International Journal of Engineering & Technology, 7 (2.24) (2018) 20-23

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Sentiment Analysis using Logistic Regression and Effective Word Score Heuristic

Abhilasha Tyagi[1], Naresh Sharma[2]

[1]Dept. of Computer Science and Engineering, SRM University (India)-201204
[2]Assistant Professor, Dept. of CSE, SRM University (India)-201204
**Corresponding Author E-mail: abhanuty004@gmail.com*

Abstract

Sentiment Analysis is a method for judging somebody's sentiment or feeling with respect to a specific thing. It is utilized to recognize and arrange the sentiments communicated in writings. The web-based social networking sites like twitter draws in a huge number of clients that are online for imparting their insights in the form of tweets or comments. The tweets can be then classified into positive, negative, or neutral. In the proposed work, logistic regression classification is used as a classifier and unigram as a feature vector. For accuracy, k fold cross validation data mining technique is used. For choosing precise training sample, tweet subjectivity is utilized. The idea of Effective Word Score heuristic is likewise presented to find the polarity score of words that are frequently used. This additional heuristic can speed up the classification process of sentiments with standard machine learning approaches.

Keywords: *Sentiment; Unigrams; Polarity; Machine Laerning; Twitter.*

1. 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 reviews define summarized contemplation of author, tweets are additional easygoing and partial to 140 characters of content. For the most part, tweets are not as insight fully created as reviews. However, they still offer companies an extra avenue to accumulate feedback. Customers can utilize sentiment analysis to look into items or administrations before assembly a buy. Organizations can likewise utilize this to accumulate basic feedback about issues in recently launched items.

In previous researches like Pang et al. [1] movie reviews are examined over many classifiers. This work of Pang et al. [1] 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).

1.1 Sentiment

A sentiment is defined as a view or opinion that is expressed. It is a feeling of someone that he/she expresses either in textual or verbal form. A sentiment can be defined as a personal positive or negative feeling.

1.2 Sentiment Analysis

Sentiment analysis is defined as the extraction of information from a piece of text. Sentiment analysis means telling the state of mind of a speaker, essayist, or other subject regarding any fact or the general relevant extremity or enthusiastic answer to a report, communication. The attitude might be a result or assessment, full of feeling state, or the planned passionate statement. In the planned work, we have twitter as an area of sentiment analysis and tweets are arranged into positive, negative and neutral opinion or sentiment with the help of construct models.

Many machine learning approaches are used to classify someone's emotion or feeling. Different algorithms can be utilized to do sentiment classification.

2. Sentiment Analysis Techniques

The following techniques are the general approaches for analyzing sentiments:

- **Manuscript based**

Manuscript level sentiment analysis can use any machine learning approach (supervised or unsupervised) to analyze any sentiment. Another intriguing strategy decides the PMI of expressions as positive or negative just to register the polarity of an expression**.**

- **Sentence based**

Sentence based analysis divides the sentence into smaller similar phrases. This type of analysis is useful to deal with special type of sentences like conditional, negation sentences.

- **Aspect-based**

Aspect-based analysis of sentiment defines that a product has many aspects or features or properties which in result have differ sentiments.

- **Lexicon based sentiment analysis**

Lexicon based sentiment analysis is an attractive research area that uses methods such as WordNet distance to label the sentiment as fine and dire.

3.Basic Sentiment Analysis Tools

Various tools that can be used to track user sentiment are:

- **Google Alerts:** It is used to monitor any search related queries. You can record your content and get usual updates via email. It is very simple and useful to track competitors and influencers.
- **People Browser:** It is used to find the mentions of any brand, organization and opponent and examine sentiment. To compare the quantity of mentions throughout the marketing campaigns, people browser is used.
- **Google Analytics:** It is a controlling tool for discovering which channels influenced the subscribers and buyers. Create custom reports, annotations to maintain uninterrupted records of the marketing and web design actions, as well as higher segments to breakdown visitor data and gain valuable insights on their online experiences.
- **Tweet statics:** To graph the stats of Twitter this tool is used. It is a freely available tool. You just need to enter any Twitter handle for the resulted graph
- **Pagelever:** It is a tool that is used to measure activity of twitter. It is used to measure the accuracy of content sharing and consuming on the platform like Facebook.
- **Social Mention:** This tool is useful for recording the mentions for some identified keywords in any blog, video, audio, event or hashtags. It also defines whether the mention is positive, negative, or neutral.
- **Marketing Grader:** It is used to analyze the aspects of your overall efforts in marketing. It let you know where you succeed and where to improve. It also helps in understanding mobile marketing, competitive benchmarking and overall analysis.

4. Classifiers

Following basic classifiers are considered.

1) Naive Bayes: It is a family of algorithm that is used to construct classifier that assigns labels to the instances of problems.

2) Support Vector Machines (SVM): SVM is used for classification purpose. It can also be used in regression analysis and outlier detection. It constructs a hyper-lane in high dimensionality space.

3) Logistic Regression: It is used to determine the output or result when there are one or more than one independent variables. The output value can be in form of 0 or 1 i.e. in binary form.

5. 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 [2].

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 [3].

Bing Liu. 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 [8].

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 [10]

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. [13].

Lo Y.W, 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 favorable 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 [16]

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 [19].

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 [20].

Hassan S. 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 [21].

6. Proposed Work

Before implementing the research work some concepts should be well known. These concepts are defined below:

- **Subjectivity filtering:** Subjectivity filtering is used to gain accurate result when the training dataset is small. To classify a tweet, TextBlob is used which classify the tweets into subjective or objective. The tweets which have score less than a specific threshold are then removed. After this filtering, the remaining tweets are used to train datasets. As subjectivity threshold increases, number of tweets will filtered
- **Preprocessing:** Preprocessing is done to remove the data which is not useful to reduce the feature space. In most of the researches, preprocessing steps are same but here some more steps are considered.

1) Basic steps: The emoticons in any comment or tweet are stripped of. Users often include twitter usernames in their tweets in order to direct their messages. Usernames (e.g. @Chinmay) and URLs present in tweets are also stripped of because they do not help in sentiment classification. Apart from full stops, which are dealt in the next point, other punctuations and special symbols are also removed. Repeated whitespaces are replaced with a single space. Stemming is also performed in order to reduce the size of the feature space.

2) Full Stops: Full stops in the model are usually replaced by a space. However, it is observed that casual language in tweets is often seen in form of repeated punctuations. For example, this is so cool...wow. This format is taken into consideration, and two or more occurrences of (.) and (-) is replaced with a space. Also, full stops are also quite different in usage. Sometimes, there is not any space in between sentences. For example, "It's raining…Feeling awesome". A single event of a full stop is replaced with a space.

3) Removing Hash-tags: Hash-tag in a tweet is used to emphasize a particular word or sentence, for example #thisisgood. Removal of these hash-tags is important because these hash-tags do not define any sentiment. Thus pre-processing is done and hash-tags before any word are removed.

4) Repeated letters: A tweet can be written in very informal way. For example, hello can be written as helloo or hellooo in a tweet. Pre-processing is done to remove such repetitions.

6.1 Effective Score of a Word Approach

EFWS is a heuristic approach for which the training dataset is filtered on the basis of subjectivity threshold. Due to the additional preprocessing steps in proposed work, the resulted work will be more robust. In proposed work, a list of frequent words is maintained with dictionary of stop words. The polarity score will also maintained with the score ranges from -5 to 5. The dictionary has approximately 2500 words. A fact is assumed here that no two synonyms have same polarity score. For example, a word having polarity score 3 cannot have a synonym word with polarity score5.

7. Conclusion

As it is observed these days, that many individuals' posts surveys with respect to any item, movie, game or occasion via web-based networking media stages. For this, it is essential for the organizations to define particular sentiments of such surveys keeping in mind the end goal to realize that what individuals think about the item. The projected method utilizes one such stage called twitter to play out the sentiment categorization. The info is taken as tweets in the wake of verifying the client. The current framework has utilized Naive Bayes and Support Vector Machine classifiers to classify the sentiments. The classifiers utilized as a part of proposed framework are Support Vector Machine, Naïve Bayes. For highlighting features Logistic Regression is used and for defining labels Effective score of a word is utilized.

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, 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.

[24] Zhao L., Ren Y., Wang J., Meng L., and Zou C., Research on the opinion mining system for massive social media data.

[25] Lo, Y.W, Potdar, V, A review of opinion mining and sentiment classification framework in social network.

[26] Peter D. Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.

[27] Amit Goyal and Hal Daume III , Generating Semantic Orientation Lexicon using Large Data and Thesaurus.

[28] Jonathon Read, John Carroll, Weakly Supervised Techniques for Domain- Independent Sentiment Classification.

[29] Emma Haddi, Xiaohui Liu, Yong Shi, The Role of Text Pre-processing in Sentiment Analysis.

[30] Hassan Saif, Yulan He and Harith Alani, Semantic Sentiment Analysis of Twitter.

[31] Efthymios Kouloumpis, Theresa Wilson, Johanna Moore, Twitter Sentiment Analysis: The Good the Bad and the OMG!

[32] Oaindrila Das, Rakesh Chandra Balabantaray, Sentiment Analysis of Movie Reviews using POS tags and Term Frequencies.

[33] http://dataminingwarehousing.blogspot.in/2008/10/data-mining-steps-of-datamining.html

[34] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[35] http://www.thearling.com/text/dmwhite/dmwhite.htm

[36] https://blog.semanticlab.net/2013/09/10/techniques-and-applications-forsentiment-analysis/

[37] R. Narayanan, B. Liu, and A. Choudhary, Sentiment analysis of conditional sentences, in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 Volume 1, Stroudsburg, PA, USA, 2009, pp. 180189.

[38] D. Davidov, O. Tsur, and A. Rappoport, Semi-supervised recognition of sarcastic sentences in Twitter and Amazon, in Proceedings of the Fourteenth Conference on Computational Natural Language Learning, Stroudsburg, PA, USA, 2010, pp. 107116.

[39] N. Jindal and B. Liu, Identifying comparative sentences in text documents, in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 2006, pp. 244251.

[40] https://www.iprospect.com/en/ca/blog/10-sentiment-analysis-tools-track-socialmarketing-

[41] S.V.Manikanthan and D.Sugandhi " Interference Alignment Techniques For Mimo Multicell Based On Relay Interference Broadcast Channel " International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume- 7 ,Issue 1 –MARCH 2014.

[42] T. Padmapriya and V. Saminadan, "Distributed Load Balancing for Multiuser Multi-class Traffic in MIMO LTE-Advanced Networks", Research Journal of Applied Sciences, Engineering and Technology (RJASET) - Maxwell Scientific Organization , ISSN: 2040-7459; e-ISSN: 2040-7467, vol.12, no.8, pp:813-822, April 2016.