detecting the right classes, and covering an appropriate number of cases for each class, the concepts of *precision* and *recall* can be employed. These are defined by the formulae below, as depicted on [52]:

$$Precision = \frac{Correct\ Prediction\ for\ Class}{Total\ Prediction\ for\ Class}$$

$$Recall = \frac{Correct\ Prediction\ for\ Class}{Total\ Entries\ for\ Class}$$

Precision indicates the rate at which a classifier makes a correct prediction, or the percentage for which its predictions are correct. A high-precision classifier for the positive class will have high true positives while low false positives. Recall narrates so as to how many forecasts and predictions for particular class are made, once contrasted with the total available cases pertaining to that class. A high recall classifier will have high true positives while low false negatives therefore covering all entries tagged as “positive”. The above formulae can be articulated utilizing the above confusion matrix, as depicted for the “positive” class:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

There is an innate trade-off amongst precision and recall: increasing the precision variable of a classifier, it becomes more specific and therefore more “conservative” while predicting, ultimately, lowering recall. On the contrary a high recall classifier might be adjusted for predicting more “generously”, at the cost of precision. Consider the loan applications data set to be an example for a high-precision and low-recall loan application classifier will make few loan approvals, while taking correct decisions comprehensively most of the time. On the contrary, A low-precision and high-recall loan application classifier will make incorrect predictions at a greater frequency, and is more prospective to identify a positive loan application.

Evaluation Metrics: Accuracy and F-Measure

Combined metrics integrating precision and recall statistics are sometimes employed to report research outcomes. The *accuracy* states the complete classifier precision transversely across entire classes, and is defined by the formula:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + True\ Negatives + False\ Negatives}$$

In other words: rate for correct predictions over total i.e. whole predictions. Accuracy is 1 if classification bears no errors as reported by the classifier. It however agonizes from the same concerns visualized on misclassification rates, for a data set where a classification category contains many more instances than another can create biased results. To alleviate this issue, the harmonic mean of variables like precision, recall and *F-measure* is often computed and used:

$$F - Measure = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

The selection of performance measurement should take into consideration the data set and orientation of the prediction problem being investigated: classification precision variable might be of more importance than recall variable, for example on diagnostic systems which is equipped with a high risk of misclassified incidences might prefer increased precision for classification at the cost of recall. In the literature, F-measure is a common metric for reporting classification performance outcomes, as presented in [44],accuracy is often reported for such cases where data set represents a balanced number of instances for both positive and negative classes, as depicted in [37,38].

2.2. Opinion Mining

Opinion Mining is a novel and exhilarating domain of research concerned with extracting opinion related facts from textual data repositories. It maintains the prospective for multiplicity of interesting applications both in commerce and academic arenas, and poses innovative cerebral challenges, which lingers to attract substantial research attention. This section articulates the research domain of opinion mining while introducing, its major motivations, baseline tasks and challenges by discussing them in great detail. Finally, the *SentiWordNet* lexicon for opinion mining is depicted, along with its potential benefits, applications and limitations are deliberated.

2.2.1. Opinions in Text

Information related to people's opinions could be a very significant component for more precisely accurate decision making in a diversified set of domains. Companies, for example, maintain a keen interest in identifying what are the opinions of their customers in accordance with the newly launched product on a marketing campaign. Consumers on the contrary would benefit by accessing other people's opinions alongside reviews on a particular product they are planning to procure, as recommendations from different users tend to play vital part in influencing the purchasing decisions. Knowledge of different user's opinions is also an important feature in the political dominion, where for instance, one can identify the sentiment directed towards a completely or partially new section of legislation, or an individuals like politicians or activists.

The internet is obviously a gigantic source of publicly available user produced content dedicated to expression of opinions pertaining to any topic of interest. Sources for opinions are not restricted to particular review sites, and are encapsulated in user blogs, discussion forums and online social networks. Furthermore, opinions are usually expressed in textual format, making it an enriched ground for the applicability of text mining and related approaches to analyze natural language. Therefore, the motivating requirement to analyze huge volumes of opinion material, coupled with advent in natural language processing along with machine learning frameworks gave rise to research effort that is directly focusing in the evolving domain of *Opinion Mining*.

Opinion Mining is attributed with application of computational methods for the identification and measurement of *opinions, sentiments* and their *subjectivity* in a text [39]. A text document can be visualized as a collection pertaining to both objective and subjective statement; here objective statement refers to factual information encapsulated in the text, while subjectivity is concerned with the expressivity of opinions, along with their evaluations and speculations.

In an effort to map the happenings of the evolving domain of opinion mining a research survey by [39] categorizes the domain into two generic sub-domains i.e. classification and extraction. Classification is entailed with research directly related to identification in the first place if a segment of text could be categorized as one of the two i.e. subjective or objective, if the texts is subjective, the ability to correctly forecast and predict the text's polar orientation. A similar formulation is presented in [16], where the primary objectives of opinion mining are categorized into 1) identifying the degree to which a particular text is objective or may be subjective; 2) identifying whether its expressivity is positively or negatively biased, if a text is certainly subjective; and 3) identifying the degree of *strong points* of the polarity for a given subjective piece of text.

For the resolutions of this research, mainline focus of this analysis is towards the predictive perspectives of opinion mining concerned with the tasks pertaining to subjectivity identification and sentiment classification of text. It is acknowledged though that opinion mining is a pertinent part of this domain and one that completely goes hand in hand along with opinion detection and its classification methods, therefore any pertinent research on opinion mining will be articulated where suitable to the argument.

2.2.2. Subjectivity Detection

In order to identify subjectivity in text in an automated manner, a computational model that requires a formalization of what is understood by the concept. In [50], the subjectivity of a sentence is defined based upon previous work in linguistics and literary theory. Firstly, there are *subjective elements*: the linguistic expressions that characterize private states of mind. Characterizing subjective features is not an inconsequential task; they may surface in text as single words, expressions or entire sentences, may depend on the context, and may also be obvious in text format. A subjective component expresses the opinions, thoughts and conjectures of a *source*, that is the document author or somebody mentioned in the text. Lastly, a subjective part has a *target*, or the object being referred to.

There have been several approaches proposed in the literature to detect elements of subjectivity on text. In [50] a method is proposed on the basis of an exploring word relations learned from an annotated corpus of subjective expressions. Subjectivity is annotated manually at expression, sentence and document stages, and employed to train identification approaches based on term presence and term collocation, or term's position in the text relative to each other.

Another approach to detecting subjectivity is proposed on [38], where machine learning classification models are trained to predict and forecast objective or subjective sentences on the basis of training set comprising extracted documents from the internet. The subjective data set comprises of 5000 textual extracts from film reviews, while the objective data set is developed from 5000 extracts pertaining to film plot summaries.

A comparable approach is depicted in [54] where a Naïve Bayes classifier is trained to detect subjective documents on the basis of data set of news known a priori to contain objective (news or business sections) while subjective (editorials or letters to the editor) type content, with promising results. The approach is stretched to sentence-level opinion mining by comprising parts of speech, sentence resemblance measures and including the existence of semantically orientated terms from a subset belonging to a manually annotated seed words.

2.2.3. Sentiment Classification

Sentiment classification deals with identifying what, if any, are the sentiment's *orientations* of the opinions that are encapsulated in a particular document. It is considered in a generic perspective that a document being inspected is known to represent opinion, such as a product review, and that the document's opinion is stated as a single entity [39].

Parts of Speech

POS information is supposed to be a significant indicator of sentiment expression. The presence of adjectives means more sentence subjectivity [25] and adjectives and adverbs are better than adjectives alone [4]. Only adjective features classification results are not as remarkable and adverb, verbs and nouns along adjectives improve sentiment classification [37]. The employment of parts of speech for a pre-processing stage for stemming features for opinion extraction has also been observed in numerous other sentiment classification experiments.

Unigrams, Bi-grams, adjectives, POS and Position

In [37] a series of experimentations utilizing unigrams, bi-grams, adjectives, and unigrams with POS position information classes pertaining to word vectors for sentiments classification of films reviews. Word vectors, for each entry is mapped to a term identified in the corpus of documents and value of a given terms corresponds to an enumeration of term presence or an enumeration of relative term frequency. Traditional approach for text mining used for sentiment classification. Binary presence performs better than frequency-dependent word vectors, articulating that feature presence, rather than frequency is more significant to opinion identification. Using only bigrams the accuracy actually falls than only unigrams. Accuracy improves if all the frequently occurring words from all parts of speech are taken, than only adjectives.

Phrases, Syntactic Part of Speech Patterns

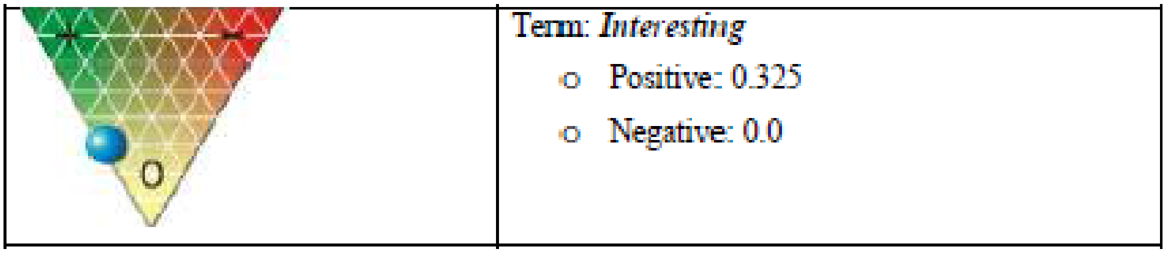
In [48] a phrasal lexicon was extracted from reviews based on two word part of speech patterns with adjective because adjectives are considered better indicators for opinion information. Learn polarity of each phrase such that positive phrases co-occur more with "excellent" and negative phrases co-occur more with "poor" than rate a review by the average polarity of its phrases. It used phrases instead of word and got best results for automobiles domain and worse results for movies domain.

Bi-tagged phrases are sentiment rich bigrams based on two-word POS patterns in [1] extends the [48] work and utilize these phrases in supervised classification algorithms as features. Bi-tagged phrases features as new dimension instead of the bigrams and combination with unigrams are experimented. Best F-measure observed of combination of bi-taggedwith unigrams on movies reviews.

2.2.4. Lexical Resource: SentiWordNet

One interesting perspective of approaches dependent on word list is that it doesn't necessarily requires training data to make predictions, as it relies only on pre-defined sentiment lexicons, therefore being applicable to scenarios where training data is not present. For this reason these methods are often labeled as *unsupervised learning* approaches [39].

One example of a lexical resource considered to support in opinion extraction tasks is *SentiWordNet*[16]. The objective of the SentiWordNet is to provide sentiment bias information at the term level. English terms information is a derivative of the WordNet database using the semi-automatic methods. In SentiWordNet, each term have the numerical values of its positive and a negative polarity in the range from 0 to 1. Sum of these polarity score shows the subjectivity of the term and lower score means that the term is less subjective. In figure 3, a term “interesting” is illustrated with its positive and negative extracted from SentiWordNet.



Term: *Interesting*

- Positive: 0.325
- Negative: 0.0

Figure 3: SentiWordNet Sample Score (http://sentiwordnet.isti.cnr.it)

Applying SentiWordNet

Issues stemming from ambiguity in word sense also ascend on opinion mining related problems. Data in SentiWordNet is stratified according to the parts of speech tags, and there exists considerable variances in the level pertaining to objectiveness a synset might exist, depending on its attributed grammatical role. Raw level of word sense disambiguation (WSD) was followed such that part of speech information of a term is considered to accurately apply SentiWordNet scores.

Sentiment features were new dimension introduced by [15] generated from subjective terms present in SentiWordNet. They experimented on movie reviews to validate that addition of sentiment features with content free features and content specific unigrams and bigrams could enhance the classification performance.

Limitation of SentiWordNet

Objective words are used in [27]with revised score along with subjective word of SentiWordNet to enhance classification performance. This work reevaluates objective sentiment words in the SentiWordNet because more than 90 % of the words are objective. According to the experiments, the average accuracy is 71.89% for the original SentiWordNet and 76.02% for the revised SentiWordNet. The revised SentiWordNet outperforms the original SentiWordNet evaluated by 4.1%.

2.2.5. Combining Approaches

Visualizing various methods for carrying out sentiment classification encapsulates various types of sentiment centric information from documents, it is pertinent to mention that the contribution in the literature to combining different aspects of text in order to achieve better results.

Ensemble of features (POS based & Word relation) and ensemble of classifiers is studied in [53]. They proposed to improve classification performance by the ensemble of features and not only just combining them. Classification performance of machine leaning algorithm often varies from domain to domain; hence combine their outputs for a better integrated output. Three ensemble methods used for features sets and classification algorithms and for both named as Fixed Combination, Weighted Combination and Meta-classifier combination. They achieved better with the ensembles word relational features than ensemble of POS-based feature. They experiments on the same 5-dataset of reviews that we used in this research.

2.2.6. Different Features Considered in Literatures

Following are the different features considered in literature as shown in table 3:

Study	**Features**
Turney 2002 [48]	Phrases based on POS patterns
Pang, et al. 2002 [37]	Unigrams, Bigrams, POS and Position
Hung, et al. 2013 [27]	SentiWordNet Subjective Terms and Objective (after revised score)Terms
Xia, et al. 2011 [53]	POS(Adjectives, Adverbs, Nouns and Verbs) and Word Relations (Unigrams, Bigrams & Dependencies)
Dang, et al.2010 [15]	SentiWordNet Subjective Sentiment features , Content-specific unigrams & bigrams and Content-free features
Agarwal, et al.2013 [1]	Unigrams, Bigrams and Bi-Tagged Phrases (based on POS patterns)

Table 3: Different features used in literature

2.4. Conclusion

The research domains of text and opinion mining were explored comprehensively. Text mining carries out the computational treatment of text for extraction of novel information, and leverages techniques from machine learning, natural language processing, information retrieval and computational linguistics. Applications of text mining to knowledge discovery were surveyed, based on exploratory analysis and other traditional data mining approaches and techniques.

The representation of documents for performing text mining was studied in more details, with the words vector, bag of words approach being the most popular methods for representing text for machine learning, demonstrating comprehensively efficacious empirical results.

The research area pertaining to opinion mining was introduced. It is a new area of research leveraging elements from data mining, textual data mining alongside natural language processing (NLP), and a wide spectrum of applications for extracting opinions from documents is probable, as articulated in this chapter. These methods range from ameliorating business intelligence (BI) in organizations to information extraction / retrieval systems, followed by recommender systems and more effective online advertising and spam detection.

A SentiWordNet lexical resource was introduced, with a depiction of its building blocks alongside potential employments. SentiWordNet is an add-on of the famous WordNet database of terms alongside their relationships, and is a freely available lexicon of terms sentiment information, which can be employed in opinion mining research where numerous similar approaches were developed in an ad-hoc fashion.

The ultimate outcome is the different designs of a sets pertaining to features that leverage sentiment information and applied to sentiment identification and classification problems in literature.

3. DESIGNING FEATURES

3.1. Introduction

SentiWordNet [16] is english language lexical resource used as a tool to perform sentiment classification, from textual documents as much as possible sentiment information needs to be devised as the set of sentiment features. Then, once a feature set is generated from text documents with SentiWordNet, these sentiment features in the input to a classification algorithm and classification performance results can be analyzed. SentiWordNet database and its structure is look at in this chapter and the data sets of product and film reviews are analyzed in detail. The considerations for data preparation and step wise text preprocessing will initiate the requirements to generate an informative text mining exercise.

The outcome of the current chapter is a specification of the feature sets that takes amazon products and film reviews as the starting point to captures sentiment and content specific information present in the text. These feature sets as an input can be used to train supervised learning classifier to perform sentiment classification.

3.2.Data Sets of Reviews

3.2.1 The Multi Domain Sentiment Data Set

The multi-domain sentiment dataset of product reviews (book, DVD, electronic, and kitchen appliances) taken from amazon. This sentiment dataset was first used by John Blitzer, Mark Dredze and Fernando Pereira in their research presented in [3]. The data set contains 1000 negative and 1000 positive labeled reviews for each domain. Each review have the information consists of a rating (0-5 stars), a reviewer name and location, a product name, a review title and date, and the review text. Reviews with ratings > 3 were labeled positive; those with rating < 3 were labeled negative. The dataset is public and available at

- URL: http://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html

3.2.2 The Polarity Data Set

The polarity data v2.0 set is a set of film review documents available for research in sentiment orientation mining and opinion analysis. This data set was first introduced by Bo Pang and Lee as a research on sentiment classification using machine learning methods in [37]. It comprises 1000 positive labeled and 1000 negative labeled film reviews extracted from the Internet Movie Database Archive [38]. A film review from the polarity data set already go through multiple preprocessing tasks aiming at standardizing the text [38]:

- All text initially converted to lowercase.
- Each sentence corresponds to a new line.
- All HTML are removed from the text i.e. documents only contains plain text.
- Text is also free from ratings information: rating information corresponds to labels
- In a positive review corresponds to a letter “B” grade or above while “C” or lower is designated as negative review.

The dataset is publically available at http://www.cs.cornell.edu/people/pabo/movie-review-data/

3.3. Proposed Model

In the next sections of this chapter, the SentiWordNet structure was assessed in details, and considerations were made on challenges and limits of how opinion relevant information can be gathered. The approach for a feature set proposed in this section however starts from the principle that the features obtained should capture diverse aspects of document sentiment.

Syntactic Grammar and Phrases

Sentence structure is viewed in terms of the constituency relation [36] as shown in figure 4. Sentence encapsulates phrases that are the combination of words those act as a single POS in a sentence. The noun and verb phrases are considered as the important constituent phrases in each sentence.

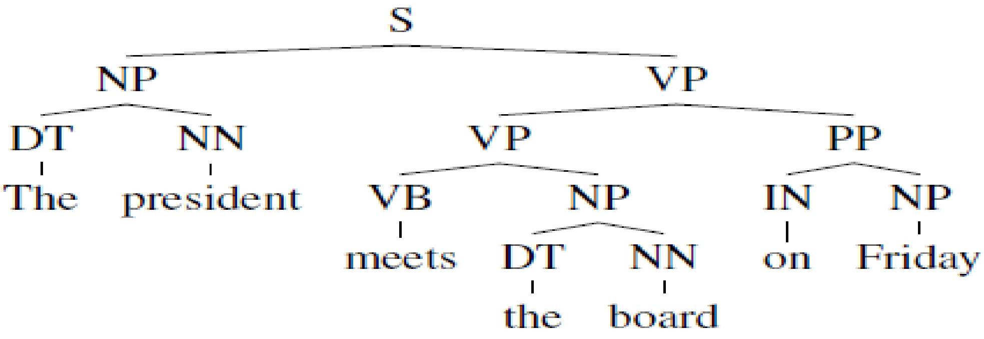


Figure 4: Phrase Structure Grammar- constituency relation

Noun phrases are the important key words as object and subject while carrying the most important information. Verb is the skeleton of any sentence but mostly verbs are objective in meaning. However when verbs are combined with their dependents they form verb phrases that are more meaningful.

Unigram appeared in text classification represent the actual content in BOW model and also witnessed by [37] as best performance achieved by unigrams. Unigrams also mostly utilized in sentiment analysis features due to its benchmarked performance in literature all across.

Currently most frequent verb phrases, noun phrases along the unigrams is proposed as features to better represent the content of text. The limitation of SentiWordNet could be reduced by considering the proposed features.

The table 4 shows the currently proposed features and how better these feature represent the previously following review. Phrases "the foulest one" and "mistaken" were also tagged by the proposed method but discarded due to their lower frequency in corpus.

> *"I've had my share of dutch ovens in my time, and I have to say that this is the foulest one yet. I thought I smelled a good deal when I got it, but boy was I mistaken"*

Feature Type	**Features Words**
Sentiment Feature	"good"
Unigrams	"get", "think", "say", "time", "boy", "smell", "oven", "yet", "deal", "dutch", "share"
Syntactic Phrases (NPs & VPs)	"get", "think", "say", "time", "boy", "smell", "oven", "have to say", "my time", "a good deal", "dutch ovens"

Table 4: Addition of new feature type and feature words

Once a feature set is generated it can be used as starting point to train supervised learning algorithm of sentiment classification. Our proposed system mainly consists of following three phases as shown in figure 5 and 6.

1. Text Pre-processor
2. Feature Extractor
3. Sentiment Classifier

Text Pre-Processor and Feature Extractor are covered here in this chapter however Sentiment Classifier will be covered in the next chapter.

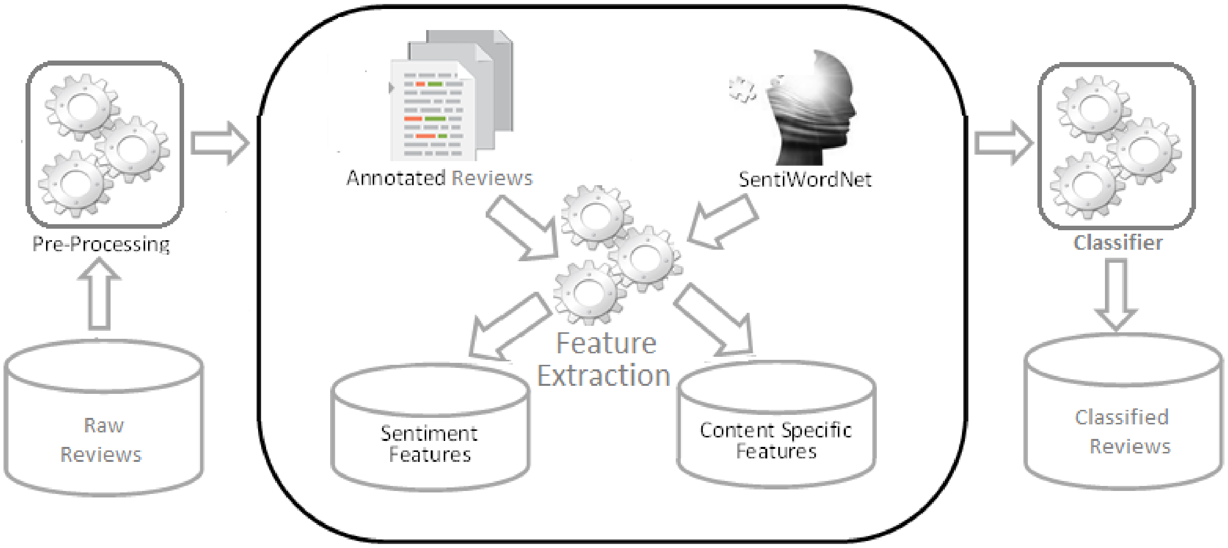


Figure 5: High Level Proposed Model

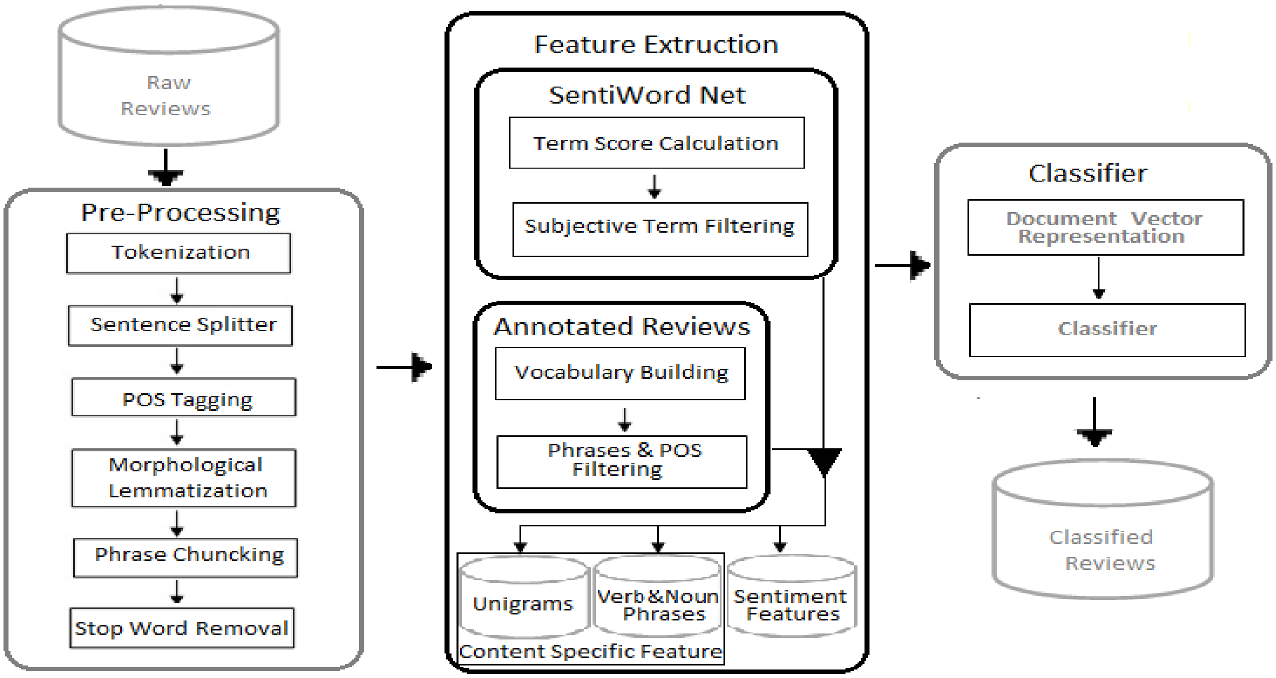
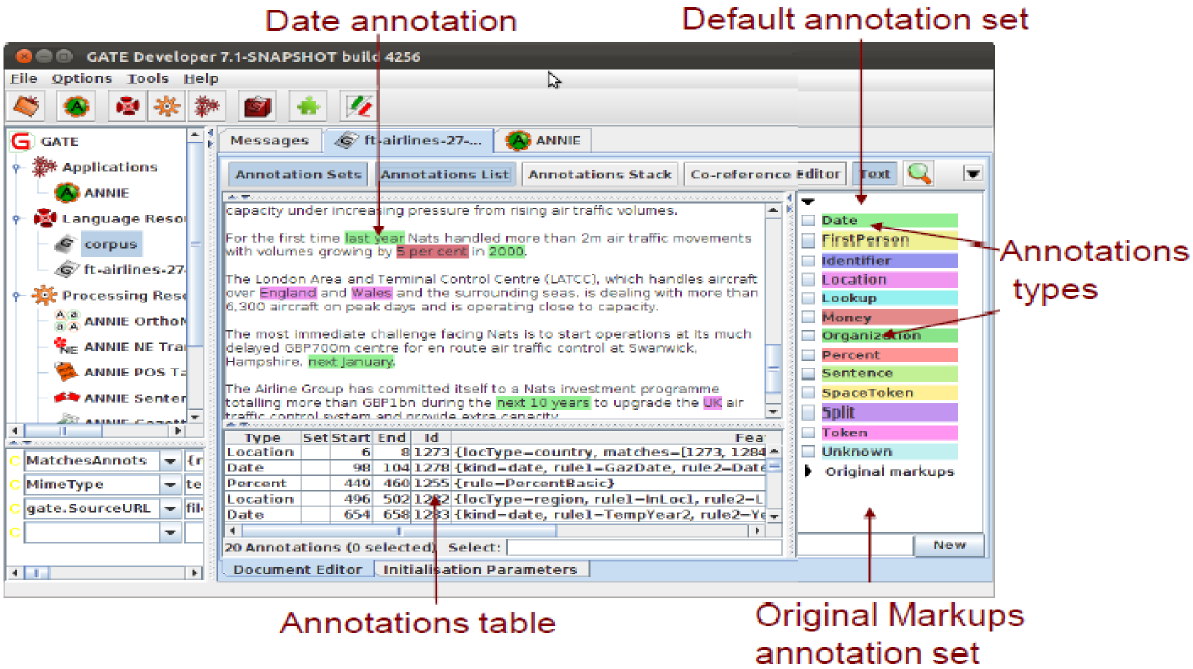


Figure 6: Detailed Proposed Model

3.4. The Preprocessing Process

General Architecture for Text Engineering-GATE[11] for text preprocessing is used which is open source and widely used by many research communities. Purpose of the process is the transform the text that can be used for further text engineering activities. A Nearly-New Information Extraction System (ANNIE) is the information extraction application available in GATE and used in our preprocessing process with defaults options. The preprocessing activities used in this process are:

- **Tokenization:**ANNIE English Tokenizer used for tokenization.
- **Sentence Splitter:**ANNIE Sentence Splitter used to set the scope of sentences.
- **Part of Speech (POS) Tagging:**ANNIE POS Tagger used to identify each token with appropriate part of speech.
- **Morphological Analyzer:**GATE Morphological Analyzer used with ANNIE to consider the root of each word instead of original string in feature extraction.
- **Noun Phrase and Verb Phrase Chunker:** Extraction of Noun and Verb Phrase based on the syntactic pattern from text.
- **Stop Word Removal:**Most frequent and less important stop words are also removed using the list of 233 word list.



capacity under increasing pressure from rising air traffic volumes.

For the first time last year Nats handled more than 2m air traffic movements with volumes growing by 5 per cent in 2000.

The London Area and Terminal Control Centre (LATCC), which handles aircraft over England and Wales and the surrounding seas, is dealing with more than 6,300 aircraft on peak days and is operating close to capacity.

The most immediate challenge facing Nats is to start operations at its much delayed GBP700m centre for en route air traffic control at Swanwick, Hampshire, next January.

The Airline Group has committed itself to a Nats investment programme totalling more than GBP1bn during the next 10 years to upgrade the UK air traffic control system and provide extra capacity.

Type	Set	Start	End	Id	Fea
Location		6	8	1273	{locType=country, matches=[1273, 1284
Date		98	104	1278	{kind=date, rule1=GazDate, rule2=Date
Percent		449	460	1255	{rule=PercentBasic}
Location		496	502	1282	{locType=region, rule1=InLoc1, rule2=L
Date		654	658	1283	{kind=date, rule1=TempYear2, rule2=Ye

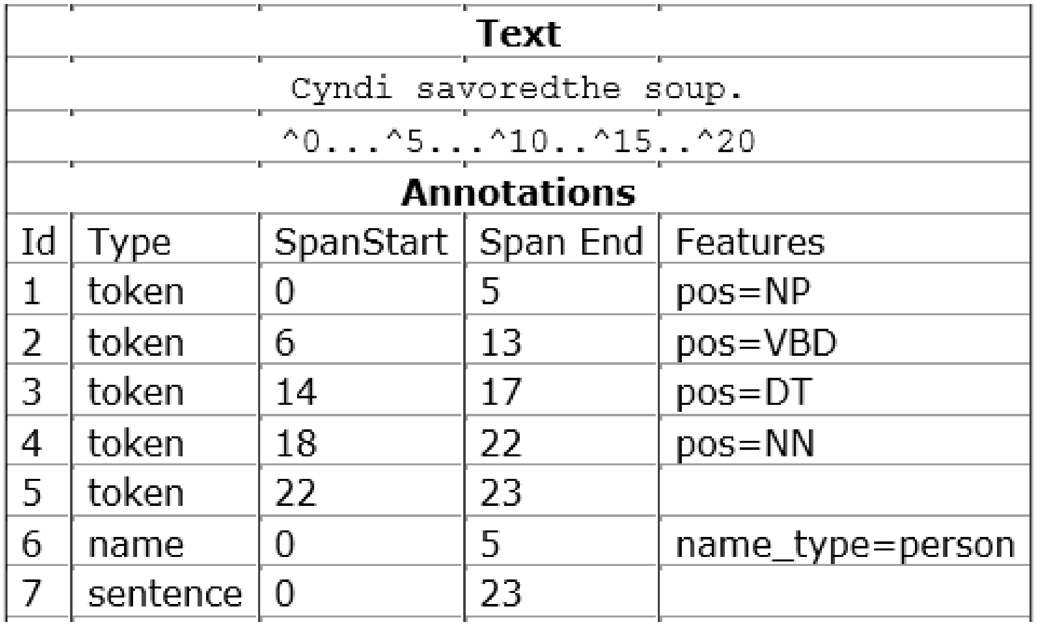
20 Annotations (0 selected) Select:

Figure 7: GATE GUI [11]

3.4.1. Output of Prepossessing: The Annotated Reviews

In GATE the Language Resource (LR) are of two types; Documents and Corpus and documents are members of the Java Set corpus. A Feature Map is the Java Map associated with both document and corpus, stored with attribute/value information. Arbitrary information also associated with the feature maps via the annotation model used for kinds of documents. The text content in a document is present with one or multiple annotation sets. Documents are exhibited with content plus annotations along the present features.

The example shown below as figure 8, a single sentence is illustrated after the preprocessing activities tokenization; sentence splitter followed by part-of-speech tagging and name entity recognition. Each token showed with part of speech (POS) as feature in the figure 8.



Text

```
Cyndi savoredthe soup.
^0...^5...^10..^15..^20
```

Annotations

Id	Type	SpanStart	Span End	Features
1	token	0	5	pos=NP
2	token	6	13	pos=VBD
3	token	14	17	pos=DT
4	token	18	22	pos=NN
5	token	22	23	
6	name	0	5	name_type=person
7	sentence	0	23	

Figure 8: GATE Annotation Example

After preprocessing reviews document have annotations encoded and preserved in xml format as:

```
<Annotation>
<Feature>
<Name className='java.lang.String'>Id</Name>
<Value className='java.lang.String'>2</Value>
</Feature>
```

```
<Feature>
<Name className='java.lang.String'>Type</Name>
<Value className='java.lang.String'>Token</Value>
</Feature>
<Feature>
<Name className='java.lang.String'>StartNode</Name>
<Value className='java.lang.String'>1</Value>
</Feature>
      ...
```

3.5. The Feature Extraction Phase

A customized application build using Object Oriented Programming C# for reading and analyzing XML type SentiWordNet and annotated documents.

3.5.1. The SentiWordNet Database

SentiWordNet is a database contains list of english terms with opinion scores for these terms. SentiWordNet was built from the WordNet version 2.0. A semi-supervised method is used to build it by obtaining opinion polarity scores of seed terms from a subset that have known opinion polarity. Similar meaning terms are grouped into *synsets* and a glossary define the relevant meaning of the associated terms. There is numerical values are associated with the synset to show its positive, negative and objectiveness bias. This numerical value indicates the synset positive or negative score, ranging from 0 to 1. A synset have a numerical ID which uniquely identified it for a specific part of speech. A synset appeared with only four possible parts of speech i.e. Adjective (a), Noun (n), Verb (v) and Adverb (r).

In SentiWordNet following information regarding a synset is available:

- *PosScore:* Synset positive score.
- *NegScore:* Synset negative score.
- *ObjScore:* Synset objective score.

Following scoring rule been applied in SentiWordNet:

$$PosScore + NegScore + ObjScore = 1$$

The objectiveness of a synset is calculated as:

$$ObjScore = 1 - (PosScore + NegScore)$$

SentiWordNet Structure

It is provided as a text file in which terms of similar meaning and of same part of speech are grouped in a synset. The table 5 describes a synset and the available columns information for each entry in the SentiWordNet database.

Field	Description
POS	There are only four possible part of speech that may appear with a synset as: • adjective (a) • noun (n) • verb (v) • adverb (r)
Offset	A synset have a numerical ID which uniquely identified it for a specific part of speech.
PosScore	A numerical value indicates the synset positive score, from 0 ~ 1.
NegScore	A numerical value indicates the synset negative score, from 0 ~ 1.
Synset	It is the terms list for this synset.

Table 5 : Record Structure of SentiWordNet Database

How to mine polar information from the SentiWordNet, the table 6 presented the rows as appeared in SentiWordNet:

POS	Offset	PosScore	NegScore	Synset
a	1001456	0.375	0.125	Casual, everyday
n	13488485	0.0	0.125	Pull, twist, wrench
v	1248670	0.125	0.0	Truss, tie_up, bind, tie_down
r	326136	0.375	0.25	Dreamily, dreamfully, moonily

Table 6: Sample SentiWordNet Data

Part of Speech (POS) Designation

Data is categorized in SentiWordNet as per part of speech of english terms, as seen in Table 6. A synset have considerable differences in terms of level of objectiveness, using the grammatical role i.e. part of speech. So, part of speech information from the source documents being essential needed, so that accurate score from SentiWordNet can obtained. To achieve this, POS *tagging* algorithm can be employed to automatically classify words from the source documents into categories based on POS.

Each term with a relevant POS tag has been associated which indicates its role in the sentence, such as verb, noun, adjective, etc. For simplicity and later comparison we convert the tagger POS designation into SENTIWORDNET format and following table 7 equivalent POS.

SentiWordNet POS	Included POSs from Tagger
A	JJ, JJR, JJS, JJSS
R	RB, RBR, RBS
V	VB, VBD , VBG, VBN, VBP, VBZ, MD
N	NN,NNP, NNS, NNPS

Table 7: Tagger POS conversion for POS present in SentiWordNet

Term Score Calculation

When evaluating scores for a given term using SentiWordNet, an issue arises in determining to what specific WordNet synset the term belongs to and which score to take into account. Every entry in the SentiWordNet takes the form term#sense. Obviously, different word senses can have different polarities as shows in table 8:

POS	ID	PosScore	NegScore	term#sense
R	00011093	0.375	0	well#1
R	00012531	0.5	0	well#3
R	00013092	0.75	0	well#6
R	00013626	0.125	0.25	well#12
R	00012129	0.667	0.333	well#13

Table 8: SENTIWORDNET Score against Sense

To calculate a term score different prior polarity calculation formula are discussed by [22] but we used AVE in our experiments however both the First and Average are frequently used in previous studies , defined as follows:

FS (First Sense). The first sense is considered from the n-senses for the given term against for the specific POS as.

$$PosScore = PosScore_1 and NegScore = NegScore_1$$

AVE (Average). It calculates the mean of all the n-senses for the positive and negative score of the given term against for the specific POS as.

$$PosScore = \frac{\sum_{i=1}^{n} PosScore_i}{n} and NegScore = \frac{\sum_{i=1}^{n} NegScore_i}{n}$$

Subjective Terms Segregation

A term with higher PosScore value considered as positive else considered as negative, However if sum of both PosScore and NegScore values more than a given threshold value is considered as subjective. If the averaged positive and negative scores for a term are below than given threshold, it is assumed that a decision cannot be made on term orientation and the term is said to be objective.

Usually subjective terms are more meaningful for sentiment classification task and are used as features. We consider threshold 0.5 with both FS and AVE prior polarity calculation formulato differentiate subjective and objective terms and we designate a term with specific POS either subjective or objective with following formula also we compare both FS and AVE to estimate no of subjective terms in SentiWordNet as sown in table 9.

$$Term_i = \begin{cases} subjective \quad if\,(PosScore_i + NegScore_i) > 0.5 \; \boldsymbol{AND} \; (PosScore_i \,!= \, NegScore_i) \\ \qquad\qquad\qquad\qquad else \\ objective \end{cases}$$

	Subjectivity $(PosScore_i + NegScore_i) > 0.5$		Also removing Neutrals $PosScore_i \,!= \, NegScore_i$	
	No of Subjective Terms	**No of Objective Terms**	**No of Subjective Terms**	**No of Objective Terms**
FS	13849	141438	13052	142235
AVE	12351	142936	11678	143609

Table 9: No of Subjective and Objective

FS have more subjective terms than AVE even after neutral terms are excluded from the sentiment features.

3.5.2. Getting Features from Annotated Document

Vocabulary Building

A vocabulary of terms was constructed for each dataset using four part of speech groups i.e. Adjectives (a), Adverb (r), Verb (v) and Noun (n) as shown in table 10, each row in the vocabulary have following format

POS: Term: Frequency: IsPhrase

Vocabulary Size	books	Dvd	electronics	kitchen	movies
Unigrams	15782	14737	8249	7705	38988
Phrases	54625	53297	31142	24987	150133
Total	70407	68034	39391	32692	189121

Table 10: Vocabulary Size

3.5.3. Content-Specific features

Unigram appeared in text represent the actual content in BOW model and also from [37] unigrams were considered as best in performance. We proposed noun and verb phrases are the new dimension of features based on the syntactical patterns with Unigrams are used as content specific feature after filtering less frequent terms. The part of speech group Adjectives (a), Adverb (r), Verb (v) and Noun (n) are filtered based on the frequency with the threshold such as:

POS {‘a’, ‘r’, ‘v’} frequency >3 and
POS{ ‘n’} frequency > 4 { as mostly noun are not sentiment bearing words}

Filtered size	books	Dvd	electronics	Kitchen	movies
Unigrams	4403	3938	2498	2361	13424
Phrases	4408	4607	2601	2132	11700

Table 11: Content Specific Features Size

3.5.4. Sentiment features

We form the final sentiment features by filtering such that remove such sentiment features (terms with frequency < 2) in the vocabulary. Sentiment features are considered as base line feature.

3.5.5. Features Sets construction with/without Feature Selection

Most frequent content specific features are gathered after filtering and then combined with above finalized sentiment features. Some content specific features are already parts of sentiment features, so we also remove duplication at this stage. Ultimately we formulate the four combinations from sentiment features and content specific features. To select more discriminative feature sets we calculate F-Score for each feature with the tool [9] available in LIBSVM as python script i.e. fselect.py and then such feature are selected having greater 0.002 F-Score value. Abbreviations and their description for all features sets and no of features are as shown in table 12 and 13.

Feature Set	Description
Senti	Sentiment Features
SentiUni	Joint Sentiment and Unigrams Features
SentiPhr	Joint Sentiment and NP & VP Features
SentiUniPhr	Joint Sentiment , Unigrams and NP & VP Features
SelSenti	Selected Sentiment Features
SelSentiUni	Selected Joint Sentiment and Unigrams Features
SelSentiPhr	Selected Joint Sentiment and NP & VP Features
SelSentiUniPhr	Selected Joint Sentiment, Unigrams and NP & VP Features

Table 12: Feature Set Descriptions

Features Set	Books	Dvd	Electronics	Kitchen	Movies
Senti	935	855	312	302	1886
SentiUni	4878	4354	2649	2502	14021
SentiPhr	5278	5392	2889	2419	13441
SentiUniPhr	7951	7679	4432	3952	21846
SelSenti	101	124	68	61	273
SelSentiUni	599	645	488	477	2211
SelSentiPhr	613	704	564	457	1971
SelSentiUniPhr	935	1016	785	713	3213

Table 13: No of Feature for a feature sets

3.6. Conclusion

In this chapter lexical resource SentiWordNet [16], and the data set of different reviews were analyzed in more details, with the objective of determining how to best use SentiWordNet to build a model that represent opinion information from text documents. The highlighted need to analysis done through avail of natural language processing techniques such as part-of-speech tagging to enrich the model, as well as potential limits of using lexical SentiWordNet score information.

Raw WSD becomes relevant, since terms may carry multiple meanings with potentially different opinion bias depending on context and their use within a sentence. Domain-specific knowledge is also an issue, since a different bias may be indicated than the more commonly use and seen. The above issues naturally impose restrictions to the effectiveness of SentiWordNet for sentiment classification. Addition of Content specific information is the proposed solution with sentiment information to enrich the contextual background of each domain.

The outcome of this chapter is the specification of features that reflects opinion information resultant with SentiWordNet help, and a proposed process for obtaining the features having the original reviews data sets as a starting point. From this specification, the process can be employed with help of third party tools GATE, and C#, Visual Studio 2010 for the generation of SentiWordNet based features.

4. CLASSIFICATION EXPERIMENT

The analysis presented in this chapter starts by looks at the common available classification methods with their highlighted key characteristics to decide how a classifier to be chosen and its setting for a particular task. The data set should study in terms of dimensionality, size and data characteristics and specially type of data it contains. Most methods or tools availableat this time consider only numerical data, and specifically categorical information missing need to be talked before training. We discussed in details what algorithm we choose and our experimental setup is arranged according to it.

4.1. Classifier Techniques Considerations

In the section of this chapter,mostcommon classification algorithms were gauged, with emphasis on their fundamentalenthusiasm, applicability and optimistic aspects. It can be seen from the algorithms inspected that each implements a specific heuristic to address the lack of information regarding the unknown real distribution of data. Therefore, each technique makes some assumptions on how the predicted classes can be separated: for example Naïve Bayes algorithm on the probabilistic independence of attribute occurrence.

Summary of the findings about the classification algorithms is presented in the table 14.

Algorithm	Positive Aspects	Negative Aspects
Naïve Bayes	• Model is easy to interpret and efficient computation	• Assumption of attributes being independent not necessarily valid
Maximum Entropy	• Simple to understand and implement • Easy interpretation	• Potentially slow as training data increases
Support Vector Machines	• Very good performance on experimental results • Low dependency on data set dimensionality	• Categorical or missing values need to be pre-processed • Difficult to interpret the resulting model

Table 14: Survey of Classification Methods

As noted earlier in this chapter, aims at presenting popular classification methods and their application for the data-driven prediction, along representationfor the element of current common classifier techniques, with many more constantly being developed. Other methods based on the same principle present on SVM, of finding a separating hyper plane can be understood in the “*linear discriminant”*class of methods. A variety of methods applied to the classification of textual data is surveyed in [44].

4.2.1. Choosing a Classifier

Each algorithm must be look for it runtime characteristics, as performance of the wholeproject will depends on the time cost of algorithm. Heavy training needs also increase the time cost and limit themining task usefulness. In some cases, the high dimensionality also limits the performance of the application with particular method. The explanatory capabilities of the method be observed, and may be a key factor in the choice of data mining algorithm, depending on the expected results by the end usersfor data mining workout.

To conclude, it would be crucial to minimizethe classification error when choosing a classifier along betterruntime performance. Defining which classifier will make best , how manyfactors it will be dependent on related to the availability of data, such as class label distribution in the whole population - in principle an unknown fact - and how closely that distribution is characterized in the data attributes available for training.

The size of data used for training is also significant, since larger training sets inclinethe performance of a classifier to the best obtainable performance. In addition, induction bias has one part in picking an algorithm or classifier since a specific heuristic existent in one technique may better distinguish classes than other methods for a specific problem.

It is observedthat the available data is the single most important source of information and intuition on determining a classifier, and the bettermethodology in the literature for selecting the best performing method for a problem in hand.

4.2.2. Why SVM

- Mostly in literature of text classification SVM is used as the front line algorithm.

- Specially, extensive experiments performed by [37] to compare performance of Maximum Entropy, Naïve Bayes and Support Vector Machine (SVM) on movies reviews data sets.
- SVM consistently outperforms Maximum Entropy and Naïve Bayesian.

4.4. Experiment Setup

4.4.1. Text Representation

Before implementing machine learning techniques on text data, a systematic and structured form of text or document representation needed which should capture as much as possible document statistical information.

Bag Of Features:

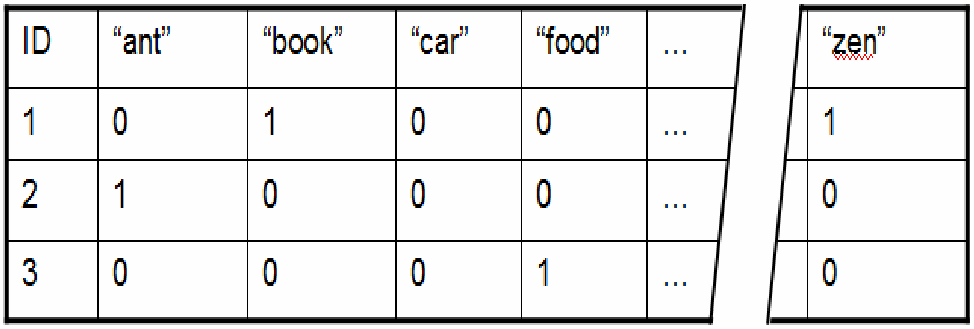
Combination of both Bag of words and phrases are used for the document vector. The number of distinct features corresponds to number of columns or length of record in a single in a word vector in the document collection. If document size are comparatively larger or and also with richer documents than this number can increase quite rapidly, and it is common to get very high dimensional spaces for the word vector with attributes count is in thousands.

Bernoulli document model:

In [37] both terms frequency and term presence are used for feature weighting but binary term presence shows better result. When Term Presence is considered for feature value and each document/review is converted into vector for the present features such as

$$Feature_i = \begin{cases} \quad 1 \;\; if \; (present \; in \; document) \\ \qquad\qquad else \\ null \end{cases}$$

When a word is present in a given document, a non-zero value is present representing term presence. This can be, for instance, a binary value indicating a term has occurred in a given document. A partial word vector data set is represented in the figure 9.



ID	“ant”	“book”	“car”	“food”	...	“zen”
1	0	1	0	0	...	1
2	1	0	0	0	...	0
3	0	0	0	1	...	0

Figure 9: Example Word Vector

Class and Attribute Label:

To assign each feature a numeric index value, than the reviews datasets are in the format such like:

```
<class_label><Feature1> :< value1><Feature2> :< value2>...
.
.
.
```

As there are only two labeled review classes, so numeric +1 and -1 as class label i.e.

class_label = +1 (for review labeled as positive)

class_label = -1 (for review labeled as negative)

4.4.2. LIBSVM

LIBSVM is integrated software introduced by [8] for C-SVC and nu-SVCsupport vector classification, epsilon-SVR and nu-SVRregression, and one-class SVM of distribution estimation. It supports the multi-class classification. LibSVM library publically available at:

- http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

In this exercise we use C-SVC for multiclass classification and select the linear kernel available in LIBSVM. As in text classification there is quite large dimension of feature space and also such classification problems are linearly separable.

4.4.3. Validation & Evaluation

We consider cross-validation because there are limited amount of data for training and testing, such that consider swapping the roles of training data and testing data. We use 10-Fold Cross

Validation in our experiments such that split the data into 10 equal partitions and repeat the process for 10 times such that in each run 10% of data i.e. one partition used for testing and remaining 90 % i.e. nine partitions used for training and ensure that each partition is used for testing at once. Accuracy is taken as an evaluation measure for classification in the case of balance dataset as in our experiments.

For 10 fold cross validation in LibSVM parameter (–v = 10) is selected.

4.5. Conclusion

This chapter discussed aspects of classification algorithms to be chosen into account when picking a classifier for a learning task, what data characteristics effect algorithm performance and its explanatory capabilities.

This chapter also provides a detailed description our experiment using LIBSVM. This experiment is conducted specially to evaluate the use of SentiWordNet for document-level sentiment classificationas a tool. The document representation and feature weighting is selected to perform the traditional classification setup. There is limited scope for classifier algorithms fine tuning for the parameters used in it,becauseby experiment to test multiple algorithm and with multiple setting for achieving the best performance is out of scope of this research.

5. RESULTS, COMPARISON AND FUTURE WORK

This chapter concludes this dissertation's research. The obtained results are presented followed by the detail discussion and comparison with previous literature, research objectives of our thesis and achievements are highlighted.

Experiment results by using SentiWordNet for sentiment classification are reviewed, with concluding remarks. Opportunities for future research work are presented at the end of this chapter.

5.1. Results

Table 15 shows the results for all dataset and the bold face represents the best result.

Measure	Features Set	Books	Dvd	Electronics	Kitchen	Movies
AVERAGE ACCURACY	Senti	69.25	72.25	72.10	74.40	73.20
	SentiUni	74.95	75.15	77.80	78.45	84.40
	SentiPhr	73.35	75.75	77.05	77.80	82.50
	SentiUniPhr	76.45	76.65	78.70	79.35	85.25
	SelSenti	71.85	74.65	73.10	75.65	81.60
	SelSentiUni	83.50	82.95	83.30	85.15	86.95
	SelSentiPhr	82.00	84.15	85.55	84.75	86.90
	SelSentiUniPhr	**85.20**	**85.40**	**86.05**	**86.80**	**89.30**

Table 15: Accuracy results of different feature set

5.2. Comparison with Previous Work

We compared our approach results as shown in table 16 with previous studies with the reference to the approach followed in this paper. The objective of this research was to find the most representative features for the sentiment classification. Mainly we compare our results to those machine learning based approaches which were built on the lexicon resource i.e. SentiWordNet.

There is lot of semantic orientation approaches available those have used SentiWordNet. Other significant addition of this research is new feature dimension syntactic phrases with SentiWordNet for the sentiment classification.

Approach	books	Dvd	electronics	kitchen	Movies
Phrases and PMI [48]	_	_	_	_	65.83
Unigrams with presence[37]	_	_	_	_	82.90
SentiWordNet Subjective and Objective (after revised score) Words [27]	_	_	_	_	78.50
Unigrams and Bi-Tagged Phrases [1]	_	_	_	_	89.40 F-measure
SentiWordNet sentiment, Content-specific unigrams & bigrams and Content-free features [15]	78.85	80.75	83.75	84.15	_
Ensemble of unigrams, bigrams & dependencies [53]	78.35	81.00	83.35	86.75	87.25
Our Approach(sentiment, Content-specific unigrams , NP & VP)	**85.20**	**85.40**	**86.05**	**86.80**	**89.30**

Table 15: Results Comparison

Initially we compare our work with the most renounced and foundation work done by [48] in sentiment analysis. He followed the unsupervised approach to determine the semantic orientation of text based on the PMI value of the phrases. The relevance of his work with our approach is that it’s the first bench mark in sentiment analysis which is based on the phrases. He extracted two word phrases based on the POS patterns with adjective and adverbs. Phrases PMI value was determined with the phrase co-occurrence of “Excellent” and “Poor”. He done his experiment on four domains with limited no of reviews he best achieved his results on automobiles i.e. 84%. But as comparison with our work he obtained worse results on movies domain i.e. 65.83%.

When we compare our work with [37] different types of features were experimented including individual unigram, individual bigrams, combination of unigrams and bigrams, adjectives, unigrams with POS and Position with term frequency and term presence. They best achieved accuracy 82.9% on Movie reviews using unigrams and their feature presence as weight. One thing notable is here they proved that only top adjective shows less results with the unigrams of all POS. Second the tried combination of unigrams and bigrams in combination which not shows the best result due to huge features dimension any bigrams. Finally we want to mentioned that they not followed any specific feature selection method i.e. may be their results was better than the mentioned. We establish in this research different combination taking the sentiment features as base line and our results are much better than their results on the same domain of movies reviews.

Objective words are used in [27]with revised score along with subjective word of SENTIWORDNET to enhance classification performance. This work reevaluates objective sentiment words in the SentiWordNet because more than 90 % of the words are objective. Reassign a proper sentiment value and efficiently integrate them with subjective word to use for sentiment classification.Sentence semantic orientation is calculated based on the no of positive and negative word it contains. Huge number i.e. 25,886 of movies reviews utilized to reassign the orientation to the objectives words. According to their results, they achieved average accuracy 71.89% with original SentiWordNet and accuracy 76.02% with revised SentiWordNet. The revised SentiWordNet outperforms the original SentiWordNet evaluated by 4.1%. We included the content specific features along SENTIWORDNET features to complement sentiment classification. Our procedure to get the content specific features is less computationally expensive. They achieved best accuracy 78.5% on Movie Reviews and however our accuracy on Movies reviews is 89.30%.

Another motivation for the phrasal features with unigrams to address the sentiment classification got from [1]. They enhance the [48] work and get the bi-tagged phrases based on POS based nine patterns. They also utilize these phrases in supervised classification algorithms as features. It is noted that individual bi-tagged phrases features results are even less than the bigrams and combination results with unigrams are better. Information gain is used to reduce the dimensionality and noise and prominent features shows the best results. One thing notable that they publish their results with f-measure with 10 fold cross validation on movies reviews. Our experiment used the most easy phrase extraction preprocess activities and extracting bi-tagged phrases also we utilized these phrases with sentiment features. Our results are still comparable with their 89.4% f-measure i.e. we achieved 89.3% accuracy.

Ensemble of features (POS based & Word relation) and ensemble of classifiers is studied in [53]. They proposed to improve classification performance by the ensemble of features and not only just combining them. Classification performance of machine leaning algorithm often varies from domain to domain; hence combine their outputs for a better integrated output. They use three classifier in their experiments includes NB, MaxEnt and SVM (as base-classifier with linear kernel).Three ensemble methods used for features sets and classification algorithms and for both named as Fixed Combination, Weighted Combination and Meta-classifier combination. They not mentioned about size of features and features selection strategy for their experiment. Results that are comparable with our approach are Joint feature and ensembles of features of POS and dependency word relation. We also learned from their features because it highlights lexical POS aspects and word relation aspects with unigrams, bigrams and dependency trees. They achieved better with the ensembles word relational features than ensemble of POS-based feature. They experiments on the same 5-dataset of reviews that we used in this research. Our results much better than their approach even we exclude much computation of ensemble or integration of different results.

The most comparable and motivational study to this research was carried by [15]. They worked on three types of features i.e. F1-content free (lexical, syntactical and structural), F2-content specific (unigrams and bigrams) and F3-sentiment features (subjective terms from SentiWordNet). Sentiment features were new dimension introduced by them and experimented to validate that addition of these sentiment features could enhanced the classification performance. of considered by and they exclude nouns in SENTIWORDNET features. But one thing noted here that F1+F2 still better than F1+F3 and selected F1+F2 and selected F1+F2+F3 shows little difference in results because F2 diminish the effect of F3. There is a need to re-think on content-specific features to get a good match. We exclude the content free features as looks neutral from their results of selected F1+F2 and selected F1+F2+F3. We proposed the new dimension of content specific features with noun and verb phrases and considered the sentiment features as base line features. Also use unigram with all POS for content specific features with phrasal features. F-Score for feature selection is the new approach followed by us as compared to IG. On the multi domain product reviews dataset they achieved best accuracy with Selected f1+f2+f3: 78.85~84.15 which is less against our feature set “SelSentiUniPhr” on the same dataset 85.20~86.80%. Especially we mentioned here that our results for all the domains are consistent.

5.3. Conclusion

It is cleared from our results sentiment classification performance enhanced with the inclusion of content specific features with sentiment lexicon features. If individual sentiment features were used, text having zero or less quantity of sentiment features shows ambiguous sentiment orientation. Previously different individual features and joining scheme of the features are researched but our features sets proved as best representative features. More significant performance improvement noticed after feature selection.

Individual sentiment features are smaller in size and easy to establish as compare to content specific features but it added valuable representation to text. Also performance is more concerned for sentiment classification therefore our combination of sentiment features and content specific features perform much better than all other feature set.

Our approach provides classification accuracy of 85.2~89.3% by

1. Including new feature dimension of content specific syntactic Noun and Verb phrases with sentiment features extracted from SentiWordNet subjective terms.
2. Provide a more compact vocabulary by considering morphological roots and accurate unigrams and with only four POS groups.
3. F-Score is a relative new and simple method of feature selection to find most discriminative features

5.4. Future Work

Addition of more pre-processing capabilities such as dependency parsing for word relations features could be used with sentiment features. TF-IDF for feature weight and other features selection options could be explored.

Neural Networks and other latest classifier may be adapted and ensemble of these classifiers can be experimented with our features baseline. We are also interested to validate our framework on other than English language dataset and multilingual dataset.

This framework is general in nature so can easily be adapted to other domain dataset e.g. blogs, emails and tweets etc. as well as on other than product reviews datasets. We are also interested in future to validate our framework on other than English language dataset and multilingual dataset based on the availability of respective lexicon resource.

APPENDIX A – *Stop Word List*

"	and	Does	he's	Myself	t's	until	who's
#	anybody	doesn't	Her	No one	Tends	unto	whoever
$	anyone	Doing	here's	Nor	Than	upon	whom
'	anything	don't	Hers	Of	That	uses	whose
's	aren't	Don't	Herself	Ones	that's	via	with
(	associated	During	Him	Oneself	that's	vs.	without
)	became	Eddo	Himself	Onto	the	was	wo
*	because	Else	His	Others	the	wasn't	won't
+	becomes	et	How	Ought	their	we	would
,	been	etc	Howbeit	Our	theirs	we'd	wouldn't
.	beside	everybody	If	Ours	them	we'll	you
/	c'mon	everyone	Inasmuch	Ourselves	themselves	we're	you'd
:	c's	everything	Indicated	Per	there's	we've	you'll
;	came	followed	Indicates	Pp	there's	went	you're
?	can't	follows	Into	Provides	thereupon	were	you've
I'd	cannot	for	Is	Qv	these	weren't	your
I'll	causes	from	isn't	Regarding	they	what	yours
I'm	changes	gets	Isn't	Regards	they'd	what's	yourself
I'vet	com	gives	it'd	Says	they'll	when	yourselves
Inc.	comes	goes	it'll	Seemed	they're	whenever	
[	concerning	got	it's	Seems	they've	where	
]	considering	gotten	Its	Seen	this	whereafter	
_	containing	greetings	Itself	Selves	those	where's	
`	contains	had	Keeps	Shall	thru	whereas	
against	could	hadn't	Knows	She	to	whereby	
allows	couldn't	happens	Lest	Should	took	wherein	
although	detail	has	let's	shouldn't	toward	whereupon	
among	did	hasn't	Looks	Since	towards	whether	
amongst	didn't	haven't	Ltd	Something	tries	which	

Table 16: List of Stop Words

REFERENCES

[1] Agarwal, Basant, Namita Mittal, and Erik Cambria. "Enhancing Sentiment Classification Performance Using Bi-Tagged Phrases." Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on. IEEE, 2013.

[2] Alpoydin E. “Introduction to Machine Learning”, MIT Press, 2004.

[3] Blitzer, J., M. Dredze, and F. Pereira. "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification." Proceedings of the Association for Computational Linguistics. ACL Press, 2007. 440-447.

[4] Benamara, Farah, et al. "Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone." ICWSM. 2007.

[5] Boser B, Guyon I, Vapnik V. “A Training Algorithm for Optimal Margin Classifiers”, Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pp. 144-152, March, 1992.

[6] Brants T. "T&T-a statistical part-of-speech tagger", Proceedings of the sixth conference on Applied natural language processing, 224-231, 2000.

[7] Brill, Eric. "A simple rule-based part of speech tagger." Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992.

[8] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST) 2.3, 2011.

[9] Chen, Yi-Wei, and Chih-Jen Lin. "Combining SVMs with various feature selection strategies." Feature extraction. Springer Berlin Heidelberg, 2006. 315-324.

[10] Cui H, Mittal V, Datar M. "Comparative Experiments on Sentiment Classification for Online Product Reviews." Proceedings of the National Conference on Artificial Intelligence. AAAI Press, 2006. 1265-1270.

[11] Cunningham, Hamish. "GATE, a General Architecture for Text Engineering." Computers and the Humanities, 2002: Volume 36, Issue 2, pp 223-254.

[12] Conrad J, Al-Kofahi K. “Effective document clustering for large heterogeneous law firm collections”, Proceedings of the 10th international conference on Artificial intelligence and law, pp. 177-187, 2005.

[13] Corney M. de Vel, O., Anderson A, Mohay G. "Gender-preferential text mining of e-mail discourse", 18th Annual Computer Security Applications Conference, 2002.

[14] Dorre J, Gerstl P, Seiffert R. “Text mining: finding nuggets in mountains of textual data”, Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999.

[15] Dang, Yan, Yulei Zhang, and Hsinchun Chen. "A lexicon-enhanced method for sentiment classification: An experiment on online product reviews." Intelligent Systems, IEEE 25.4, 2010: 46-53.

[16] Esuli, Baccianella, Stefano, Andrea, Sebastiani, and Fabrizio. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." LREC. Vol. 10, 2010.

[17] Fayyad, U. Piatetsky-Shapiro, G. Smyth, P, "From Data Mining to Knowledge Discovery in Databases". AI Magazine, Vol. 17; No. 3, (1996) pages 37-54.

[18] Feldman R, Dagan, I. “Knowledge discovery in textual databases (KDT)”, Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), pp. 112-117, 1995.

[19] Feldman R, Fresko M, Hirsh H, Aumann Y, Liphstat O, Schler Y, Rajman M. (1998) “Knowledge Management: A Text Mining Approach”, Proceedings of the 2nd international conference on Practical Aspects of Knowledge Management – PAKM’98, Basel, Switzerland, Oct. 1998.

[20] Forman, George. "An extensive empirical study of feature selection metrics for text classification." The Journal of machine learning research 3 (2003): 1289-1305.

[21] Forman G, Kirshenbaum E, Suermondt J, “Pragmatic Text Mining: Minimizing Human Effort to Quantify Many Issues in Call Logs”, Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’06-Philadelphia), pp. 852-861, August 2006.

[22] Gatti, Lorenzo, and Marco Guerini. "Assessing sentiment strength in words prior polarities." ." arXiv preprint arXiv:1212.4315. 2012.

[23] Ghani R, Probst K, Liu Y, Krema M, Fano, A. “Text mining for product attribute extraction”, ACM SIGKDD Explorations Newsletter, Vol. 8, No. 1, pp. 41-48, 2006.

[24] Hand D, Mannila H, Smyth P, 2001, “Principles of Data Mining”, The MIT Press, Cambridge Massachusetts, 2001.

[25] Hatzivassiloglou, Vasileios, and Janyce M. Wiebe. "Effects of adjective orientation and gradability on sentence subjectivity." Proceedings of the 18th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2000.

[26] Horrigan J. “Online Shopping”, Pew Internet and American Life Project – Research Report, February, 2008.

[27] Hung, Chihli, and H. Lin. "Using objective words in SentiWordNet to improve sentiment classification for word of mouth." Published by the IEEE Computer Society, 2013: 1-1.

[28] Joachims, T. “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”. Proceedings of the European Conference on Machine Learning (ECML), Springer, 1998.

[29] Kao, A., Quach, L., Poteet, S., Woods, S. "User assisted text classification and knowledge management", Proceedings of the twelfth international conference on Information and knowledge management, 524-527, 2003.

[30] Kolcz J, Alspector E. "SVM-based filtering of e-mail spam with contentspecific misclassification costs", Proceedings of the Workshop on Text Mining (TextDM’2001), 2001.

[31] Kroeze, J. H., Matthee, M. C., and Bothma, T. “Differentiating data- and text-mining terminology”. Proceedings of the 2003 Annual Research Conference of the South African institute of Computer Scientists and information Technologists on Enablement Through Technology, September 17 - 19, 2003.

[32] Marwick A D. “Knowledge Management Technology”, IBM Systems Journal, Vol 40, No 4, pp. 814-830, 2001.

[33] McKnight. "Text Data Mining in Business Intelligence." Information Management Magazine, January 1st, 2005.

[34] Meyer, T.A. and Whateley, B. "SpamBayes: Effective open-source, Bayesian based, email classification system", Proceedings of the First Conference on Email and Anti-Spam (CEAS), 2004.

[35] Nasukawa T, Nagano T. “Text Analysis and Knowledge Mining System”, IBM Systems Journal, Vol. 40, No. 4, 2001.

[36] Noam Chomsky, Syntactic Structures is a book in linguistics by American linguist first published in 1957.

[37] Pang, Bo, Lillian Lee, and ShivakumarVaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computer Linguistics, 2002.

[38] Pang B, Lee L. "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", Proceedings of the ACL, 2004."

[39] Pang B, Lee L. “Opinion Mining and Sentiment Analysis”, Foundations and Trends in Information Retrieval, Vol. 2, Nos. 1-2, pp. 1-135, 2008.

[40] "Provost, J. ""Naive-Bayes vs. Rule-Learning in Classification of Email"", Technical Report 99 - University of Texas, 1999."

[41] Rogati M, Tang Y. “High-performing feature selection for text classification”, Proceedings of the ACM 11th international conference on Information and knowledge management, pp. 659-661, 2002.

[42] Salvetti F, Lewis S, Reichenbach C. "Automatic Opinion Polarity Classification of Movie Reviews." Colorado Research in Linguistics (Boulder: University of Colorado) 17, no. 1 (June 2004).

[43] Scott, Sam, and Stan Matwin. "Feature engineering for text classification." ICML. Vol. 99. 1999.

[44] Sebastiani F. “Machine Learnining in Automated Text Categorization”, ACM Computing Surveys, Vol. 34, pp. 1-47, 2002.

[45] Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management 24.5 (1988): 513-523.

[46] Stavrianou, A., Andritsos, P., and Nicoloyannis, N. “Overview and semantic issues of text mining”. SIGMOD Record, Vol. 36, Sep. 2007, 23-34, 2007.

[47] Toutanova K, Manning C. “Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger”. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[48] Turney P. “Thumbs up or Thumbs down? Sentiment Orientation Applied to Unsupervised Classification of Reviews”, Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics – ACL, 2002.

[49] Wang H, Wang S. "A Knowledge Management Approach to Data Mining for Business Intelligence." Industrial Management and Data Systems 108, no. 5 (2008): 622-634.

[50] Wiebe J, Bruce R, Martin M, Wilson T, Bell M. "Learning Subjective Language." Computational Linguistics 30, no. 3 (2004): 277-308.

[51] Weiss S, Indurkhya N. “Predictive Data Mining – A Practical Guide”, Morgan-Kauffman Publishers, San Francisco, California, 1998.

[52] Weiss S, Indurkhya N, Zhang T, Damerau F, “Text Mining – Predictive Methods for Analyzing Unstructured Information”. Springer, 2005.

[53] Xia, Rui, ChengqingZong, and Shoushan Li. "Ensemble of feature sets and classification algorithms for sentiment classification." Information Sciences 181.6, 2011: 1138-1152.

[54] Yu H, Hatzivassiloglou V. “Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying Polarity in Sentences”, Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pp. 129- 136, 2003."