SENTIMENT ANALYSIS AND OPINION MINING FOR TWITTER

[1] Nikita Dandwate, [2]Prof. Sarika Solanke

[1]MTech 2nd year student, [2]Assistant Professors,
[1,2] Department of Computer Science and Engineering,
[1,2]Deogiri Institute of Engineering and Management Studies,
Aurangabad (MH), India

Abstract: Sentiment analysis is defined as the category of natural language processing-based and computational technique. It is then used to sense, extract and illustrate information, which is subjective, and is expressed in a given part of text. The main intention of sentiment analysis is to classify the writer's attitude towards different topics into various categories like positive, negative or neutral. In past few years, on the other side, have witnessed the arrival of social networking websites, microblogs and Web applications and accordingly, an extraordinary growth in user-generated data is composed for sentiment mining. Data like Tweets, web-postings etc., all express thoughts on numerous topics and events, offer vast opportunities to study and analyse human feelings and sentiment. Twitter sentiment analysis however has emerged as a hot research topic in past few years. Most of existing solutions to Twitter sentiment analysis only consider textual information of Twitter messages, and probably fails when short or confusing messages or conversations appear. Current studies display that sentiment diffusion patterns on Twitter have very close relations with sentiment polarities of the Twitter messages. Therefore, in this paper we focus on how to fuse textual information of Twitter messages and sentiment Analysis patterns to obtain better performance on Twitter data.

IndexTerms **– Sentiment Analysis, Aspects, Opinion Mining, Social Network, Text Classification, Sentiment Polarities, Twitter Data.**

I. INTRODUCTION

Textual data in the world can be largely categorized into two types: facts and views. Opinions or views are so significant that whenever we need to decide, we hear other peoples' opinions. During decision making process, what others think has always been a vital piece of information. This is true for the individuals as well as for the organizations [9].

The Web has vividly changed the way that all express their views or opinions. All can now post opinions or reviews of products at many merchant sites and can put their views on almost anything. For a company, it may no longer be necessary to conduct surveys, organize focus groups, or employ external consultants in order to find consumer opinions about its products because the content which is generated on the Web can already give them such information. Social Media has become a central site where people express their opinions and views on various recent or trending topics. Emerging topics and news are instantly followed by many people thus providing opportunity to gauge the relation between expressed public sentiments and various events [10].

An opinion (which is expressed by means of opinion words) is a positive and/or negative sentiment, attitude, emotion or appraisal round an aspect. Optimistic and negative are called sentiment or opinion orientations. Daily huge amount of data is generated but not all of data is beneficial. So, by sentiment analysis we could reduce amount of data filter only useful data and generate diffusion models for same. To categorize the sentiment polarity of a Twitter message as positive, neutral and/or negative. The author also stated the main purpose is to classify a writer's attitude towards various topics into positive, negative and/or neutral categories [12]. Sentiment Analysis (SA) is an enduring field of research in the area of text mining. SA is the computational behavior of sentiments, opinions and subjectivity of text. Sentiment analysis at different levels like document-level or at the sentence level is valuable in many claims, but it does not provide the necessary detail needed for many use cases. In a document or sentence, the author may describe both positive and negative aspects of the product object, although the general sentiment on the object may be either positive or negative. To obtain these minutiae we need to shift to an additional fine-grained level of study and spread over the aspect level sentiment analysis. Aspect-level SA aims to categorize the sentiment with respect to the specific aspects of entities. Aspect-level sentiment analysis is further classified as: explicit and implicit. Consider example, in the sentence, ''The signal quality of this phone is amazing'', the aspect is ''signal quality'' of the object represented by ''this phone''. In this example the aspect is an explicit aspect as it appears in the sentence. In the sentence, ''This phone is too light'', the aspect is ''weight'' which is an implicit aspect as it does not appear in the sentence, but it is implied [9]. With the speedy growing user-generated content on the internet, instinctive sentiment analysis of online customer reviews has become a hot research topic recently, but due to variety and wide range of products and amenities being studied on the internet, the supervised and domain-specific copies are often not practical. As the number of reviews expands, it is essential to develop an efficient sentiment analysis model that is capable of extracting product aspects and defining the sentiments for these aspects. Also, the presentation of other traditional text sentiment analysis algorithms drops drastically when applied to predict sentiment polarities of Twitter messages [1].

II. LITERATURE SURVEY

1. Dataset Description

The Author collected the Twitter tweet and retweet data through their collaborations on research with Beijing Intelligent Starshine Information Technology Corporation, a leading big data collection and mining service provider in China. In their system, each tweet or retweet is assigned a sentiment label: +1 (positive), 0 (neutral) or -1 (negative). The Author, in his system ignored

retweets without added comment information in sentiment labelling and sentiment analysis and obtained labelled dataset which contains over 100; 000 tweets and retweets in total. After collecting enough Data, a Repost Cascade Tree (A directed, acyclic labelled graph) was created to capture the relationships between tweets and its re-tweets. Then a Repost Diffusion Network is created to describe how users interact with each other. Sentiment Reversal Technique was then used for identifying the sentiment polarities between the Tweet (Parent Tweet) and retweet (Child Tweet). In his work the Author tried combining ideas from sentiment diffusion and text sentiment analysis to analyse how sentiment diffusion information can help improve text sentiment analysis in social networks [1].

2. Joint Segmentation and Classification Framework

The researcher proposed a joint segmentation and
classification framework for sentence-level sentiment classification. It is broadly known that phrasal information is vital for sentiment classification**.** Author tried to overcome the drawback of existing sentiment classification Algorithms. He projected a joint framework for sentence-level sentiment cataloguing which simultaneously produces valuable segmentations and forecasts sentence-level polarity based on the segmentation results. Precisely, they established a candidate generation model to produce segmentation candidates of a sentence; a segmentation ranking model to score the helpfulness of a segmentation candidate for sentiment classification; and a classification model for predicting the sentiment polarity of a segmentation. They skilled the joint framework directly from sentences annotated with only sentiment polarity, without using any syntactic or sentiment annotations in segmentation level. Unlike present sentiment classification algorithms that build sentiment classifier based on the segmentation results from bag-of-words or separate segmentors, the proposed method simultaneously produces beneficial segmentations and predicts sentence-level polarity based on the segmentation results [2].

3. Aspect Level Sentiment Analysis

The researcher has done a Survey which focused on aspect-level sentiment analysis, where the goal was to find and aggregate sentiment on entities mentioned within documents or aspects of them. Aspect Level Sentiment Analysis was chosen Sentiment Analysis has gathered a lot of attention in past few years. The equivalent growth of the field has caused in the rise of several subareas, each addressing a different level of analysis or research question. An in-depth outline of the current state-of-the-art was offered, screening the tremendous growth that has already been made in finding both the target, which can be an entity as such, or some aspect of it, and the corresponding sentiment. Aspect-level sentiment analysis yields very fine-grained sentiment data which can be beneficial for applications in several domains. Existing solutions are categorized based on whether they offer a method for aspect detection, sentiment analysis, or both. Furthermore, a breakdown based on the sort of algorithm used is provided. To enable the quantitative evaluation of the numerous proposed methods, a call is made for the standardization of the evaluation methodology that comprises the usage of shared data sets. Semantically rich concept-centric aspect-level sentiment analysis is discussed and identified as one of the most promising future research directions. When considering the future of aspect-level sentiment analysis, they presented a move from traditional word-based approaches, towards semantically rich concept-centric aspect-level sentiment analysis [3].

4. Emoticon Smoothed Language Models

The researcher made a depth research in field of Twitter Sentiment Analysis which has a goal to discover the attitude or Opinion of the Tweets, which is typically formulated as a machine learning based text classification problem. The Author presented the drawback of using the fully Supervised models and how we can overcome them with the model presented by Author. For fully supervised manually labelled data is used as training data set while some models use noisy labels such as hashtags or emoticon s. Thus, there is a very limited data available as a training dataset. Also, it is a very time-consuming and labor-intensive methods. Henceforth, the finest strategy is to utilize both manually labelled data and noisy labelled data for training. However, a challenge of how to seamlessly integrate these two different kinds of data into the same learning framework is still exists. Author presented a novel model, called emoticon smoothed language model (ESLAM), to handle this challenge. The elementary idea is to train a language model grounded on the manually labelled data, and then usage the noisy emoticon data for smoothing. Experiments on real data sets established that ESLAM can effectively integrate both kinds of data to outperform those approaches using only one of them. Tests on real data sets show that author's ESLAM method can competently integrate both kinds of data to outperform those methods using only one of them. ESLAM method is general enough to integrate other kinds of noisy labels for model training, which will be tracked in their upcoming work [5].

5. Target dependent Sentiment Analysis

Target-dependent sentiment analysis on Twitter has fascinated growing research attention. Most previous effort rest on syntax, such as automatic parse trees, which are subject to noise for informal text such as tweets. Here researcher presented that competitive results can be attained without the use of syntax, by extracting a rich set of automatic features. A method is proposed where they split a tweet into a left context and a right context according to a given target, using distributed word representations and neural pooling functions to extract features. Both sentiment-driven and standard embeddings are used, and a rich set of neural pooling functions are explored. Sentiment lexicons are used as an additional source of information for feature extraction. Authors experiential that in a standard evaluation, the conceptually modest method gave a 4.8% absolute development over the state-of-the-art on three-way targeted sentiment classification, achieving the best reported results for this task They proposed a method, which is independent of external syntactic analysers, which gives better performance compared to the best previous method that uses syntax. This way solves the potential restriction of syntax-based method by avoiding the influence of noise by automatic syntactic analyser [6].

6. Annotate Sample Average – Distant Supervision Approach

The approach here is based on classification of tweets into polarity classes which is considered as a popular task in sentiment analysis. State-of-the-art clarifications to this tricky are based on supervised machine learning models trained from manually annotated examples. A disadvantage of these approaches is the high cost involved in data annotation. But here two freely available resources that can be exploited/use to solve the problem which are: 1) large amounts of unlabelled tweets obtained from

the Twitter API and 2) preceding lexical data in the form of opinion lexicons. Here the Writer tried to suggest Annotate-Sample-Average (ASA), a distant supervision method that uses these two resources to generate synthetic training data for Twitter polarity classification. Positive and negative training instances are produced by sampling and averaging unlabelled tweets containing words with the corresponding polarity. Polarity of words is determined from a given polarity lexicon. Their experimental results show that the training data generated by ASA (after tuning its parameters) produces a classifier that performs significantly better than a classifier trained from tweets annotated with emoticons and a classifier trained, without any sampling and averaging, from tweets annotated according to the polarity of their words. This method annotates tweets according to the polarity of their words, using a given polarity lexicon and generates balanced training data by sampling and averaging tweets containing words with the same polarity. ASA is based on the lexical polarity hypothesis: because further assumption is made which is tweets are short messages, opinion words are strong indicators of the sentiment of the tweets in which they occur, and therefore tweets with at least one word with a certain known prior polarity are more likely to express the same polarity on the message level than the opposite one [7].

7. Preprocessing Techniques and their interaction with Twitter Sentiment Analysis

Pre-processing is assumed to be the first step in text classification and choosing this very initial step as right pre-processing techniques can improve classification effectiveness. In this methodology the Author tried to experimentally compare 16 commonly used pre-processing techniques on two Twitter datasets for Sentiment Analysis, employing four popular machine learning algorithms, namely, Linear SVC, Bernoulli Naïve Bayes, Logistic Regression, and Convolutional Neural Net- works. Then the Author evaluated the pre-processing techniques on their resulting classification accuracy and number of features they produce. They found that techniques like lemmatization, removing numbers, and replacing contractions, improve accuracy, while others like removing punctuation do not. Finally, in order to investigate interactions desirable or otherwise between the techniques when they are employed simultaneously in a pipeline fashion, an ablation and combination study is contacted. The results of ablation and combination showed the significance of techniques such as replacing numbers and replacing repetitions of punctuation. They examined a significant number of pre-processing techniques, which have not been evaluated in a comparative study in the past and tested them in two datasets. Each technique was evaluated in four representative machine learning algorithms on accuracy. Moreover, they also distinguished some performance categories based on the results and counted the resulting number of features for each technique. Finally, an ablation study was performed for all, as well as for the high-performance techniques, in order to determine their interactions. [8].

8. Opinion Mining

The researcher has proposed a research survey that covers techniques and approaches that promise to directly enable opinion-oriented information seeking systems. Here mainly they focused on methods that seek to address the new challenges raised by sentiment aware applications, as compared to those that are already present in more traditional fact-based analysis. They also included material on summarization of evaluative text and on broader issues regarding privacy, manipulation, and economic impact that the development of opinion-oriented information-access services gives rise to. To facilitate future work, a discussion of available resources, benchmark datasets, and evaluation campaigns is also provided to us. The major goal of Author in this survey has been to cover techniques and approaches that promise to directly enable opinion-oriented information-seeking systems, and to convey to the reader a sense of excitement about the intellectual richness and breadth of the area. With the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. The sudden eruption of activity in the area of opinion mining and sentiment analysis, which deals with the computational treatment of opinion, sentiment, and subjectivity in text, has thus occurred at least in part as a direct response to the surge of interest in new systems that deal directly with opinions as a first-class object [9].

9. Text Processing of Twitter Data

Twitter sentiment analysis offers establishments ability to monitor public feeling towards the products and events related to them in real time. The primary step of the sentiment analysis is the text pre-processing of Twitter data. Most existing researches about Twitter sentiment analysis are focused on the extraction of new sentiment features. However, to select the pre-processing method is ignored. Here Author deliberated the properties of text pre-processing method on sentiment classification performance in two types of classification tasks and summed up the classification performances of six pre-processing methods using two feature models and four classifiers on _ve Twitter datasets. Their researches show that the accuracy and F1-measure of Twitter sentiment classification classifier are improved when using the pre-processing methods of expanding acronyms and replacing negation, but barely changes when removing URLs, removing numbers or stop words. The Naive Bayes and Random Forest classifiers are more sensitive than Logistic Regression and support vector machine classifiers when various pre-processing methods were applied. In this methodology six different pre-processing methods that affect sentiment polarity classification in the Twitter are discussed. Experimental results indicated that the removal of URLs, the removal of stop words and the removal of numbers minimally affect the performance of classifiers; furthermore, replacing negation and expanding acronyms can improve the classification accuracy. Therefore, removing stop words, numbers, and URLs is appropriate to reduce noise but does not affect performance. Replacing negation is effective for sentiment analysis. We select appropriate pre-processing methods and feature models for different classifiers for the Twitter sentiment classification task [11].

III. PROPOSED SYSTEM

(i) System Architecture

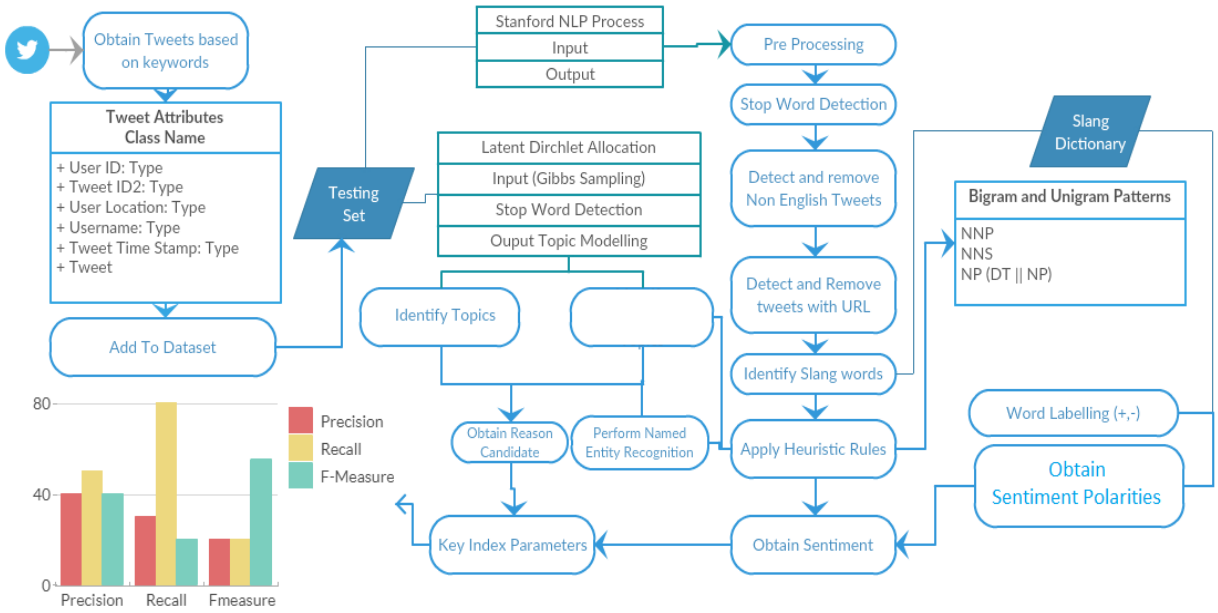

Figure 1. System design

As Shown in Figure 1. System start execution with Input as Users Post on Twitter.

Once system receives input using Historical Data it performs 4 functions as mentioned below.

- Finding multi-word aspects and using heuristic rules
- Employing proposed iterative bootstrapping with A-score metric
- Aspect pruning
- Detecting implicit aspects Using result of these 4 functions we obtain Sentiments behind the Tweet, Key Index Parameters, Identify Precision, Recall, F-Measure.

ii)System Flow:

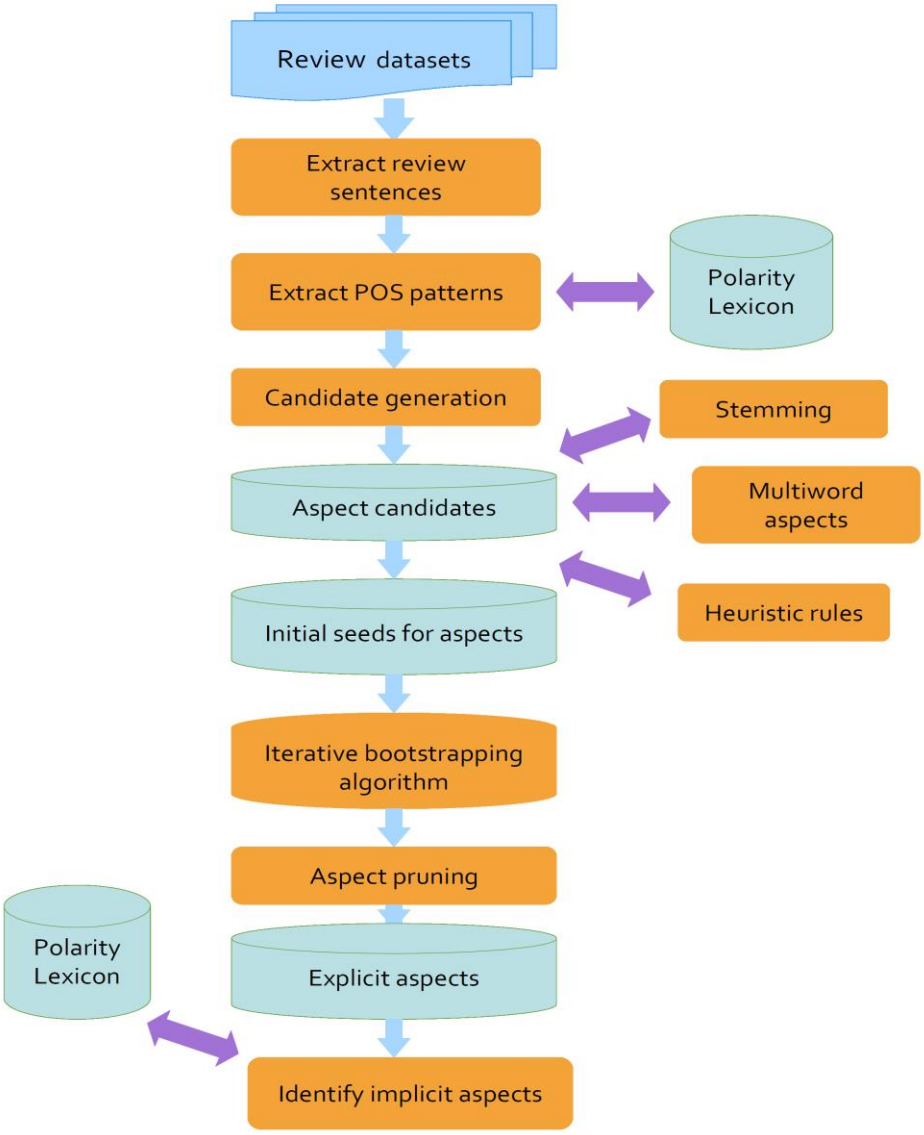

Figure 2. Flow Diagram

The input to the model is a dataset of reviews and the output is a set of aspects from the reviews. Firstly, the model segments the reviews into sentences, then for each sentence POS tagging is performed, and candidates for aspects and words corresponding to the polarity lexicon are extracted. Then a stemming step is used for each aspect candidate, and single- and multi-word aspects are extracted by utilizing a generalized statistical measure. Some heuristic rules are applied to filter less informative aspects. Then a bootstrapping algorithm is employed, based on a newly defined metric and an unsupervised initial seed set, to find aspects with the highest information. Not all aspects detected by the bootstrapping algorithm are genuine and useful aspects. There are also some redundant ones. Therefore, the model uses aspect pruning to remove these incorrect aspects. Finally, a new graph-based approach for extracting implicit aspect is introduced. Below, we discuss each of the functions in the proposed aspect detection model**.**newsfeed. This is a summary of what's been happening recently among their friends on Facebook. Every action their friends take is a potential newsfeed story. Facebook calls these actions "Edges." That means whenever a friend posts a status update, comments on another status update, tags a photo, joins a fan page, or RSVP's to an event it generates an "Edge," and a story about that Edge might show up in the user's personal newsfeed.

Part-of-Speech tagging and stemming

The model starts with extracting review sentences, and then for each of the sentences POS tagging is utilized, and candidates for aspects are extracted and stemmed. A Part-Of-Speech Tagger (POS Tagger) is a software package that reads text and assigns parts of speech tags to each word, such as noun, verb, adjective, etc. Here we focus on five POS tags: NN, JJ, DT, NNS and VBG, for nouns, adjectives, determiners, plural nouns and verb gerunds respectively. Stemming is used to select one single form of a word

instead of different forms. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. In this work we use the Stanford software package for both POS tagging and stemming.

POS patterns and candidate generation

Based on the observation that aspects are nouns, in the model we extract combination of noun phrases and adjectives from review sentences. We use several experimentally extracted POS patterns which we introduce as heuristic combinations. From Table 3.1 below, heuristic combinations of the first row selects the candidate aspects from the noun phrase patterns like ‘‘NN’’, ‘‘NNS’’, ‘‘NN NN’’ and etc. The second row uses patterns like ‘‘JJ NN’’, ‘‘JJ NNS’’, ‘‘JJ NN NN’’ and so on. The third row of the selects candidates based on the pattern ‘‘DT JJ’’, and the last row of the table uses heuristic patterns like ‘‘DT VBG’’, ‘‘VBG NN’’ and ‘‘NN VBG NN’’.

Description	**Pattern**
Nouns	Unigram to four-gram of NN and NNS
Nouns and adjectives	Bigram to four-gram of JJ, NN and NNS
Determiners and adjectives	Bigram of DT and JJ
Nouns and verb gerunds	Bigram to trigram of DT, NN, NNS and VBG

Heuristic combination POS patterns for candidate generation

Multi-word aspects

In the review sentences, some aspects that people talk about have more than one single word, ‘‘battery life’’, ‘‘signal quality’’ and ‘‘battery charging system’’ are examples. This step is to find useful multi-word aspects from the reviews. A multi-word aspect is represented by a = a1, a2, an where a1 represents a single word contained in a, and n is the number of words contained in a. In this work we propose a generalized version of FLR method to rank the extracted multi-word aspects and select the importance ones. FLR is a word scoring method that uses internal structures and frequencies of candidates (FLR: Frequencies and Left and Right of the current word). One of the advantages of the FLR method is its size-robustness, that it can be applied to small corpus with less significant drop in performance than other standard methods like TF and IDF, because it is defined using more fine-grained features**.**

The FLR for an aspect a is calculated as:

$$LR(a) = (lr(a_1) * lr(a_2) * \ldots * lr(a_n)) \wedge (1/n)$$

Where,

f(a) - is the sentence frequency for aspect a, in other words it is the number of sentences that contain aspect a, and

LR(a) - is the LR score of aspect a which is defined as a geometric mean of the scores of subset single-words as: In this equation, each ai represents a single-word in the multiword aspect a, and n is the number of single-words in a.

As an example, consider the word ‘‘Sound’’, for which the Type-LR score is calculated as follows:

$$l(Sound) = 1, \quad r(Sound) = 4, \quad lr(Sound) = \sqrt{4}$$

And Token-LR score for "Sound" is:

$$l(Sound) = 2, \quad r(Sound) = 10, \quad lr(Sound) = \sqrt{20}$$

As the Type-LR can reflect the number of different types of words connected to the current word, were in our methods we use Type-LR score.

CONCLUSION

In this article, we have studied the various well-known techniques of Sentiment Analysis in various fields like politics, marketing, weather forecasting and Social Media. Mining sentiment polarities expressed in Twitter messages is a meaningful while challenging task. Most of the existing solutions to Twitter sentiment analysis only consider textual information of Twitter messages and cannot achieve satisfactory performance due to unique characteristics of Twitter messages. To this end, we analysed Sentiment Analysis and find some interesting patterns for classification of Sentiments as Positive, Negative and Neutral.

REFERENCES

[1] Lei Wang, Jianwei Niu, Shui Yu, "SentiDiff: Combining Textual Information and Sentiment Difffusion Patterns for Twitter Sentiment Analysis", Journal of Latex Class Fies, VOL 14, NO. 8, August 2018
[2] Duyu Tang, Bing Qin, Furu Wei, Li Dong, Ting Liu, and Ming Zhou," A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification", IEEE/ACM Transactions on Audio, Speech, And Language Processing, VOL. 23, NO. 11, November 2015
[3] Kim Schouten and Flavius Frasincar," Survey on Aspect-Level Sentiment Analysis", IEEE Transactions on Knowledge and Data Engineering.
[4] Bing Liu," Sentiment Analysis and Opinion Mining", Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.
[5] Kun-Lin Liu,Wu-Jun Li, Minyi Guo, "Emoticon Smoothed Language Models for Twitter Sentiment Analysis", Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence
[6]Duy-Tin Vo and Yue Zhang, "Target-Dependent Twitter Sentiment Classification with Rich Automatic Features", Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015).
[7]Felipe Bravo-Marquez, Eibe Frank and Bernhard Pfahringer, "Annotate-Sample-Average (ASA): A New Distant Supervision Approach for Twitter Sentiment Analysis", Conference Paper · September 2016.
[8] Symeon Symeonidis, Dimitrios Effrosynidis, Avi Arampatzis, " A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis", Expert Systems With Applications 110 (2018) 298–310.
[9] Bo Pang Lillian Lee, "Opinion Mining and Setiment Analysis", Foundations and Trends in Information Retrieval, Vol. 2 pp- 1 – 35.
[10]Hao Wang, Dogan Can, Abe Kazemzadeh, Francois Bar, Shrikanth Narayanan," A System for Real-Time Sentiment Analysis of 2012 U.S. Presidential Election Cycle", Proceedings of 50th Annual Meeting of the Association for Computationl Linguistics, pp – 115 -120
[11] Zhao Jianqiang and Gui Xiaolin," Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis", Comparison Research on Text Pre-processing Methods.
[12] Ghazalech Beigi, Xa Hu, Ross Maciejwski and Huan Liu, "An Overview of Sentiment Analysis in Social Media and its Applications in Disaster Relief"