Text Mining at Feature Level: A Review

Tanya Shruti[1], Manish Choudhary[2]

[1]M.tech Scholar, Department of CSE, YIT College, Jaipur, Rajasthan, India
[2]Assistant Professor, Department of CSE, YIT College, Jaipur, Rajasthan, India

***Abstract*—*Text Mining is the technique that helps users to find out useful information from a large amount of text documents on the web or database. Most popular text mining and classification methods have adopted term-based approaches. The term based approaches and the pattern-based method describing user preferences. This review paper analyse how the text mining work on the three level i.e sentence level, document level and feature level. In this paper we review the related work which is previously done. This paper also demonstrated that what are the problems arise while doing text mining done at the feature level. This paper presents the technique to text mining for the compound sentences.*

Keywords*—*Text Mining, Sentiment Analysis, Sentiment level, Compound Sentences, Feature Analysis.

I. INTRODUCTION

Text Mining [7] is the technique, by which automatically extracting information from different written resources. Text mining is different from web search. In search, the user is typically looking for something that is already known and has been written by someone else. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down. Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge. The problem of Knowledge Discovery from Text (KDT) [1] is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasingly significant role in emerging applications, such as Text Understanding. Text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. As a result, text mining is a much better solution for companies. To date, however, most research and development efforts have centered on data mining efforts using structured data. The problem introduced by text mining is obvious: natural language was developed for humans to communicate with one another and to record information and computers are a long way from comprehending natural language. Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle such as slang, spelling variations and contextual meaning. However, although our language capabilities allow us to comprehend unstructured data, we lack the computer's ability to process text in large volumes or at high speeds.

II. METHODS AND MODELS USED IN TEXT MINING[11]

Text mining methods is based on how text document are analyzed. In these methods of text mining text document analyzed on the basis of term, phrase, concept and pattern. Based on the information retrieval there are four methods, 1) Term Based Method (TBM). 2) Phrase Based Method (PBM). 3) Concept Based Method (CBM). 4) Pattern Taxonomy Method (PTM).

A. Term Based Method

Term in document is used to determine content of text. In Term Based Method each term in document is associated with value known as weight, which measure importance of term i.e. terms contribution in document. Word having semantic meaning is known as term and collection of such terms contributes meaning to document. Term based methods suffer from the problems of polysemy and synonymy. Polysemy means a word has multiple meanings and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want. Information retrieval provided many term-based methods like supervised and traditional term weighting methods to solve this challenge.

B. Phrase Based Method

Phrases are less ambiguous and more discriminative than individual term so in phrase based method document is analyzed on phrase basis. In process of analysis of document phrases are profile descriptor of document. Phrases are collection of semantic terms so carries more information than single term. Over many years this is hypothesis that phrase based approach performs better than term based approach, as phrase may carry more semantic than term. Using data mining algorithms it is definite to obtain various phrases but it is difficult to use these phrases effectively to answer what user want. It is difficult because phrases have fewer occurrences in document and phrases comprise large number of noisy with redundant terms. As phrases are collection of terms those can be considered as sequence of terms and hence to find sequence of terms sequential pattern mining algorithm is used. Algorithm extracts frequent sequential patterns, here pattern used as words or phrase which is extracted from document.

C. Concept Based Method

Most of text mining techniques are based on word and/or phrase analysis of text. It is important to find term that contributes more semantic meaning to document this concept is known as concept based method. Only the importance of term within document is captured in statistical analysis of term based method. In concept based method the term which contributes to sentence semantic is analysed with respect to its importance at sentence and document levels. The model tries to analyze term at sentence and document level by efficiently finding significant matching term rather than single term analysis.

D. Pattern Based Model

In pattern based model document is analysed on pattern basis i.e. pattern of document is formed by analyzing is-a-relation between terms to form taxonomy. Taxonomy is tree like structure The pattern based approach can improve the accuracy of system for evaluating term weights because discovered patterns are more specific than whole documents. To generate PTM document split into paragraphs. In pattern taxonomy the nodes represent frequent patterns and their covering sets. The edges are “is-a” relation. Smaller pattern in taxonomy are usually more general because they could be used in both positive and negative documents. Larger patterns in taxonomy are usually more specific since they may be used in positive documents. The semantic information will be used in the pattern taxonomy to improve the performance of using closed patterns in text mining.

III. RELATED WORK [2, 3, 4, 5, 6, 8, 10, 12]

Pang et al. [2002], presented a work based on classic classification techniques. It aims to identify that machine learning algorithms can produce good result or not when opinion mining is computed at document level. He presented the results using nave bayes maximum entropy and support vector machine algorithms and shown the good results as comparable to other ranging from 71 to 85% depending on the method and test data sets. When he used movie reviews as a data set the all three method did not perform well. **Turney [2002],** presented a work based on distance measure of adjectives found in whole document with known polarity i.e. excellent or poor. The author presents a three step simple unsupervised algorithm for classifying reviews as recommended (thumbs up) or not recommended (thumbs down). In the first step; the adjectives are extracted Second step, the semantic orientation is captured by measuring the distance from words of known polarity .Third step, and the algorithm counts the average semantic orientation for all word pairs and classifies the review. It appears that movie reviews are difficult to classify. **Riloff and Wiebe [2003],** proposed a method called bootstrap approach to identify the subjective sentences and achieve the result around 90% accuracy during their tests. It used high precision classifier unannotated data to automatically create large training set. It used extraction pattern learning algorithm to identify more objective sentences. Author goal is to classify individual sentences as subjective or objective at the document level. The extraction patterns perform well and achieve better precision range. **Yu and Hatzivassiloglou [2003]**, separated opinions from facts at document and sentence level. They proposed a Bayesian classifier which was used to classify documents as subjective (editorials) vs objective (news articles). They also proposed three unsupervised statistical techniques for detecting opinions at sentence level. They performed three class classification, positive vs negative vs neutral, and compared their system performance with human evaluation over 400 sentences and achieve 97% accuracy at the document level and 91% accuracy at sentence level.**Wilson et al**.**[2004],** It presented the first experimental results classifying the strength of opinions and other nested clauses using boosting, rule learning, and support vector regression. It pointed out that not only a single sentence may contain multiple opinions, but they also have both subjective and factual clauses .It is also important to identify the strength of opinions. **K Denecke [2008],** performs opinion mining at document level of movie domain. The author used SentiWordNet and follows average scoring method. The scores of individual words in documents are aggregated to compute final score. For calculating score of word, the score of all synsets is calculated and averaged to generate final score through rule. The technique works well at document level. For movie domain feature based opinion mining will be more appropriate as users could be interested in any specific aspects of movie based on his choice. **S. Agrawal [2012] ,** presents the summarization on the basis of features of movies. The sentences which contain the specific feature are computed through technique to express opinion in the form of ratings. The authors proposed the method which generates ratings on the basis of individual features. The technique could not work well in case of compound sentences in which there is opinion on different features is described about product or services. Hence, in such cases, segmentation of sentence into clauses or simple sentences based on feature is required to better results. It also uses prior polarity lexicon to start with contextual polarity identification. **Yuefeng Li et.al [2015],** presents an innovative model for relevance feature discovery. It discovers both positive and negative patterns in text documents as higher level features and deploys them over low-level features (terms). It also classifies terms into categories and updates term weights based on their feature and their distributions in patterns. Substantial experiments using this model on RCV1, TREC topics and Reuters-21578 show that the proposed model significantly outperforms both the state-of-the-art term-based methods and the pattern based methods.

IV. LEVEL OF SENTIMENT ANALYSIS

Sentiment analysis or opinion mining is the computational observation of user's opinions, appraisals, and emotions toward entities, events and their attributes. Opinions are important because whenever we want to make a decision about any product or services we have need to know others opinion about that product or services. Sentiment analysis depends on opinoted text which is commented by user.

Textual information may be broadly classified into two main types –

Facts: Facts are objective based expression about entities, events and their properties.

Opinion: Opinions are usually subjective based expression that determines people's sentiment or feelings.

Sentiment analysis are mainly divided into document level, sentence level and feature level/attribute level/aspect level / phrase level to find whether the given text is providing positive opinion ,negative opinion or neutral .This is also known as 'sentiment polarity prediction'. Hence sentiment analysis is carried out into three levels [2] [3],

I. Document level
II. Sentence level
III. Feature level

1.1 Document level

It is classifying the opinionated text given by the user in whole document as positive, negative or neutral about a certain subject or object. Hence subjective or objective classification is necessary in document level classification .The problem arise in this classification when the informative text is to extract for deducing sentiment of the entire document. In document level classification each document focuses on single objects and contains opinion from a single opinion holder.

1.2 Sentence level

This type of classification refer to calculate the polarity of each sentence as shown in fig. 2.1.The sentence level classification mainly focused on two things [4].First one is ,to identify that the opinionated sentence is objective or subjective .The second one is ,to identify the opinionated sentence is positive ,negative or neutral. The assumption is taken at sentence level is that a sentence contain only one opinion for e.g.,

"The picture quality of this phone is good."

However, it is not true in many cases like if we consider compound sentence for e.g.

"The picture quality of this phone is amazing and superb battery life, but the screen is too small".

It expresses both positive and negative opinions and we say it is a mixed opinion. For "picture quality" and "battery life", the sentence is positive, but for "screen", it is negative. It is also positive for the camera as a whole.

1.3 Feature level sentence classification

The feature level sentiment classification is a more pinpointed method to opinion mining .This type of classification mainly focused on feature of particular product or services .It give the opinion based on the feature of the object .Analysis of the object based on their feature called as feature based sentiment analysis .It extract the feature of the object and conclude the opinion in the form of positive ,Negative or neutral, then group the feature synonyms and produce the summarization report [8]. Liu used supervised pattern learning method to extract the features of the object for identification of opinion orientation. To identify the orientation of opinion author used lexicon based approach. This approach basically uses opinion words and phrases in a sentence to identify the opinion. The working of lexicon based approach is described in following steps.

- Identification of opinion words
- Role of Negation words
- But-clauses

V. COMPOUND SENTENCES

The following methodology we use to determine the opinion in compound sentence

2.1 Sentence classification

In the sentence classification we go to individual compound sentences to determine whether a sentence is subjective or objective.

2.2 Segmentation of the document into sentences

By the help of sentence delimiter the document is segmented into individual sentences. We have to use rule based pattern matching to identify sentence boundary.

2.3 Determining the opinionated sentence

We will use boot strap approach proposed by Riloff and Wiebe [5] for the task of subjective sentences identification. It will use a high precision and low recall classifiers to extract a number of subjective sentences collected from various movie review sites.

2.4 Semantic Orientation

There are various tools for text mining like Stanford CoreNLP, Weka, and Rapid Miner etc. SentiWordNet tools can use for determine semantic strength for text mining. It determines the strength of text in the form of positive, negative or neutral. For Example:-

“This movie is good”- Positive

“Actor was not good”- Negative

“The movie is good but songs is not good”- Neutral or Mixed.

2.5 Feature Extraction from Text

From the opinioted text we have to extract the feature. In previous text it is about movie and other text is about actor of the movie so we can see that first is positive opinion and other text is negative opinion. Here movie, actor, music, songs, story etc. can be termed as a feature of the movie. For mobile phone camera, picture, look etc cost, etc may be feature of the mobile phone. The lexicon based approach and pattern based approach can be used to feature extraction from the text.

VI. RESULT

We implemented this method using Stanford CoreNLP tool and SentiWordNet tool using java Programming languages. We use Movie review as a dataset. We select movie from dataset which contain 23 sentences and 200 words as a text. It generates the opinion based on the feature of the text. The accuracy is varies because it depends on sentence sentiment whether it is positive or negative and sentence structure.

VII. CONCLUSION

We conclude that Text Mining is difficult for compound sentences. The users can use any words or sentences which is difficult to identify. Text mining at the feature level is not an easy task. Many reviews site where the users post their comment about any product or services or movies based on that comments to identify whether it is positive or negative it is also a challenging task to handle.

REFERENCES

[1] Berry Michael W., (2004), “Automatic Discovery of Similar Words”, in “Survey of Text Mining: Clustering, Classification and Retrieval”, Springer Verlag, New York, LLC, 24-43.

[2] Haralampos Karanikas and Babis Theodoulidis Manchester, (2001), “Knowledge Discovery in Text and Text Mining Software”, Centre for Research in Information Management, UK.

[3] B. Pang, L. Lee, and S. Vaithyanathan, 2002. Thumbs up? Sentiment classification using machine learning techniques,” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.79–86

[4] P.Turney 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceeding of 40th annual meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417--424.

[5] E. Riloff, and J. Wiebe, 2003. Learning Extraction Patterns for Subjective Expressions, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Japan, and Sapporo.

[6] H.Yu, and V.Hatzivassiloglou, 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences, published in ACM digital library EMNLP.

[7] T. Wilson, J. Wiebe, and R. Hwa, 2004. Just how mad are you? Finding strong and weak opinion clauses. In: the Association for the Advancement of Artificial Intelligence, pp. 761--769.

[8] K. Denecke. 2008. “Using SentiWordNet for Multilingual Sentiment Analysis,” in Proceedings of the International Conference on Data Engineering (ICDE 2008), Workshop on Data Engineering for Blogs, Social Media, and Web 2.0, Cancun. IEEE

[9] Vishal Gupta and Gurpreet S. Lehal. 2009 “A Survey of Text Mining Techniques and Applications” in JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1.

[10] S.Agrawal and T.J.Siddiqui, 2012 “Feature based Star Rating of Reviews: A Knowledge-Based Approach for Document Sentiment Classification” in International Journal of Hybrid Information Technology Vol. 5.

[11] Sonali Vijay Gaikwad, Prof Archana Chaugule and Swapnil Kulkarni, 2014 “PERFORMANCE COMPARISON FOR TEXT MINING METHODS: REVIEW” in International Journal of Advanced Engineering Research and Studies E-ISSN2249–8974.

[12] Yuefeng Li, Abdulmohsen Algarni, Mubarak Albathan, Yan Shen, and Moch Arif Bijaksana, 2015 “Relevance Feature Discovery for Text Mining” in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 6.