Sentimental Analysis Using Fuzzy and Naive Bayes

Ruchi Mehra[1], Mandeep Kaur Bedi[2], Gagandeep Singh[3], Raman Arora[4], Tannu Bala[5], Sunny Saxena[6]

Webtunix Solutions Pvt. Ltd. Mohali, India

webtunix@gmail.com, info@webtunix.com

Abstract. Sentimental Analysis is the best way to judge people's opinion regarding a particular post. In this paper we present analysis for sentiment behavior of Twitter data. The proposed work utilizes the naive Bayes and fuzzy Classifier to classify Tweets into positive, negative or neural behavior of a particular person. We present experimental evaluation of our dataset and classification results which proved that combined proposed method is more efficient in terms of Accuracy, Precision and Recall.

Keywords: Sentiment analysis, opinion mining, classification, machine learning, Natural language processing, deep learning, neural networks, Artificial Intelligence, Behaviour Analysis

I. INTRODUCTION

Sentiment analysis and opinion mining is the arena that examine people's sentiments, opinions, feelings from texts generated by the users. It is the dynamic exploration areas in natural language processing and is also broadly studied in web mining, data mining and social media analytics as sentiments are crucial influencers of behaviors of human [1]. As the quick growth of social media such as Facebook, Twitter and online review sites such as Amazon, IMDB, Yelp, sentiment analysis allurements rising at attentions from both exploration and business communities. Sentiment analysis is a moderately new zone, which contracts with mining user opinion repeatedly. For example, positive sentiment is, "Learning programming language is fun" alternatively, a negative sentiment is "it's a horrible, to learn programming language". Objective texts are supposed not to be stating any sentimentality, such as headlines of news [2]. There are various methods in which data of social network can be leveraged to give a superior indulgent of user opinion such as problems are at the heart of natural language processing (NLP) and data mining research. According to definition, the sentiments can be of any type i.e. positive, negative, or neutral sentiment, or a numeric rating score stating the intensity of the sentiment (e.g., 1–5 stars) in review sites like Yelp and Amazon. Sentiment analysis tasks are derived from the five components of the sentiment quintuple [3]. For example, the task of text level sentiment classification aims at the third component while discounting the other aspects. In same way, the task of fine-grained opinion mining focuses on the first four components of the quintuple [4]. From last 15 years, machine learning determined approaches almost control tasks of sentiment analysis. As feature demonstrate greatly touches the performance of a machine learner, number of studies in literature emphasis on actual types with domain expertise by hand and careful engineering. But this can be evaded by demonstration learning algorithms, which repeatedly discover discriminative and instructive text representations from data. Deep learning is a type of representation learning approach. It learns multiple levels of representation with nonlinear neural networks, each of which converts the representation at one level into a representation at a higher and more abstract level. The learned representations can be logically used as structures and applied for detection or classification tasks [5].

A. Data Characteristics: Twitter is a social microblogging service for social networking that allows its users post real time messages known as tweets. Tweets have many exclusive characteristics, which includes new challenges and figure up the means of carrying sentiment analysis on it as related to other domains. Following are some key characteristics of tweets [4]:

- **Message Length:** The maximum length of a Twitter message is 140 characters. This is dissimilar from existing sentiment classification research that concentrated on classifying longer texts, such as reviews of products and movies.

- **Writing technique:** Occurrence of improper spellings and cyber slang in tweets is more frequently as compared to other domains. As the messages are quick and short, people use acronyms, misspell, and use emoticons and other characters that convey special meanings [5].

• **Availability:** The amount of data available is immense. More people tweet in the public domain as compared to Facebook (as Facebook has many privacy settings) thus making data more readily available. The Twitter API facilitates collection of tweets for training.

• **Topics:** Twitter users post messages about a range of topics unlike other sites which are designed for a specific topic. This differs from a large fraction of past research, which focused on specific domains such as movie reviews [6].

• **Real time:** Blogs are updated at longer intervals of time as blogs characteristically are longer in nature and writing them takes time. Alternatively, post of tweeter being restricted to 140 letters and are updated very frequently. This provides a more real time feel and signifies the first reactions to events [13,14].

We now describe some basic terminology related to twitter:

• **Emoticons:** These are pictorial representations of facial expressions using punctuation and letters. The perseverance of emoticons is to define the user's mood.

• **Target:** Twitter users make use of the "@" symbol to refer to other users on Twitter. Users are automatically alerted if they have been mentioned in this fashion [15].

• **Hash tags:** Users use hash tags "#" to mark topics. It is used by Twitter users to make their tweets visible to a greater audience.

• **Special symbols:** "RT" is used to indicate that is used to indicate that it is a repeat of someone else's earlier tweet.

II.RELATED WORK

Pak and Paroubek [5] have classified the tweets as three categories like objective, positive and negative. In order to gather a corpus of objective posts, they fetched text messages from Twitter accounts of standard newspapers and magazine. Their classifier is established on the multinomial Naïve Bayes classifier that uses N-gram and POS-tags as features. Barbosa et al. [6] too classified tweets as objective or subjective and then the subjective tweets were classified as positive or negative. The feature space used included features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS of words.

Mining for entity opinions in Twitter, Batra and Rao [7] used a dataset of tweets spanning two months starting from June 2009. The dataset has roughly 60 million tweets. The entity was extracted using the Stanford NER, user tags and URLs were used to augment the entities found. A corpus of 200,000 product reviews that had been labeled as positive or negative was used to train the model. By Using mentioned corpus model calculated the possibility that a given unigram or bigram was utilized in a positive context and the probability that it was being used in a negative context.

Bifet and Frank [8] used Twitter streaming data provided by Firehouse, which gave all messages from every user in real-time. They investigated with fast incremental methods that were compatible to deal with data streams: stochastic gradient descent, multinomial naive Bayes and the Hoeffding tree. Therefore they concluded that SGD-based model, used with a suitable learning rate was the best.

Agarwal et al. [9] approached the task of mini sentiment from twitter, as a 3-way task of categorizing sentiment into positive, negative and neutral classes. They investigated with three types of models named as unigram model, a feature based model and a tree kernel based model. They designed a new tree representation for tweets for the tree kernel based model. In this, the feature based model t uses 100 features and the unigram model uses over 10,000 features. They finalized that features that associate prior polarity of words with their parts-of-speech tags are most important for the classification task. The tree kernel based model outperformed the other two.

III.METHODOLOGY

Number of tweets & Facebook posts are required to collect tweets and store them in real JSON layout, fairly easy to convert into different data models depending on our storage. Methodology that we follow in our work is given below in the form of flow chart

- We need to create an account that interacts with the Twitter API in order to have access to Twitter data programmatically.
- We use Twitter REST API to interact with their service.

- We need to use the OAuth interface in order to authorize our app to access Twitter.

After that collecting a number of tweets and stored them in JSON as mentioned above, the structure of a tweets and posts are likely to be looking as follow:

- Text: text of tweet itself.
- Created_at: creation date.
- Favorite_count, retweet_count: number of favourites & retweets.
- Favorited, retweeted: boolean stating whether authenticated user have favourited or retweeted this tweet.

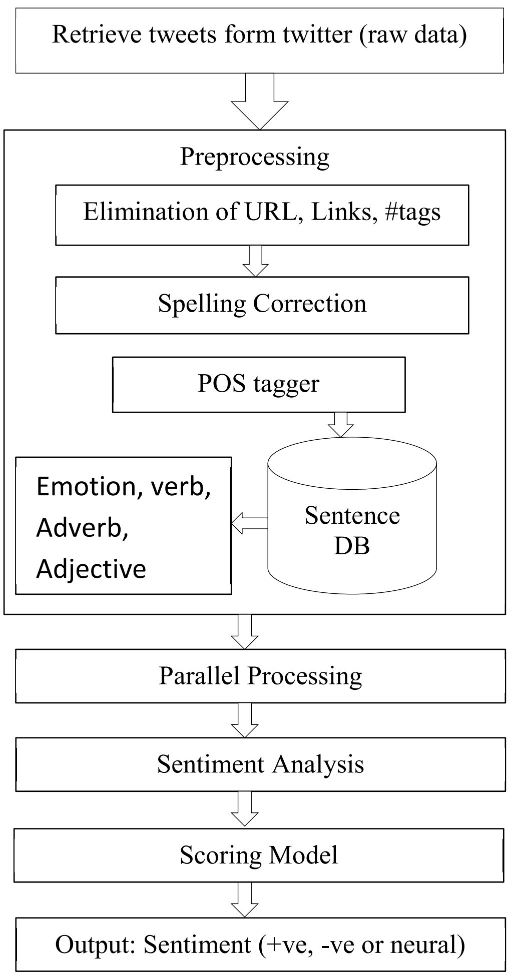

Fig 1.1 Flow diagram of proposed work

- Lang: acronym for language (e.g. “en” for english).
- ID: tweet identifier.
- Place, coordinates, geo: geo-location data if accessible.
- User: author’s full profile.
- Entities: list of entities like URLs, @-mentions, hashtags & symbols.
- in_reply_to_user_id: user identifier if tweet is a reply to a definite user.
- in_reply_to_status_id: status identifier id tweet is a reply to a definite status

A lot of information are involved in collected tweets. We start our analysis by breaking the text down into words (Tokenization one of the most basic steps in text analysis) to split a stream of text into smaller units called tokens (words or phrases). We cannot recognize: @-mentions, emoticons, URLs and #hash-tags as distinct token. We can use Named Entity Recognition to overcome the problem of pre-processing names

We should eliminate stop words, some words particularly common in every language that don’t convey a particular meaning, especially if taken out of context. Stop-word removal is one important step that should be considered during the pre-processing stages using available lists (e.g. NLTK provides a simple list for English stop-words). All of the following should be handled in pre-processing:

1. Convert tweets to lower case.
2. Eliminate URLs.
3. Eliminate "@username".
4. Eliminate #hashtags.
5. Remove punctuation at starting and ending of the tweets.
6. Replace multiple whitespaces with a single whitespace.

Moreover, filtering step contains the following functions:

1. Remove stop words.
2. Remove repeating letters.
3. Words must start with an alphabet.

Twitter extraction and data cleaning are accomplished, whereas the next will be ontology model building and sentiment analysis.

The first process is to build the ontology model using the data extracted from the social media platform, Twitter and Facebook. Defining ontology helps in sharing a common understanding of knowledge among people and software agents, and makes domain knowledge reusable. This process has tasks such as social media data extraction and data cleaning, identifying negative sentiment in tweets text.

The ontology model can classify tweets with the negative sentiments to do sentiment analysis. SentiStrength tool is used to identify the tweet with the negative sentiments. The tool can identify the polarity of words in the tweet sentences. The tweets with their higher negative polarity compared to their negative polarity were considered as tweets with negative sentiments. Word context is a bag of three word tokens: the previous word, the word itself, and the next word (n-gram). We just needed to annotate the expressions with their contextual polarity.

We define the Semantic Orientation (SO) of a word as the difference between its associations with negative words. We will build our own list of negative terms. Use of dictionary for good and bad words is the modest form of sentiment analysis. Each word in a sentence has a score, typically +1 for less negative sentiment, 0 to negative, and -1 for more negative. Then, we merely merge scores of each word in sentence to get a concluding sentiment total. Consider a text as a “bag of words” is other method. We give each text as a 1 by N vector, where N is the size of vocabulary. Next step is to apply naïve Bayes classifier.

We use fuzzy functions to calculate overall sentiment score. To mitigate some sentimental parameter ambiguities, the result of the entire sentiment analysis model is processed by the fuzzy classifier where the classification decision is achieved by using linguistic variables: more negative, negative or less negative.

IV. RESULTS AND DISCUSSIONS

In our proposed technique we perform the comparison of two techniques with our proposed method. Following is the two algorithms that we perform in our work.

A. Naïve Bayes:

Naive Bayes classifiers are components of simple probabilistic classifiers constructed on applying Bayes' theorem with strong (naive) independence assumptions between the features in machine learning [10].

Naive Bayes has been studied broadly since the 1950s and remains a standard method for text categorization, the problem of judging documents as fitting to one group or the other with word frequencies as the features. With suitable pre-processing, it is modest in this domain with more innovative methods including support vector machines. It also finds request in automatic medical diagnosis [11].

B. Fuzzy

Fuzzy set theory is an enhancement of conventional set theory that contracts with the idea of partial truth. Age processing is performed using fuzzy logic. Following is the tabular and graphical representation of our result performed [12].

Table 1.1 Comparison table

		Negative	Neutral	Positive
Accuracy	Naive Bayes	86.1111	92.3077	88.2923
	Fuzzy	91.6665	81.060	84.658
	Proposed	100	96.15385	96.634
Recall	Naive Bayes	84.2105	100	89.0764
	Fuzzy	85.714	92.3077	83.5805
	Proposed	100	100	100
Precision	Naive Bayes	88.889	84.6154	88.3903
	Fuzzy	100	76.3636	92.1212
	Proposed	100	92.3077	93.2692

C. List of Parameters

- *Accuracy*

This refers to the ability of the classifiers to correctly measure the sentimental analysis. This is defined as the ratio of correctly classified data to the total classified data.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

where, TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

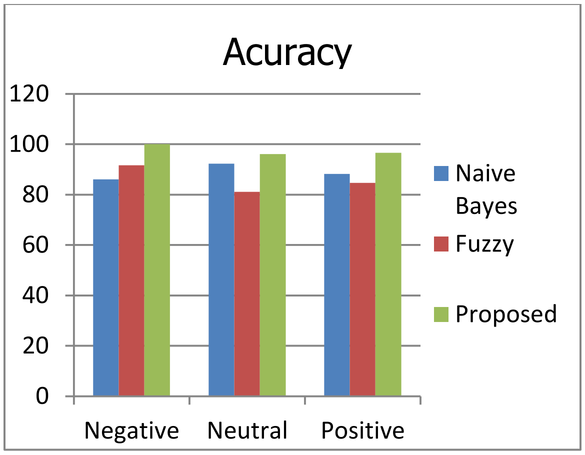

Fig 1.2 Accuracy comparison

- *Precision*

Precision is the division of retrieved instances that are significant. Precision precedes all retrieved documents into account, but it can also be estimated at a specified cut-off rank, considering only the highest results returned by the system. In the fig 1.3 we compare the precision value of three different classifiers.

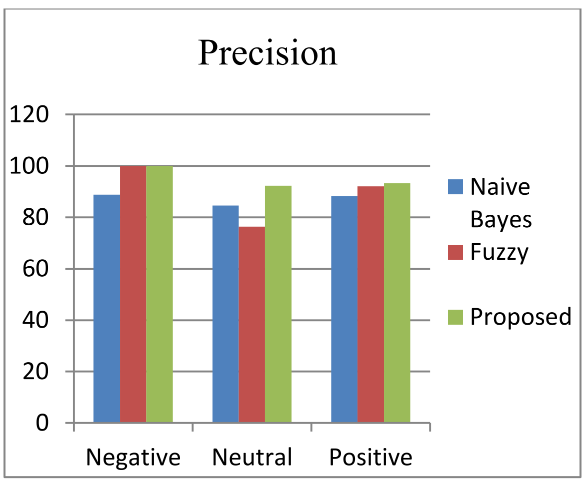

Fig 1.3 Precision comparison

- *Recall*

While recall is the fraction of applicable instances that are recovered. In binary classification, recall is known as sensitivity. So it can be observed at as the possibility that a relevant document is recovered by the query.

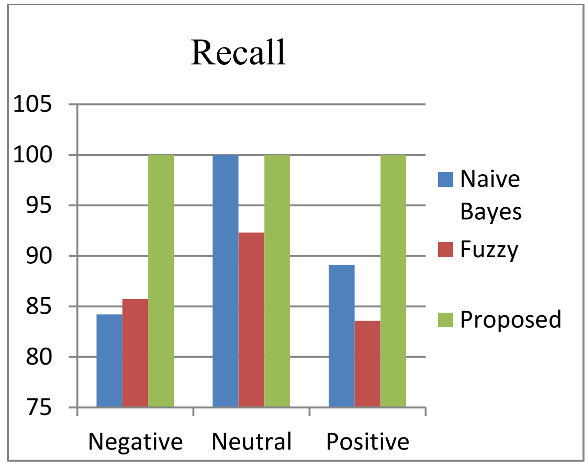

Fig 1.4 Recall comparison

V. Conclusion and Future Work

In this paper we have presented a way in machine learning techniques can be applied to large sets of data to establish membership, in this case positivity, negativity and neutral. We have looked at common process in NLP that can help us derive the meaning or context of a given phrase. We have demonstrated how to collect an original corpus for sentiment classification and the refinement that is needed with such data. We have applied a hybrid of naive Bayes and Fuzzy classifier to this set conduct sentiment analysis and have found this process to be successful. On analysis of our results we have confirmed that proposed algorithm offer better performance when conducting the classification process supporting results.

REFERENCES

[1] 1. Pang B, Lee L. Opinion mining and sentiment analysis. Found Trends Inf Retr 2008, 2:1–135.

[2] Liu B. Sentiment analysis and opinion mining. Synth Lect Human Lang Technol 2012, 5:1–167.

[3] Feldman R. Techniques and applications for sentiment analysis. Commun ACM 2013, 56:82–89.

[4] Manning CD, Schütze H. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press; 1999

[5] A. Pak and P. Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320–1326.

[6] R. Parikh and M. Movassate, “Sentiment Analysis of UserGenerated Twitter Updates using Various Classification Techniques”, CS224N Final Report, 2009

[7] A. Go, R. Bhayani, L.Huang. “Twitter Sentiment Classification Using Distant Supervision”. Stanford University, Technical Paper, 2009

[8] J. Read. “Using emoticons to reduce dependency in machine learning techniques for sentiment classification”. In Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2005

[9] L. Barbosa, J. Feng. “Robust Sentiment Detection on Twitter from Biased and Noisy Data”. COLING 2010: Poster Volume, pp. 36-44.-

[10] B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. ACL.

[11]P. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. ACL.

[12] C M Whissel. 1989. The dictionary of Affect in Language. Emotion: theory research and experience, Acad press London.

[13]T. Wilson, J. Wiebe, and P. Hoffman. 2005. Recognizing contextual polarity in phrase level sentiment analysis. ACL.

[14] Bo Pang and Lillian Lee. “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.” 2004. Proceedings of ACL, pp. 271–278.

[15] Moreno, M., Jelenchick, L., Egan, K., Cox, E., Young, H., Gannon, K., et al.: Feeling Bad on Facebook: Depression Disclosures by College Students on Social Networking Site. Depression and Anxiety 28, 447–455, 2011