ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.4

Sentiments Analysis of Twitter Data using K- Nearest Neighbour Classifier

Abhilasha Tyagi[1], Naresh Sharma[2]
M.Tech Student[1], Assistant Professor[2]
Department of Computer Science & Engineering
SRM University, India

Abstract:
Twitter is the most used social platform to express views regarding any subject. Twitter data plays an important role in sentiment classification. This Paper presents an approach for classification of sentiments using k-nn classifier with bag of words approach as feature selector. The result obtained shows that k-nn approach gives a higher accuracy as compared to polarity based classification of sentiments.

Keywords: Sentiments analysis, Twitter, k-nearest neighbour, Polarity based classification.

I. INTRODUCTION

Micro blogging has turned into an extremely mainstream communication tool among web users. Many users share sentiments on various parts of life, everyday on prevalent sites, for example, Twitter and Facebook. Prodded by this development, companies and media associations are progressively looking for approaches to mine these web-based social networking for information about what individuals consider for their organizations and items. Political gatherings might be intrigued to know whether individuals bolster their events or not. Associations that are social need to know individuals' assessment on verbal confrontations. The information can be obtained from micro blogging administrations, which users post sentiments on numerous parts of their life regularly. However, micro blogging information is unique in relation to normal content because it is extremely noisy in nature. A great deal of fascinating work is done keeping in mind the end goal to recognize feelings or sentiments from Twitter micro blogging information too. We intend an approach to automatically extract sentiment from a tweet. It is extremely supportive in nature that it enables input to be aggregated without any manual intercession. There are many researches in the area of sentiment classification. Mainly its greater part has concentrated on characterizing larger pieces of text, similar to reviews. Tweets (micro blogs) are dissimilar from reviews essentially due to their motivation: while tweet is easygoing and partial to 140 characters of content. For the most part, tweets are not as insight fully created as reviews. Organizations can likewise utilize this to accumulate basic feedback about issues in recently launched items. In previous researches like Pang et al. [2] movie reviews are examined over many classifiers. This work of Pang et al. [2] is provided as a base in many research and many researchers have used the basic procedure in many areas. With the substantial scale of topics talked about on Twitter, it would be tremendously difficult to manually gather enough information to train a sentiment classifier for tweets. We run classifiers trained on emoticon data not in favor of a test set of tweets (which could conceivably have emoticons in them).

II. LITERATURE REVIEW

Pang B. in 2002 measured the problem of characterizing reports by general feeling, for example, deciding if an audit is positive or negative. Utilizing film surveys for information, that basic machine learning strategies absolutely outflank user created baselines. Nonetheless, the techniques that were utilized are Naive Bayes, Support Vector Machine and Maximum Entropy. These techniques don't do well on classification of sentiments as on customary subject related classification of sentiments. The author finishes up by looking at factors that make the sentiment classification issue much difficult. Pang B. and Lee L in 2004, expressed that analysis of sentiments tries recognizing the perspectives of a basic content span; an example is classifying a film audit as positive or negative. To decide the polarity, the creator defines a novel machine-learning strategy which applies content classification strategies to only the subjective parts of the text document. Separating these parts can be executed utilizing proficient strategies for discovering least cuts in graphs which significantly encourages consolidation of cross-sentence relevant limitations. Hassan S., et al, described in their paper that analysis of sentiment over Twitter gives companies a very best way to analyze the public’s sentiment for their brand, business, products, etc. An extensive variety of features and techniques for preparing opinion classifiers for Twitter datasets have been inquired about as of late with varying. Bing Liu., et al, 2012 expressed that Opinions are vital to every single human activity and are key influencers of our practices. This isn't valid for people yet additionally valid for organizations. Opinions and its connected ideas, for example, feelings are the concern of investigation of sentiments. The initiation and fast development of the field match the online networking e.g., surveys, sites, micro blogs, Twitter, and social networks, on the grounds that without precedent for mankind's history, an immense volume of obstinate information is being recorded in computerized frames. Since mid 2000, sentiment examination has become a standout amongst the most dynamic research regions. Emma H., et al., clarified that it is trying to comprehend the most recent patterns and synopses the state or general sentiments about items because of the huge assorted variety and size of online networking information, and this makes the need of automated and ongoing feeling extraction and mining. Mining on the web feeling is a type of sentiment examination that is dealt with as a troublesome content arrangement undertaking. Emma H., et al., clarified that it is trying to comprehend the most recent patterns and synopses the state or general sentiments about items because of the huge assorted variety and size of online networking information, and this makes the need of automated and ongoing feeling extraction and mining. Mining on the web feeling is a type of sentiment examination that is dealt with as a troublesome content arrangement undertaking. Lo Y.W, et al., in 2013 expressed that the Web has significantly changed the best approach to express sentiments on specific items that has been obtained and utilized, or for administrations that we have gotten in the different businesses. Sentiments or surveys can be effortlessly posted, for example, in dealer destinations, audit entryways, web journals, Internet gatherings, and considerably more. This information is usually alluded to as client created service or client created media. Both the item makers, and in addition potential clients are extremely intrigued by this online verbal, as it gives item makers data on their clients different preferences, and additionally the positive and negative remarks on their items at whatever point accessible, giving them better learning of their items restrictions and favourable circumstances over contenders; and furthermore giving potential clients helpful and direct data on the items or potentially administrations to help in their buy basic leadership procedure. Bing L. and Chan. K. in 2014 clarified that amount of clients shared what they think on small scale blog administrations. Twitter is critical stage for take after estimation of opinion which is an exceptionally difficult issue. Public feeling examination is an exceptionally basic to investigate, break down and sort out clients sees for better basic leadership. Sentiment examination is procedure of recognizing good and bad sentiment, feelings and examinations in content. It is valuable for product users to examine the conclusion of items, or organizations need to screen people in general opinion of their brands. Gautam G. and Yadav P in 2014 define another method for communicating the feelings and emotions of people. In general it is a way which tremendously measures the information where clients can see the emotions of different clients which are categorized into various classes of sentiments and are progressively developing as a main factor in basic leadership. The work defined is useful to view the data as the quantity of tweets where sentiments or emotions are either good or bad, or neutral. John C and Jonathon R. specified an imperative sub-errand of emotions examination is classification of polarity, in which content is delegated being good or bad. Machine learning methods can play out this classification adequately. Be that as it may, it require a substantial corpus of preparing information, and various examinations have shown that the great execution of supervised models is reliant on a decent match between the preparation and testing information as for the area, subject, and time period. Pitifully supervised procedures utilize an extensive accumulation of unlabelled content to decide sentiment, thus their execution might be less subject to the area, subject.

III. PROPOSED METHOD

The paper proposed a method to classify sentiments of an individual using k-nn classifier. The steps taken are described below.

A. Data

For performing any classification, we need data set. Training and testing tweets are collected from Sentiment 140. We had 16, 00,000 training tweets and approx 448 testing tweets. The file was in excel format and converted in suitable format. We used MATLAB tool for our implementation. From the data, we took two main characteristics that were tweets itself and tweets labels (positive or negative).

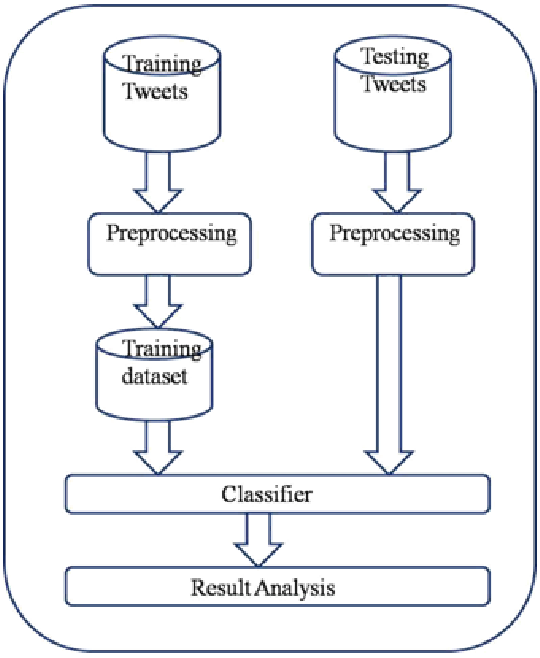

Figure.1. Block diagram of proposed method.

B. Pre-processing

Tweets collected are not in usable format. Thus basic data pre-processing steps need to be taken in order to avoid noisy values. User names, any links, hash tags, emoticons, repeated letters, were removed. The Pre-processed data did not contain any common words which do not posses any sentiment like the, about, are etc. Stemming is also done for obtaining basic word for classification. For example word national was stemmed to nation.

C. Working

After all pre-processing steps a corpus is made which has unique words alphabetically arranged. The model is trained first for polarity based classification and then for K-nn classifier. For K-nn classification, bag of word is maintained for each training and testing tweet. Then minimum Euclidian distance between training and testing bag of words is calculated. We set value of K as one-fourth of the data and choose the maximum occurred sentiment label from those one-fourth values.

IV. RESULT ANALYSIS

Implementation shows that K-nn classifier with bag of words approach works well in comparison with Polarity based classification.

Table.1. Accuracy of the Proposed Work

S.NO	METHOD	ACCURACY
1	Polarity Based Classification	77.5%
2	K-nn Classifier With Bag Of Words	79.3%

V. CONCLUSION

It can be observed from the results that k-nn classifier is a good choice for analysing sentiments. Experimental results show that k-nn classifier with bag of words feature selection outperforms over polarity based sentiment classification.

VI. REFERENCES

[1]. Alec Go, Richa Bhayani and Lei Huang. Twitter Sentiment Classification using Distant Supervision. CS224N Project Report, Stanford, pages 1-12. 2009.

[2]. Pang B, Lee L, Opinion mining and sentiment analysis in Found Trends Inform Retriev, 2 (2008), pp. 1135.

[3]. Hatzivassiloglou V, McKeown K, Predicting the Semantic Orientation of Adjectives.

[4]. B. OConnor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, from tweets to polls: Linking text sentiment to public opinion time series, in Proc. 4th Int. AAAI Conf. Weblogs Social Media, Washington, DC, USA, 2010.

[5]. J. Bollen, H. Mao, and X. Zeng, Twitter mood predicts the stock market, J. Computer Science., vol. 2, no. 1, pp. 18, Mar. 2011.

[6]. G. Mishne and N. Glance, Predicting movie sales from blogger sentiment, in Proc. AAAI-CAAW, Stanford, CA, USA, 2006.

[7]. Liu B, Sentiment analysis and opinion mining in Synth Lect Human Lang Technol (2012).

[8]. Ohana B., Tierney B., Sentiment Classification of Reviews Using SentiWordNet in 9th. IT and T Conference, Dublin Institute of Technology, 2009.

[9]. Pang B, Lillian L, Thumbs up? Sentiment Classification using Machine Learning Techniques in Proceedings of EMNLP 2002, pp. 7986.

[10]. P. Waila, Marisha, V. K. Singh, M. K. Singh, Evaluating Machine Learning and Unsupervised Semantic Orientation approaches for sentiment analysis of textual reviews Computational Intelligence and Computing Research (ICCIC), 2012 IEEE International Conference on 18-20 Dec. 2012

[11]. Duda, R.O. Hart, P.E. (1973), Pattern classification and scene analysis, Wiley, New York

[12]. McCallullum A and Nigam K, A Comparison of Event Models for Naive Bayes Text Classification 1998.

[13]. Nicholls C., Song F., Improving sentiment analysis with Part-of-Speech weighting in 2009 International Conference on Machine Learning and Cybernetics (Volume: 3), 2009.

[14]. Haddi E, Liu Xi, Shi Y, the Role of Text Pre-processing in Sentiment Analysis in ITQM2013, Science Direct, 2013.

[15]. Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe, Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.

[16]. Bing Liu, Sentiment Analysis and Opinion Mining.

[17]. Bruno Ohana, Brendan Tierney, Sentiment Classification of Reviews Using Senti- WordNet.

[18]. Pang B, Lee L, and S. Vaithyanathan Thumbs up? Sentiment Classification using Machine Learning Techniques.

[19]. Waila P., Marisha, Singh V.K., Singh M.K., Evaluating Machine Learning and Unsupervised Semantic Orientatio approaches for sentiment analysis of textual reviews.
.

[20]. Gautam G., Yadav P., Sentiment Analysis of Twitter Data using Machine Learning Approaches and Semantic Analysis.

[21]. Sun B., Ng V., Analysis Sentimental Influence of Posts on Social Network.

[22]. Bing L., Chan. K.C.C. , Public Sentiment Analysis in Twitter Data for Prediction of a Company’s Stock Price Movements.

[23] Ms. Devaki V. Ingule, Prof. Gyankamal J. Chhajed, Survey of Public Sentiment Interpretation on Twitter.