Twitter Sentiment Analysis Using Binary Classification Technique

B.N. Supriya[1], Vish Kallimani[2(✉)], S. Prakash[3], and C.B. Akki[1]

[1] Department of ISE, SJBIT, Bangalore 560060, India
[2] UTP, Seri Iskandar, Malaysia
knowvp@gmail.com
[3] Department of CSE, Dayanand Sagar University, Bangalore 560078, India

Abstract. The popularity of World Wide Web has brought a latest way of expressing the sentiments of individuals. Millions of users express their sentiments on Twitter, making it a precious platform for analyzing the public sentiment. This paper proposes a 3-step algorithm for sentiment analysis. Cleaning, Entity identification, and Classification are the 3 steps. Finally we measure the performance of the classifier using recall, precision and accuracy.

Keywords: Sentiment analysis · Recall · Precision · Heterogeneous Architecture Research Platform (HARP)

1 Introduction

During the previous years, the web has become a huge source of user-generated content. When we searched for keyword "social network" on Google, we get a hit of about 617000000 in just 0.37 s as on 09/06/2015 at 12:15 pm which speaks about its popularity.

A social network is a structure, made up of a specific group of individuals or organizations, that allow users to come closer, communicate and share information. Jacob Moreno is credited with developing the first sociograms in 1930 s to study interpersonal relationships. SixDegrees launched in 1997 was the first recognizable social network site [1]. Further, many sites like Asian Avenue, Black Planet, Ryze, Friendster, MySpace, Hi5, YouTube, Facebook, supporting various combinations of profiles publicly articulated friends were launched [1].

In contemporary, explosive growth of online social media microblogs have become a quick and easy online information sharing platform. These platforms generate rich and timely information (reviews, comments, ratings, feedbacks etc.) that requires informational filtering down to successful relevant topics and events. Twitter launched in 2009, is one such extremely popular micro blogging site where millions of users express themselves, give opinion and gets feedback via a short text message called tweets. Over 240 million tweets are being generated by the tweeter per day. This available voluminous data is being used for making decisions for enhancing profitability or for purpose. Processing of such large amount of heterogeneous data in an

P.C. Vinh et al. (Eds.): ICTCC 2016, LNICST 168, pp. 391–396, 2016.
DOI: 10.1007/978-3-319-46909-6_36

effort to uncover hidden pattern, unknown correlations gave birth to a field called as sentiment analysis.

Sentiment analysis is one of the newest research areas in computer science. Sentiment analysis is a natural language processing technique to extract polarity and subjectivity from semantic orientation which refers to the strength of words and polarity text or phrases [2]. The sentiments can be categorized into positive and negative words [7, 8]. There are two main approaches for extracting sentiment automatically which are the lexicon-based approach and machine-learning based approach [2–6].

2 Related Work

Contemporary Sarlan et al. [9] describes the design of a sentiment analysis, extracting a vast amount of tweets. This paper explains the different approaches, techniques available for sentiment analysis and also focuses on the application requirements, functionalities of developing the twitter sentiment analysis application.

Alec Go et al. [10] introduces a novel approach for classifying the sentiments of twitter message using distant supervision. In this approach the data which consists of tweets with emotions are used as noisy labels. This paper also describes the accuracy of the different machine learning algorithms.

Jiguang Liang et al. [11] introduces a sentiment classification method called AS_LDA which assumes that words in subjective documents are of two parts sentiment element words and auxiliary words. These words are further sampled according to the topics. Gibbs sampling is used to verify the performance of this model.

Harshil T. Kanakia et al. [12] proposes a method called as Twitilizer to perform the classification of tweets into positive and negative tweets using Naive Bayes classifier. This method collects the tweets from the twitter and stores it into the persistent medium. The tweets are further pre-processed to the required format. The features are then extracted by removing the stop words and punctuations from the tweet. The sentiment property returns the polarity and subjectivity of the tweets. The displays of top positive and negative sentiments are obtained by the ranking algorithm. The results are tabulated and shown statistically.

Vadim kagsn et al. [13] uses the sentiment analysis technique for forecasting the 2013 Pakistan and 2014 Indian elections. This technique was used to predict who the prime minister would be. They considered 3 leading candidates from the politics in the dataset. The datasets were collected from Resselaer Polytechnic Institute and a Twitter Indian Election Network tool was built. Later the data was analyzed using the diffusion estimation model. The estimated results were found to be correct when compared with the actual election results. This paper also concludes that twitter can be a good measure for public sentiment on election related issues.

Li Bing et al. [14] analyzed that lot of work still needs to be done on summarization and analysis techniques for social data. Motivated by this problem the authors proposed a matrix-based fuzzy algorithm called FMM system, to mine the twitter data. The paper concludes that the problem of Li Bing et al. [14] analyzed that lot of work still needs to be done on summarization and analysis techniques for social data. Motivated by this problem the authors proposed a matrix-based fuzzy algorithm called FMM system,

to mine the twitter data. The paper concludes that the problem of handling big data for data mining can be solved using FMM algorithm that adapts map reduce framework and also the speed of the execution can be increased significantly. The author has intensively worked on stock price movement through FMM system and his results are with high prediction accuracy.

Farhan Hassan khan et al. [15] presents a novel algorithm which classifies twits into positive, negative and neutral feelings. This paper gives us a distinct method for pre-processing when compared with [12]. In this algorithm the collected twitter streaming APIs are refined using different pre-processing task that removes slangs and abbreviations. The classification of the tweets is done by using the techniques such as emoticon analysis, Bag of words and SentiWordNet. The result of the algorithm shows the increase in the accuracy, precision, recall, f-measure when compared with other techniques. The framework is also further enhanced by using visual analysis of classified tweets location that is drawn on the map.

Many researchers have proposed different algorithm in recent years in order to classify the sentiments in the tweets. These researchers focused on the text mining in order to determine the sentiments. The study of existing research shows that the sentiment analysis results are not convincing.

3 Sentiment Analysis on Social Media Text

With the exponential growth of many social media such as Facebook, Flickr, Google+ etc., twitter has emerged as one of the most influential online social media service. These multiple media social networks generate a variety of data like text, video or visual etc. Our research mainly concentrates on the sentiment analysis of text only. For analysing textual sentiments, the following 3-step algorithm is being discussed here

The proposed 3-step algorithm consists of following steps:

Step 1: Input the tweet into HARP.

Step 2: Data Cleaning

The raw data that has been collected from the source will be cleaned as below:

- Check for the terms and their frequencies: The tweets may contain a single word or word n-grams. Their frequencies or presence are checked.
- Retweets and Duplicated tweets are removed
- Usernames if included should be integrated with the symbol @ before the name. Further Hash-tags and user names must be identified.
- Users often include the URL's in their tweets. This URL's must be converted to equivalent classes.
- All the text present in the tweets must be converted to a single format (Higher to lower case or vice versa).
- Spell check must be done and must be corrected if there is any.
- Stop words are deleted from the tweets.
- Slangs and abbreviations are corrected.

Step 3: Entity Identification

The cleaned data obtained from step 2 can contain a lot of distinctive properties like actors, target objects, nouns, adjectives, kinds of sentences etc. These properties are termed as entities. The identification of these entities is the task of step 3. This can be achieved as given below:

- The actors, target objects can be identified by named entity recognition and relation extraction method. The emotions can be identified by the database created manually.
- A speech tagger can be applied to a tweet which identifies the kinds of sentences.
- A part of speech tagging must be applied to a tweet that identifies the noun, verb, adjectives, adverb, interjection, conjunction, pronoun, pre-position.

Step 4: Classification

This step takes the input from the entity identification and classifies the words as positive or negative. Once the topic classification has been done, we can apply the binary classification for given data, where positive word (pw) can be initialized to 1, and negative words (nw) can be initialized to 0.

Step 5: Calculate the sentiments

Once the words are classified, the number of positive and the negative words are counted. We then check the result set (rs) by using

Rs = (tpw – tnw) / total no of words (tw).

Where tpw = total positive words,

tnw = total negative words.

$$rs = \begin{cases} Positive\ tweet,\ rs \geq 0 \\ Negative\ tweet,\ rs < 0 \end{cases}$$

The Fig. 1 shows the flow chart for the proposed algorithm

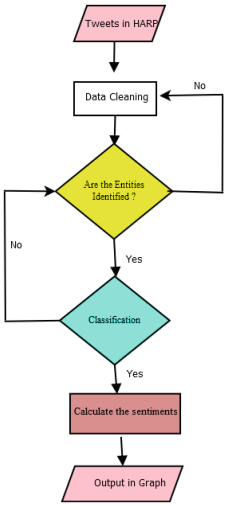

Fig. 1. Proposed flowchart

The performance of the proposed algorithm can be evaluated using accuracy, precision, recall measures.

Accuracy is defined as the ratio between correctly classified tweets from proposed classifier and manually labelled tweets. The formula for the same is

$$\text{Accuracy} = (\text{tn} + \text{tp}) / (\text{tn} + \text{fp} + \text{fn} + \text{tp})$$

Where tn = true negative, tp = true positive, fp = false positive, fn = false negative.

Precision is defined as the ratio between true positive (tp) and both true positive (tp) and false positive (fp). The formula for the same is

$$\text{Precision} = \text{tp} / (\text{tp} + \text{fp})$$

Recall is defined as the ratio between true positive and both true positive and false negative (fn). The formula for the same is

$$\text{Recall} = \text{tp} / (\text{tp} + \text{fn}).$$

4 Conclusion

In this paper, we propose a 3-step algorithm for efficient sentiment analysis. The first step cleans the data, the second step identifies the different entities and the last step uses the binary classifier to classify the tweets as positive or negative. The result can be shown through the graphs and the performance can be calculated using recall, precision and accuracy. The proposed method is simple and we expect to get better performance when compared to the other methods available in the literature however the experiment is yet to be performed.

References

1. Boyd, D.M., Ellison, N.B.: Social network sites: definition, history, and scholarship. J. Comput.-Mediat. Commun. **13**, 210–230 (2008)
2. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis. Association for Computational Linguistics (2011)
3. Annett, M., Kondrak, G.: A comparison of sentiment analysis techniques: polarizing movie blogs. In: Conference on Web Search and Web Data Mining (WSDM). Department of Computing Science, University of Alberia (2009)
4. Goncalves, P., Benevenuto, F., Araujo, M., Cha, M.: Comparing and Combining Sentiment Analysis Methods (2013)
5. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: the good the bad and the OMG!. In: International AAAI, vol. 5 (2011)
6. Sharma, S.: Application of support vector machines for damage detection in structure. J. Mach. Learn. Res. (2008)

7. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of twitter. In: Proceeding of the Workshop on Information Extraction and Entity Analytics on Social Media Data. Knowledge Media Institute, United Kingdom (2011)
8. Prabowo, R., Thelwall, M.: Sentiment analysis: a combined approach. In: International World Wide Web Conference Committee (IW3C2). University of Wolverhampton, United Kingdom (2009)
9. Sarlan, A., Nadam, C., Basri, S.: Twitter sentiment analysis. In: International Conference on Information Technology and Multimedia (ICIMU), 18–20 November 2014
10. Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification using distant Supervision (2009)
11. Liang, J., Liu, P., Tan, J., Bai, S.: Sentiment classification based on AS-LDA model. Inf. Technol. Quant. Manag., Proc. Comput. Sci. **31**, 511–551 (2014). doi:10.1016/j.procs.2014.05.296
12. Kanakia, H.T., Kalbande, D.R.: Twitilyzer: designing an approach for ad-hoc search engine. In: International Conference on Communication, Information & Computing Technology (ICCICT), 16–17 January 2015
13. Kagan, V., Stevens, A., Subrahmanian, V.S.: Using twitter sentiment to forecast the 2013 Pakistani Election and the 2014 Indian Election. IEEE Intell. Syst. (2015)
14. Li, B., Chan, K.C.C.: A fuzzy logic approach for opinion mining on large scale twitter data. In: IEEE/ACM 7th International Conference on Utility and Cloud Computing (2014). doi:978-1-4799-7881-6/14
15. Khan, F.H., Qamar, U., Javed, M.Y.: SentiView: a visual sentiment analysis framework. In: 2014 IEEE International Conference on Information Society (i-Society 2014). doi:978-1-908320-38/4