IJETST- Vol.||02||Issue||09||Pages 3221-3225||September||ISSN 2348-9480 2015

Open access Journal International Journal of Emerging Trends in Science and Technology

DOI: http://dx.doi.org/10.18535/ijetst/v2i9.14

A Machine Learning based Approach to Detect Sentiment in Twitter Data

Author
Vivek Kumar Singh
Department of Computer science
Banaras Hindu University, Varanasi-221005, India

ABSTRACT

This paper presents a machine learning based algorithmic approach to detect sentiment in Tweets posted by users on microblogging site Twitter. The experimental framework is based on use of a Naïve Bayes classifier. First of all, the standard Naïve Bayes classifier is implemented in R language and tested on two publicly available datasets comprising of sentiment labeled tweets. Then the standard Naïve Bayes classifier is modified to design a Lexicon-pooled hybrid classifier which incorporates knowledge from sentiment lexicon as well. The designs are evaluated for two feature selection schemes: tf and tf.idf. The accuracy of the different implementations is calculated and plotted diagrammatically. The proposed approach is a good and robust approach for detecting sentiment in tweets posted by users.

KEYWORDS: *Affect Analysis, Opinion Mining, Sentiment Analysis.*

1. INTRODUCTION

Internet is most popular medium of sharing information. Social media platforms have transformed the web to a more reactive, responsive and more expressiveglobal stage. Thephenomenalgrowth in web based technologies is allowing common man to express his view on international stage.There are various types of platforms available on web where user shares and expresses his views such as personal web sites, social networking sites, blogging sites and microblogging sites etc. Each of the above said sites are a special class of web platform. Personal web site allows user to develop, maintain and write on website in own style whereas blogging and microblogging sites provide a common platform for expression.

Microblogging sites are even special where a user is restricted to write a post in fix number of words, for example twitter allows a user to write a tweet in 140 characters. Twitter is the most popular microblogging site with millions of users having account. An individual twitter post is usually a narrow domain text. Thescale and potential of twitter can be easily imagined by number of tweets generated per day andit is more than 500 million. Information was never flowing at such pace as it is flowing in current times. In such a wavy pattern of information, it is difficult to encounter a wavelet of relevant information at right time. Even going through every tweet for analyzingsuch important information is not less than nightmare. Information sharing itself is hiding valuable information and this phenomenon is called information explosion.

By reading a series of Tweets a human reader can infer opinion and trend using his cognitive abilities. Automation of such cognitive process is quiet a challenging and nut-cracking task.Short length and narrow domain nature of tweets make it even difficult to process by rule based language processing techniques.To deal such short and narrow domain short textsone requires vary sophisticated machine learning techniques.

This paper presents some experimental work to detect sentiments in Tweets posted by users on microblogging site Twitter. The section 2 discusses the task of sentiment analysis and adopted methodology. It also provides algorithmic formulation of the method for computing purpose. Section 3 presents the details of two datasets used in the experiment and experimental setup developed in R language. Section 4 describes results and section 5 concludes the paper.

2. SENTIMENT ANALYSIS

Sentiment analysis is a computational task to detect sentiments associated with opinionated text. In the current work it is assumed that either a tweet can be positive or it can be negative, thus a two class, classification is required. Formally sentiment analysis task can be represented asquintuple$< O_i, F_{ij}, S_{kijl}, H_k, T_l >$ where, $O_i$is the object to be evaluated, F_{ij} is selected feature of the targeted object O_i, $S_{kij}$is the sentiment polarity of opinion holder k on j^{th} feature of the O_i object. The process involve identification of targeted object, which holds an opinion and then to identify the class of opinion, i.e. positive or negative (Liu, 2009). Among product and servicing industry sentiment analysis is a hot task in current era. Itenables a product or service provider to measure his goodwill on daily basis or larger intervals.

In this paper a Lexicon-pooled hybrid Naïve Bayes classifier is designed and used which incorporates knowledge from sentiment lexicon as well as machine learning classifier. The system is an augmentation over Naïve Bayes approach. Here, the evidence of a word belonging to a sentiment class is computed from both, the machine learning process of Naïve Bayes and the knowledge obtained from sentiment lexicon. Both the probabilities are pooled together to determine which sentiment class a given word belongs to. The designs are evaluated for two feature selection schemes:term frequency (tf) andterm frequency multiplied by inverse document frequency (tf.idf).A detailed description and the mathematics behind Lexicon pooling can be seen in Madhavi et al. (2015).

3. DATASET AND IMPLEMENTATION

3.1 Dataset

The experiments use two twitter dataset, one is emotional labeled tweets used in(Mohammad, 2012), and other dataset obtained from Sentiment140 (Go et al, 2009). Both are annotated for two classes positive and negative. Emotional Labeled Tweets, referred to as Dataset1 comprises of 1200 positive and 1200 negative tweets. Sentiment140, referred to as Dataset2, comprises of 6000 positive and 6000 negative tweets.

Table 1. Details of Dataset used

S. No.	Name of Dataset	#Positive Tweets	#Negative Tweets
D1	Emotion Labeled Tweets	1200	1200
D2	Sentiment140	6000	6000

Table 2 gives the detail of the lexicon which is used in the experiment. A popular lexicon called hashtagsentiment AffLexNegLex is used. It has 43949 words in total. It identifies 19502 positive and 24447 negative words.

Table 2. Lexicon Details

S. No.	Name of Dataset	#Positive Words	#Negative Words
1	HashtagSentimentAffLexNegLex (Unigram)	19502	24447

3.2 R Implementation

The detailed block diagram of the system is shown in figure 1. AR language program reads each tweet as a document, extracts features from the tweet. In next step program calculates feature score using machine learning classifier (Naïve Bayes) with Lexicon knowledge base. By aggregating the scores a sentiment class of the tweet is calculated.

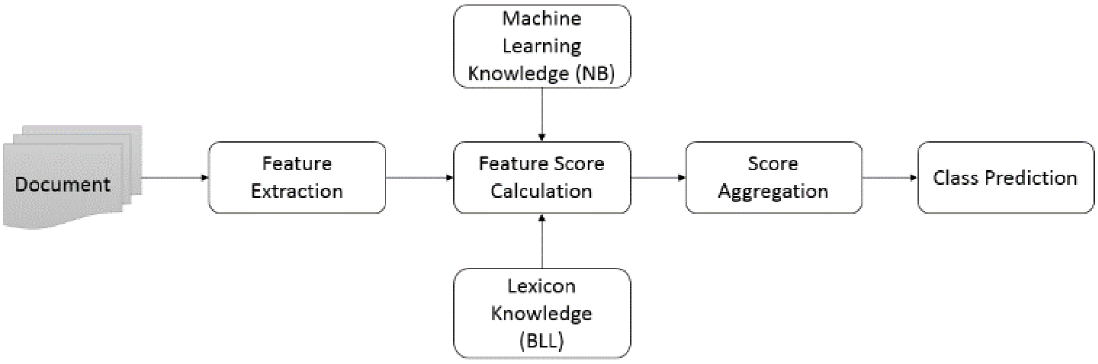

Figure 1.Architectural Block Diagram for Lexicon Pooled Naïve Bayes implementation

4. RESULTS

The table 2 shows the results for various combinations of training data, feature selection scheme and datasets. The 3 and 10 folds cross validation with two types of features, tf and tf.idf. Two classifier designs are implemented and evaluated: a normal Naïve Bayes classifier and a Lexicon pooled Naïve Bayes. For 3 fold cross validation and with tf as feature Naïve Bayes gives 98.71% accuracy for first dataset D1 and 71.30% accuracy for second dataset D2. For 3 fold cross validation and with same feature Lexicon pooled Naïve Bayes gives 90.5% accuracy for D1 and 67% accuracy for D2. For 10 fold cross validation and with tf as feature Naïve Bayes gives 98.71% accuracy for first dataset D1 and 72.60% accuracy for second dataset D2. For 10 fold cross validation and with same feature Lexicon pooled Naïve Bayes gives 92% accuracy for first dataset D1 and 68.4% accuracy for second dataset D2.

For 3 fold cross validation and with tf.idf as feature Naïve Bayes gives 85.08% accuracy for first dataset D1 and 70.35% accuracy for second dataset D2. For 3 fold cross validation and with same feature Lexicon pooled Naïve Bayes gives 71.8% accuracy for D1 and 64.6% accuracy for D2. For 10 fold cross validation and with tf.idf as feature Naïve Bayes gives 86.66% accuracy for first dataset D1 and 71.40% accuracy for second dataset D2. For 10 fold cross validation and with same feature Lexicon pooled Naïve Bayes gives 72% accuracy for first dataset D1 and 65.8% accuracy for second dataset D2.

Table 2. Performance Levels on two Different Datasets

Training Size	Feature Name	Run	D1	D2
Train: 2/3 Test: 1/3	TF	Train Doc. (p : n)	800:800	4000:4000
		NB Acc. (Avg.)	98.71%	71.30%
		Lexicon Pooled NB (Avg.)	90.5%	67.0%
Train: 9/10 Test: 1/10	TF	Train Doc.(p: n)	1080:1080	5400:5400
		NB Acc.(Avg.)	98.71%	72.60%
		Lexicon Pooled NB (Avg.)	92.0%	68.4%
Train: 2/3 Test: 1/3	TFIDF	Train Doc. (p : n)	800:800	4000:4000
		NB Acc. (Avg.)	85.08%	70.35%
		Lexicon Pooled NB (Avg.)	71.8%	64.6%
Train: 9/10 Test: 1/10	TFIDF	Train Doc.(p: n)	1080:1080	5400:5400
		NB Acc.(Avg.)	86.66%	71.40%
		Lexicon Pooled NB (Avg.)	72.0%	65.8%

The figure 2 pictorially compares Naïve Bayes and Lexicon Pooled Naïve Bayes with 3 and 10 cross validation results over tf feature set. Figure 3 pictorially shows Naïve Bayes and Lexicon Pooled Naïve Bayes with 3 and 10 cross validation results over tf.idf feature set.

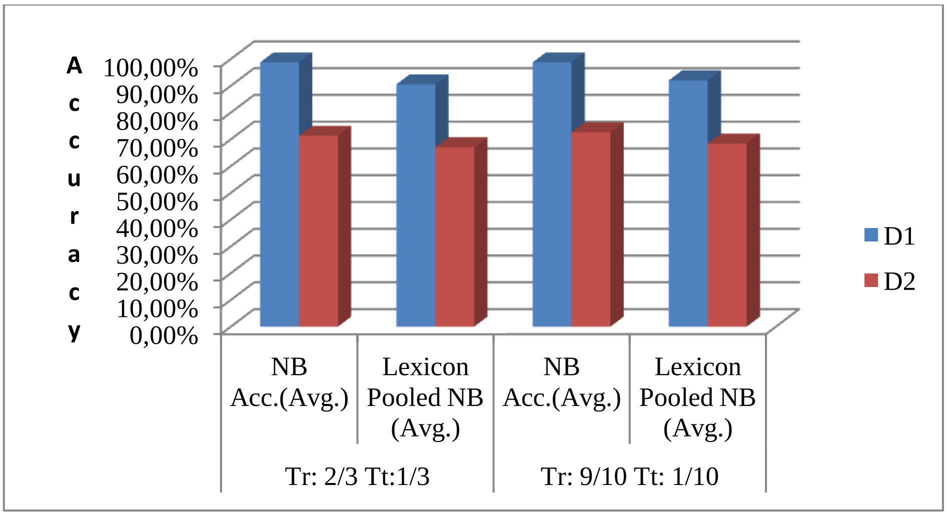

Figure 2. Accuracy Plot oftfas Features

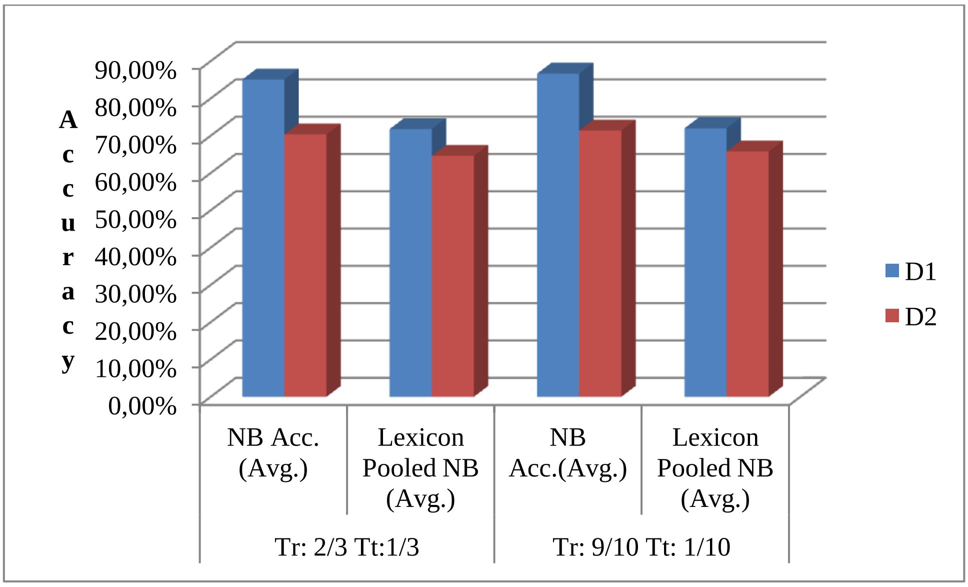

Figure 3. Accuracy Plot of tf.idfas Features

5. CONCLUSION

The experimental results obtained show that Naïve Bayes classifier gives marginally better accuracy then Lexicon Pooled Naïve Bayes classifier irrespective of the dataset, training scheme and features for sort text. For higher fold of cross validation, performance levels are closer. The short size of the tweets can be one of the reasons for the obtained results. It also remains to be seen that if a different sentiment lexicon is used in the pooling process, what will be the impact on performance levels obtained. Analyzing few previous tweets from the same user can help to analyze the system in much effective manner. If a tweet is containing a web link, exploring the page which is provided in the link can also help to analyze the sentiment of tweet.

REFERENCES

1. Go, A., Bhayani, R. and Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1-12.
2. Madhavi D., Piryani R. and Singh V.K (2015).Lexicon Ensemble and Lexicon Pooling for Sentiment Polarity Detection.to appear in IETE Technical Review, Taylor and Francis. DOI:10.1080/02564602.2015.1073572
3. Mohammad, S. M. (2012). Emotional tweets. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (pp. 246-255). Association for Computational Linguistics.