Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.8 No.S1, 2019, pp.58-62
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajcst-2019.8.S1.1957

FBSC: An Analyzing Sentiments Using Fuzzy Based Bayesian Classification

M. Karthica[1] and P. Sudarmani[2]
[1&2]Assistant Professor, Department of Computer Science, Sri Vasavi College, Tamil Nadu, India
E-Mail: karthica92@gmail.com, sudarmani087@gmail.com

(Received 15 December 2018; Accepted 25 January 2019; Available online 10 February 2019)

***Abstract* - The thriving Micro blog service, Twitter, attracts more people to post their feelings and opinions on various topics. Millions of users share opinions on totally different aspects of life on a daily basis. It observing the user's sentiment options topics in the twitter network. The sentiment classification is comparable to the user's opinions that are based on dynamic manner. An optimal Fuzzy based Bayesian classification is a capable way that has been proposed to improve the classification accuracy, unless the large amount of information on these platforms make them viable for use as data sources, in applications based on sentiment analysis. The research work developed a Fuzzy based Bayesian sentiment classification (FBSC) based dynamic online twitter search data architecture that ensures truthful positive, negative and neutral results.**

***Keywords:* Data Mining, Twitter, Sentiment Analysis, Bayesian Classification.**

I. INTRODUCTION

Data Mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data Mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, Data Mining consists of more than collecting and managing data, it also includes analysis and prediction.

Data Mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association, sequence or path analysis, classification, clustering, and forecasting. Many simpler analytical tools utilize a verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis.

The booming Micro blog service, Twitter, attracts more people to post their feelings and opinions on various topics. The posting of sentiment contents can not only give an emotional snapshot of the online world but also have potential commercial Hu .M and B. Liu, [1], Pantel .P and D. Ravichandran [2], financial Shen .D, J. Wu, B. Cao, J.-T. Sun, Q. Yang, Z. Chen, and Y. Li, [3] and sociological values A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe [4]. However, facing the massive sentiment tweets, it is hard for people to get overall impression without automatic sentiment classification and analysis. Therefore, there are emerging many sentiment classification works showing interests in tweets L. T. Nguyen, P. Wu, W. Chan, W. Peng, and Y. Zhang [5]

Topics discussed in Twitter are more diverse and unpredictable. Sentiment classifiers always dedicate themselves to a specific domain or topic named in the paper. Namely, a classifier trained on sentiment data from one topic often performs poorly on test data from another S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, [8]. One of the main reasons is that words and even language constructs used for expressing sentiments can be quite different on different topics. In social media, a Twitter user may have different opinions on different topics A. Go, R. Bhayani, and L. Huang, [9] .The, topic adaptation is needed for sentiment classification of tweets on emerging and unpredictable topics.

Textual information can be categorized into facts and opinions. Facts are objective expressions about entities, events and their properties, items of information, or state of affairs existing, observed, or known to have happened, and which is confirmed or validated to such an extent that is considered reality. Sometimes, when developing breaking research or just purchasing decision about certain product], It often look for people that had experiences in a field of our research or product we are interested in. It is completely natural that we are looking for other's opinions.

Sentiment analysis is very topical issue; both in industry and academia understand advantages of sentiment extraction from web text. Especially business companies and institutions quickly realized the importance of caring quality control as well as marketing research for selling their products and services.

The opinion mining is often associated with another research topic – information retrieval (IR). Nevertheless, opinion mining proves to be a lot difficult task. The primary reason is characteristics of the data sources. In IR, the algorithms operate on factual data, while in opinion mining input data is only subjective information. In practice, this means that opinion mining is needed to go a step further then information retrieval and analyze sentences and phrases deeper with respect to their semantics. During the facts analysis one is interested in simple characteristics and extracting it. In opinion mining the additional task is to determine the nature of opinion: whether it is positive or negative in general; what features does it describes; what features are valued, which are not etc.

The objective of this paper is to classify short Twitter contents with respect to their sentiment using data mining techniques. Twitter messages, or tweets, are limited to twitter api. This limitation makes it more difficult for people to express their sentiment and as a consequence, the classification of the sentiment will be more difficult as well. The sentiment can refer to two different types: emotions and opinions. These opinions can be divided into three classes: positive, neutral and negative. The tweets are then classified with an algorithm to one of those three classes.

II. RELATED WORK

A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe [4] authors considered a strategy of building statistical models from the social media dynamics to predict collective sentiment dynamics. The collective sentiments change without delving into micro analysis of individual tweets or users and their corresponding low level network structures. The sentiment classification starts with identifying the semantic orientation of words, and then goes to higher level text structure like the semantic orientation of sentences and documents. Several techniques are used to achieve this task: Words were directly weighted by lexicons of semantic words which were manually or automatically constructed.

L. T. Nguyen, P. Wu, W. Chan, W. Peng, and Y. Zhang [5] authors illustrated the micro blogging site Twitter generates a constant stream of communication, some of which concerns events of general interest. An analysis of Twitter may, therefore, give insights into why particular events resonate with the population. This article reports a study of a month of English Twitter posts, assessing whether popular events are typically associated with increases in sentiment strength, as seems intuitively likely. Using the top 30 events, determined by a measure of relative increase in (general) term usage, the results give strong evidence that popular events are normally associated with increases in negative sentiment strength and some evidence that peaks of interest in events have stronger positive sentiment than the time before the peak.

M. Thelwall, K. Buckley, and G. Paltoglou, [6] an author proposed the micro blogging websites have evolved to become a source of varied kind of information. This is due to nature of micro blogs on which people post real time messages about their opinions on a variety of topics, discusses current issues, complain, and express positive sentiment for products they use in daily life. In fact, companies manufacturing such products have started to poll these micro blogs to get a sense of general sentiment for their product. Many times these companies study user reactions and reply to users on micro blogs.

A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau [7] authors discussed the opinions are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are, to a considerable degree, conditioned upon how others see and evaluate the world. For this reason, when we need to make a decision we often seek out the opinions of others. This is not only true for individuals but also true for organizations. Opinions and its related concepts such as sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining.

S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, [8] authors proposed a spectral feature alignment (SFA) algorithm to align domain-specific words from different domains into unified clusters, with the help of domain independent words as a bridge. In this way, the clusters can be used to reduce the gap between domain-specific words of the two domains, which can be used to train sentiment classifiers in the target domain accurately. Compared to previous approaches, SFA can discover a robust representation for cross-domain data by fully exploiting the relationship between the domain-specific and domain independent words via simultaneously co-clustering them in a common latent space.

A. Go, R. Bhayani, and L. Huang, [9] authors introduced a novel approach for automatically classifying the sentiment of Twitter messages. These messages are classified as either positive or negative with respect to a query term. This is useful for consumers who want to research the sentiment of products before purchase, or companies that want to monitor the public sentiment of their brands. There is no previous research on classifying sentiment of messages on micro blogging services like Twitter..

X. Wan, [10] authors focused on the problem of cross-lingual sentiment classification, which leverages an available English corpus for Chinese sentiment classification by using the English corpus as training data.

Machine translation services are used for eliminating the language gap between the training set and test set, and English features and Chinese features are considered as two independent views of the classification problem. To proposed a co-training approach to making use of unlabeled Chinese data.

III. RESEARCH METHODOLOGY

Twitter comments as input which contains the JAVA framework where the dynamic fuzzy based Bayesian classification algorithm is applied to the sentiment classification. The proposed architecture diagram is described in fig. 1.

A. Twitter Connectivity

Twitter provides its own API through which developers may obtain limited streams of live tweets. This interface was used through the Twitter4j java library to filter all available live tweets for any containing any complete company name, or ticker symbol on the Dow Jones Industrial Average. All tweets matching the filter were saved along with all available metadata including timestamp, sender, re-tweet status etc. Because all of this information was collected in a real time streaming environment with very brief time windows, and no modifications to the data, this approach lends itself well to the type of moment by moment analysis that must be conducted for technical stock analysis.

1. Data availability

Another difference is the magnitude of data available. With the Twitter API, it is very easy to collect millions of tweets for training. In past research, tests only consisted of thousands of training items. Twitter users post short messages about a variety of topics unlike other sites which are tailored to a specific topic. This differs from a large percentage of the research, which focused on any domains.

The first thing is need to do is to create a sample application within twitter account, grant it access and generate authentication tokens (although, the consumer tokens will also work and it will be automatically generated upon creating the application). It can generate access tokens using the consumer keys. It can be done using the API or through the interface as well with generating using the Twitter website itself. Ensure that the access level comes the same as in consumer keys or regenerate it.

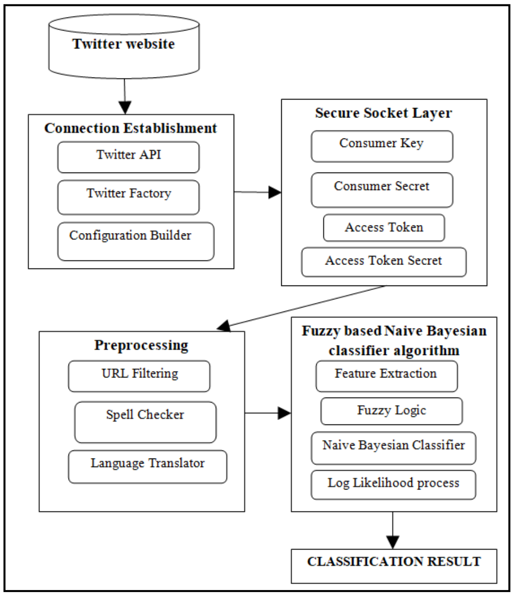

Fig.1 Proposed Architecture diagram

C. Feature Extraction Process

Machine learning, it is necessary to extract certain clues from the text that may lead to an effective correct classification. Clues about the original data are usually stored in the form of a feature vector, F = (f_1, f_2… f_n). Each coordinates of a feature vector represents one clue, also called a feature, "f_i" of the original text.

On setting out to classify a document, starts generally with depicting a very large number of words that need to be considered, even though very few of the words in the corpus are actually expressing sentiment. These extra features have two clear drawbacks that need to be eliminated. The first is that they show down the process of document classification, since there are far more words than needed. The second is that they can actually reduce accuracy, since the classifier is obliged to consider these words when classifying a document.

Regular expression based tokenize for Twitter. Focused on tokenization and pre-processing to train classifiers for sentiment, emotion, or mood.

1. Data Cleaning

A tweet has to be cleaned so that it doesn't affect the accuracy of the model. Among the cleaning steps:

1. Remove web links
2. Remove hash tags
3. Remove quotes (@psg, etc.)
4. Remove punctuations

A stemmer is a process for removing the commoner morphological and in flexional endings from words in English. For example the word 'running' will become 'run': we won't count 'running' and 'run' separately in the twitter streaming dataset.

Pre-processing the data is the process of cleaning and preparing the text for classification. Online texts contain usually lots of noise and uninformative parts such as HTML tags, scripts and advertisements. In addition, on words level, many words in the text do not have an impact on the general orientation of it.

Keeping those words makes the dimensionality of the problem high and hence the classification more difficult since each word in the text is treated as one dimension. Here is the hypothesis of having the data properly pre-processed: to reduce the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis.

The whole process involves several steps: online text cleaning, white space removal, expanding abbreviation, stemming, stop words removal, negation handling and finally feature selection. All of the steps but the last are called transformations, while the last step applying some functions to select the required patterns is called filtering.

E. Fuzzy based Bayesian Classification

Fuzzy based Bayesian classifiers are based around the Bayesian rule, a way of looking at conditional probabilities that allows you to flip the condition around in a convenient way. A conditional probably is a probably that event *X* will occur, given the evidence *Y*. That is normally written *P*(*X*, *Y*). The Bayesian rule allows us to determine this probability with the probability of the opposite result and of the two components individually: $P(X, Y) = P(X)\ P(Y, X)/P(Y)$. This restatement can be very helpful that are trying to estimate the probability of something based on examples of it occurring.

P (sentiment | sentence) = P(sentiment) P(sentence| sentiment)/P(sentence)

The probability that a document is positive or negative is estimated, given its contents so, that in terms of the probability of that document occurring if it has been predetermined to be positive or negative. This is convenient, because the examples of positive and negative opinions from the data are set above.

The thing that makes this a "fuzzy" Bayesian process is that a big assumption is made about how the calculation at the probability of the document occurring: that it is equal to the product of the probabilities of each word within it occurring. This implies that there is no link between one word and another word. This *independence assumption* is clearly not true: there are lots of words which occur together more frequently that either does individually, or with other words, but this convenient fiction massively simplifies things for us, and makes it straightforward to build a classifier.

The probability of a word occurring given a positive or negative sentiment are estimated by looking through a series of examples of positive and negative sentiments and counting how often it occurs in each class. This is what makes this *supervised learning* - the requirement for pre-classified examples to train on.

P (sentiment | sentence) = P (sentiment)
*P (sentence | sentiment) / P(sentence)

The dividing P (sentence) is dropped, as it's the same for both classes, and to rank them rather than calculate a precise probability. The use independence assumption to let, treat P (sentence | sentiment) as the product of P (sentence | sentiment) across all the tokens in the sentence.

So, the estimation P (token | sentiment) as

count (this token in class) +
1 / count (all tokens in class) + count(all tokens)

The extra 1 and count of all tokens is called 'add one' or Laplace smoothing, and stops a 0 finding its way into the multiplications. With an unseen token in it would score zero, if haven't it any sentence.

The classify function starts by calculating the prior probability (the chance of it being one or the other before any tokens are looked at) based on the number of positive and negative examples - in this example that'll always be 0.5 as we have the same amount of data for each. The incoming documents are tokenized, and for each class multiply together the likelihood of each word being seen in that class. The final results are sorted, and return the highest scoring class.

The Bayesian Naive Classifier selects the most likely classification V_{nb} given the attribute values $a_1, a_2, \ldots, a_n$.
The results in:

$$\mathrm{Vnb} = \mathrm{argmax}_{\mathrm{vj} \in \mathrm{V}} \mathrm{P}(\mathrm{V_j}) \pi \mathrm{P}(\mathrm{ai}|\mathrm{vj}) \quad (1)$$

Where
n= the number of training examples for which v = vj
ne = number of examples for which v = vj and a=ai
p = a priori estimate for P (ai | vj)
m = the equivalent sample size

F. Dynamic Natural Language Processing

DNLP is a technology that concerns with dynamic natural language generation (DNLG) and dynamic natural language understanding (DNLU). DNLG uses some level of underlying linguistic representation of text, to make sure that the generated text is grammatically correct and fluent. Most DNLG systems include a syntactic realize to ensure that grammatical rules such as subject-verb agreement are obeyed, and text planner to decide how to arrange sentences, paragraph, and other parts coherently.

The most well-known DNLG application is machine translation system. The system analyses texts from a source language into grammatical or conceptual representations and then generates corresponding texts in the target language. DNLU is a system that computes the meaning representation, essentially restricting the discussion to the domain of computational linguistic. DNLU consists of at least of one the following components; tokenization, morphological or lexical analysis, syntactic analysis and semantic analysis. In tokenization, a sentence is segmented into a list of tokens. The token represents a word or a special symbol such an exclamation mark. Morphological or lexical analysis is a process where each word is tagged with its part of speech.

IV. PERFORMANCE EVALUATION

To the best of the knowledge, there is no annotated dataset for opinion retrieval in Twitter. Therefore, a new dataset for this task is created. About 30 million tweets are crawled and indexed using the Twitter API. All tweets are English. Using these tweets a search engine is implemented.

The impact of the dataset size on the performance of the system is also examined. To measure the performance, The F-measure is used,

$$\mathrm{F} = (1 + \beta^2) \frac{\mathrm{precision} \cdot \mathrm{recall}}{\beta^2 \cdot \mathrm{recall} + \mathrm{recall}}$$

To compare our adaptive algorithm Fuzzy based Bayesian classification with five baseline algorithms, i.e., TASC, DT, MSVM, RF, MS3VM and CoMS3VM in Table I.

TABLE I SHOWING COMPARISON OF CLASSIFICATION ALGORITHM

Methods	Accuracy	F-score
DT	50.63%	0.3400
MSVM	50.36%	0.4624
RF	54.03%	0.5063
MS3VM	51.16%	0.4344
CoMS3VM	67.95%	0.6400
TASC	68.82%	0.6528
Proposed Fuzzy	79.24%	0.723

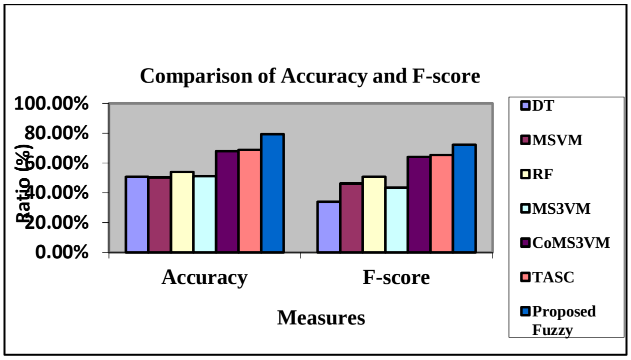

Fig.2 Comparison of Accuracy and F-score

V. CONCLUSION

Fuzzy based Bayesian classification is a capable way that has been proposed to improve the classification accuracy, unless the large amount of information on these platforms makes them viable for use as data sources, in applications based on sentiment analysis. The research work developed a Fuzzy based Bayesian sentiment classification (FBSC) based dynamic online twitter search data architecture that ensures truthful positive, negative and neutral results. The FBSC can easily analyze the sentiment word, classification information on the user input query. Machine learning algorithms (Fuzzy Naive Bayesian), can achieve high accuracy for classifying sentiment when using this method.

REFERENCES

[1] Hu .M and B. Liu, "Mining and Summarizing Customer Reviews," *Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '04)*., pp. 168-177, 2004.

[2] Pantel .P and D. Ravichandran, "Automatically Labeling Semantic Classes," *Proc. Conf. North Am. Ch. Assoc. for Computational Linguistics: Human Language Technologies (NAACL-HLT '04)*., pp. 321-328, 2004.

[3] Shen .D, J. Wu, B. Cao, J.-T. Sun, Q. Yang, Z. Chen, and Y. Li, "Exploiting Term Relationship to Boost Text Classification," *Proc. 18th ACM Conf. Information and Knowledge Management (CIKM '09)*., pp. 1637-1640, 2009.

[4] Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," *in Proc. 4th Int. AAAI Conf. Weblogs Soc. Media*., Vol. 10, pp. 178–185, 2010.

[5] L. T. Nguyen, P. Wu, W. Chan, W. Peng, and Y. Zhang, "Predicting collective sentiment dynamics from time-series social media," *in Proc. 1st Int. Workshop Issues Sentiment Discovery Opinion Mining*., pp.6, 2012,

[6] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in twitter events," *J. Am. Soc. Inform. Sci. Technol.*, Vol. 62, No. 2, pp. 406–418, 2011.

[7] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," *in Proc. Workshop Lang. Soc. Media*., 2011, pp. 30–38.

[8] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," *in Proc. 19th Int. Conf. World Wide Web*., pp. 751–760, 2010.

[9] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Computer Science Department, Stanford, USA*., pp. 1–12, 2009.

[10] X. Wan, "Co-training for cross-lingual sentiment classification," *in Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int.Joint Conf. Natural Language Process. AFNLP.*, Vol. 1, pp. 235–243, 2009.