Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 10, October -2017

A Robust Approach to Analyse Sentiments for Trending Hashtags

Ms. Priyanka Nandal

Maharaja Surajmal Institute of Technology, GGSIPU, Delhi, India

Abstract—*Interest in opinion mining or sentiment analysis has developed exponentially in the recent years. Opinion mining or sentiment analysis can be defined as the computational study of people's opinions, sentiments, attitudes, and emotions present in written form. Billions of opinions are shared on social websites every day. Therefore these sites serve as ample data source for sentiment analysis or opinion mining. In the present work twitter is used as the data source for sentiment analysis. English language is used for the analysis as it is easier to use this language as compared to other languages.*

Keywords— *Bigram, Naïve Bayes, Opinion Mining, Sentiment Analysis, Unigram*

I. INTRODUCTION

Internet users today employ microblogging as a very popular communication tool. The opinion of others greatly influences the decisions of human beings. The few of the examples of decisions which affect various aspect of the life include buying a smartphone, money investment, selection of school, etc. In the Internet era, the opinions of many people are obtained by visiting the sites like review sites (e.g. Citysearch, Consumer Reports), e-commerce sites (e.g., Flipkart, Infibeam), online opinion sites (e.g., Kirkus Reviews, IMDb, Yelp) and social media sites (e.g., Facebook, Twitter). The feedback can be collected for a particular product or a service. Organizations use opinion polls, social media and surveys as a technique for collecting feedback on their products and services. Sentiment analysis (also known as opinion mining) is the computational study of people's sentiments, opinions, attitudes and emotions present in written form. Sentiment analysis or opinion mining is the use of natural language processing, text analysis and computational linguistics to recognize and take out subjective information in original documents. Using sentiment analysis as a tool any organization can collect a large amount of information such as feedback of their product and use it for producing better products, comparison with the competitors and for advertisement purposes.

Billions of messages daily appear on the popular web sites such as Twitter, Facebook, Tumblr. These messages are in free format. A microblog is typically smaller than a traditional blog and may consist of short sentences, individual images or video links. Twitter produces immense amount of data everyday which is called as the twitter fire hose [1]. A large number of features are introduced to model and analyse such huge datasets. Sentiment analysis has been performed using twitter data [2]-[4]. The primary focus of the early research on sentiment analysis of Twitter data was on product or movie reviews at the document level [5], [6]. However, now the twitter data has been handled at phrase level [7]-[9].

To work with the enormous amount of data produced by twitter every day is a tedious task. Therefore researchers limit the amount of data in their analysis. A training set of only 216 records was employed by Pak and Paroubek [10] after analysing 300,000 records. Jansen et al. [11] analysed 149,472 tweets, but they employed a third party service for determination of the sentiments. The third party which they employed i.e. "Summize" use only 125 recent tweets for sentiment determination. They also preformed a study to determine whether Summize service is accurate. For this purpose they manually coded 2610 tweets and found that there was no significant difference in sentiment distribution between the first 125 tweets and the next 125 tweets. Go et al. [12] also manually coded 359 tweets of the few specified weeks. Various approaches have been developed to monitor real-time twitter data to see the reaction of users to events, news stories and major events [13]-[20]. There are two broad classes of techniques applied for the sentiment analysis. In the first class a sentiment lexicon is applied in terms of negative or positive related opinion for the evaluation of the text in unsupervised fashion [21]. In the second class approach, the relationship amid features of the text segment and the opinions is derived by coupling the textual feature representation with machine learning algorithms [22].

A large training set is required for models using supervised methods [23]. This training set includes labels which are assigned to training instances manually. Two major issues are examined in sentiment analysis. One issue is the features used for representing the writing and second is the techniques used for sentiment analysis [22], [23]. In the literature, many writing features have been tested such as syntactic [23], semantic [21] and stylistic features [22].

The tweets ending in positive emoticons like ":)" are treated as positive and negative emoticons like ":(" as negative. Models are built using Naive Bayes, Support Vector Machines (SVM), etc. It is reported that SVM is more preferred than as compared to other classifiers. In terms of feature space, these models are studied in combination with parts-of-speech (POS) features. Pak and Paroubek collect data following a similar epitome [10]. A different classification was performed by them which imply an approach that compares subjective data with the objective one. Subjective data in their study refers to the tweets ending with emoticons in the same manner as Go et al. [12]. Similarly objective data refers to the data (eg: twitter accounts) scraped from the internet or through web pages. They report that POS works efficiently with the Bigram Models (contrary to results presented by Go et al. [12]). Moreover, the data they use for training and testing is biased. In contrast, the author here present features that achieve an advantage over these limitations .This was mainly

@IJAERD-2017, All rights Reserved

possible because the author report results on manually commented data that does not suffer from biases. The data used in the present work is a random sample. The size of the dataset allows the author to perform validation experiments and various statistical operations. Barbosa and Feng [24] used polarity predictions from three websites to train a model and used 1000 tweets for training and other different 1000 tweets for testing. They proposed to employ syntax features of tweets like hashtags, exclamation marks and punctuation with features like POS of words. The author has implemented their approach by combining prior polarity with POS. The obtained results show that on increasing the efficiency of features for the classifier used, increases the polarity of words with POS and hence an improved accuracy. Gamon performed sentiment classification for noisy data [25]. One aim of their paper is to observe the role of POS tags. They show that feature selection and analysis contribute to increased accuracy. In the present paper the author perform feature analysis and show that the use of only 200 abstract linguistic features performs well with equivalent results as obtained from hard unigram models. The aim of the mentioned work is to get sentiment polarity for the text.

Plan of paper:

This paper is organized as follows. Section II describes the research methodology used. Section III illustrates the results and discussion. Section IV consists of the conclusion and a direction for future work.

II. RESEARCH METHODOLOGY

The flow of varied tasks for evaluating the sentiments of social media dataset is explained in this section. The first step is collecting the dataset for training. The training data collected from varied sources is subjected to pre-processing for eliminating features that do not contribute to polarity detection. Next test data is classified by feeding this training data into sentiment analysis engine. In the next step, the data is fetched from the social media using the input query for which polarity is to be detected. The sentiment analysis engine contains Naive Bayes classification algorithm that consults training data for calculating the probabilities and predicting the sentiment of the proposed query term.

A. Data Extraction

The data is extracted from social media like Twitter using Twitter API on the basis of the input query term. The twitterR package of R is used in this work which provides an interface to the Twitter web API. The tweets are extracted for three hashtags- #india strikes back, #trump, #noteban. The retrieved data is subjected to preprocessing to be used as test data to analyse sentiments using classification algorithm.

B. Pre-procesing

In pre-processing the part that does not contribute significantly to the polarity detection is eliminated. The messages of limited size typically less than 140 characters are called tweets [26]. The generic token USERNAME is replaced for tweets which consist of usernames of account holder (@nirajp). Similarly the generic token URL is replaced in place of links (http://goo.gl/nirajp). Further pre-processing of tweets is done to convert tweets to lower case characters to remove unevenness. '#' symbol is eliminated which is used to denote hash tags and the succeeding hash tag word is retrieved. Further eliminate the stop words (like the, is, a) which do not aid necessary in polarity detection. Eliminate the additional white spaces, punctuation marks and two or more repetitive letters in a word. For example, Happy represented as haaappy or haaaaaaappy to stress emotion on social media platform is converted to 'happy'. For simplicity, ensure that words are started with an alphabet. Eliminate all the words which do not start with an alphabet to reduce feature e.g. 21st, 7:30 pm.

C. Feature Extraction

For sentiment analysis feature extraction is considered as a very basic and crucial task. Improving feature extraction can often have a significant positive impact on classifier accuracy.

D. N-GRAMS

Using n-grams subsequence of n items can be defined from a given sequence. Applications of n-grams can be found in genetic sequence analysis, natural language processing, etc. N-grams method explains how to find a set of n-gram words from a given document. Unigrams (n=1), bigrams (n=2) and trigrams (n=3) are the commonly used models. Nevertheless, the value of n can extend for higher level grams.
The following example can be used to explain the n-gram model:
Text: "Its water is so transparent."
Unigrams: "Its", "water", "is", "so", "transparent".
Bigrams: "Its water", "water is", "is so", "so transparent".
Trigrams: "Its water is", "water is so", "is so transparent".

The simplest model is represented by unigrams for the n-gram approach. All the individual words present in the text constitute it. A pair of adjacent words is defined by the bigram model. A single bigram is formed from a pair of words. Similarly, n adjacent words can be extracted to form the higher order grams. The efficiency of higher order n-grams is

@IJAERD-2017, All rights Reserved

more as compared to the lower order n-grams as they can better capture the context to understand the position of the word.

E. Classification

As human beings learn from their past experiences machine learning algorithms also make decisions based on the past acquired knowledge. The applications of machine learning algorithms are in document classification and artificial intelligence. Machine learning algorithms can be used in classification. The two main steps involved are:

1. The training dataset is used to train the model.
2. The trained model is applied to the test dataset.

Any existing supervised classification method can be applied to sentiment analysis as it is a text classification problem. Naive Bayes classifier is used here for the classification of tweets. The obtained results are then compared using this model.

F. Naive Bayes

Naive Bayes classifier is based on the Bayes theorem. It is a simple probabilistic classifier. In this classification technique it is assumed that each feature in the document is independent of any other feature in the document. Naïve Bayes classifier considers a document as a collection of words and assumes that the probability of a word in the document is independent of its position in the document and the presence of other word.

The equation below shows the multinomial Naive Bayes model

$$P(c \mid d) := \frac{\left(P(c) \sum_{i=1}^{m} P(f \mid c)^{ni(d)} \right)}{P(d)} \quad (1)$$

In the above formula, f symbolizes a feature. The total features are represented by m. The count of feature fi in tweet d is represented by ni(d). Parameters P(c) and P(f—c) are calculated by maximum likelihood evaluation. For unseen features add -1 smoothing is utilized.

III. Results and Discussion

The dataset used had 100000 tweets. A ratio of 60/40 was used for training and testing. The samples were distributed randomly. Therefore the training data consisted of 60,000 tweets both positive and negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Here FN stands for False Negative, FP stands for False Positive, TN stands for True Negative and TP stands for True Positive. Using this technique the sample training data is partitioned into complementary subsets, performing the analysis on one subset (called the training set). The analysis is validated on the other subset (called the validation set or testing set). Training data was used to train the classifier. For calculating accuracy, some data was kept aside and passed to the classifier for classification. To find the accuracy testing parameters, the labels generated are compared to the actual labels of the dataset.

Table I shows the comparison of different features with their accuracy in percentage using the n grams method. The total numbers of positive and negative tweets are shown for unigram, bi-gram and trigram methods. The corresponding accuracy in percentage is also shown.

The author tested with three features. The author finally used unigram with the highest performance for sentiment analysis on three twitter hashtags - #indiastrikesback, #trump, #noteban.

TABLE I. Comparison of different features

Feature	Accuracy	Positive	Negative
Unigram	81.25	21475	18525
Bi-grams	40.69	16744	23256
Tri-grams	67.50	12452	27584

TABLE III. Analysis of trending Hashtags

Hashtag	Positive	Negative
#noteban	1221	3205
#india strikes back	3152	2951
#trump	1463	2028

@IJAERD-2017, All rights Reserved

IV. CONCLUSIONS AND FUTURE WORK

The accuracy of the results produced is highest using unigram feature extraction for negative and positive sentiment by the engine. Machine Learning is an emerging field as new and enhanced algorithms are being proposed. The latest algorithms can be used for the further work in connection to space, speed and accuracy parameters. Few of the interesting works which can be performed according to the author are:

A. Adding other datasets

The author has worked only with twitter media dataset. Other social media datasets can also be utilized like Facebook, Google+, LinkedIn, etc.

B. Context Classification

Every statement cannot be easily analysed as negative or positive. For example, the word ”kick” can behave as positive as well as negative as can be seen from the statements below. ”I love kicking ball” denotes positive, whereas ”I got kicked today” denotes negative sense. Likewise, things like sarcasm can’t be detected and lead to false classification due to the structure of the language.

C. Language Options

Other languages like Hindi, Spanish, Russian, etc. can also be included in the future work. The author has worked only with the English language in the present work.

D. Real-time Systems

The proposed current system does not perform sentiment analysis for tweets in real-time as new trends continuously develop on hashtag. Currently the system tests a static data set.

REFERENCES

[1] M. Ghiassi, J. Skinner and D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network," *Expert Systems with applications*, vol. 40, pp. 6266-6282, Nov. 2013.

[2] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment analysis of twitter data," In *Proceedings of the workshop on languages in social media*, Association for Computational Linguistics, pp. 30-38, Jun. 2011.

[3] D. Zimbra, M. Ghiassi and S. Lee, "Brand-related Twitter sentiment analysis using feature engineering and the dynamic architecture for artificial neural networks," In *System Sciences (HICSS), 49th Hawaii International Conference on*, IEEE, pp. 1930-1938, Jan. 2016.

[4] S. Rosenthal, N. Farra and P. Nakov, "SemEval-2017 task 4: Sentiment analysis in Twitter," In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 502-518, 2017.

[5] M. Hu and B. Liu, "Mining and summarizing customer reviews," In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177, Aug. 2004.

[6] L. Zhuang, F. Jing, and X. Y. Zhu, "Movie review mining and summarization," In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 43-50, Nov. 2006.

[7] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp. 347–354., Oct. 2005.

[8] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis," *Computational Linguistics*, vol. 35, pp.399-433, Sep. 2009.

[9] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, Association for Computational Linguistics, vol. 10, pp. 129–136, Jul. 2003.

[10] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," In *LREc*, vol. 10, May 2010.

[11] B.J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *Journal of the Association for Information Science and Technology*, vol. 60, pp.2169-2188, Nov. 2009.

[12] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, Dec. 2009.

[13] J. Benhardus, and J. Kalita, "Streaming trend detection in twitter," *International Journal of Web Based Communities, vol.* 9, pp.122-139, Jan. 2013.

[14] A. Bifet, and E. Frank, "Sentiment knowledge discovery in Twitter streaming Data," In *International conference on discovery science*, Springer, Berlin, pp. 1-15, Oct. 2010.

[15] C. Hsieh, C. Moghbel, J. Fang and J. Cho, "Experts vs. the crowd: examining popular news prediction performance on Twitter," In *Proceedings of ACM KDD conference*, Chicago, USA, May 2013.

[16] M. Mathioudakis, and N. Koudas, "TwitterMonitor: Trend detection over the twitter stream." In *Proceeding of ACM SIGMOD International Conference on Management of data*, pp. 1155–1158, Jun. 2010.

[17] N. Naveed, T. Gottron, J. Kunegis, and A. Alhadi, "Bad news travel fast: A content-based analysis of interestingness on Twitter," In *Proceedings of the 3rd International Web Science Conference*, ACM, p. 8, Jun. 2011.

@IJAERD-2017, All rights Reserved

[18] S. Petrović, M. Osborne and V. Lavrenko, "Streaming first story detection with application to twitter." In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 181-189, Jun. 2010.
[19] O. Phelan, K. McCarthy, and B. Smyth, "Using Twitter to recommend real-time topical news," In *Proceedings of the third ACM conference on Recommender systems*, pp. 385–388, Oct. 2009.
[20] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment in Twitter events. Journal of the American Society for Information Science and Technology," vol. 62, pp. 406–418, Feb. 2011.
[21] P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews." In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 417–424, Jul. 2002.
[22] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums," *ACM Transactions on Information Systems*, vol. 26, pp.12, Jun. 2008.
[23] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," In *Proceedings of the conference on empirical methods in natural language processing*, vol. 10, pp. 79–86, Jul. 2002.
[24] L. Barbosa, and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," In *Proceedings of the 23rd international conference on computational linguistics*, Association for Computational Linguistics. pp. 36–44, Aug. 2010.
[25] M. Gamon, "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," In *Proceedings of the 20th international conference on Computational Linguistics*, p. 841, Aug. 2004.
[26] E. Martínez-Cámara, M. T. Martín-Valdivia, M.T., L. A. Urena-López, and A. R. Montejo-Ráez, "Sentiment analysis in Twitter," *Natural Language Engineering*, vol. 20, pp.1-28, Jan. 2014.

@IJAERD-2017, All rights Reserved