Harnessing Twitter for Automatic Sentiment Identification Using Machine Learning Techniques

Amiya Kumar Dash, Jitendra Kumar Rout and Sanjay Kumar Jena

Abstract User generated content on twitter gives an ample source to gathering individuals' opinion. Because of the huge number of tweets in the form of unstructured text, it is impossible to summarize the information manually. Accordingly, efficient computational methods are needed for mining and summarizing the tweets from corpuses which, requires knowledge of sentiment bearing words. Many computational techniques, models and algorithms are there for identifying sentiment from unstructured text. Most of them rely on machine-learning techniques, using bag-of-words (BoW) representation as their basis. In this paper, we have applied three different machine learning algorithm (Naive Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVM)) for sentiment identification of tweets, to study the effectiveness of various feature combination. Our experiments demonstrate that NB with Laplace smoothing considering unigram, Part-of-Speech (POS) as feature and SVM with unigram as feature are effective in classifying the tweets.

Keywords Bag-of-words (BoW) · Machine learning algorithms · Laplace smoothing · Part-of-Speech (POS)

A.K. Dash (✉) · J.K. Rout · S.K. Jena
Department of Computer Science & Engineering, National Institute
of Technology Rourkela, Rourkela, India
e-mail: 213CS1141@nitrkl.ac.in

J.K. Rout
e-mail: 513cs1002@nitrkl.ac.in

S.K. Jena
e-mail: skjena@nitrkl.ac.in

A. Nagar et al. (eds.), *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, Smart Innovation, Systems and Technologies 44, DOI 10.1007/978-81-322-2529-4_53

1 Introduction

Sentiment analysis (SA) is concerned with automatically extracting sentiment related information from a text and aims to categorize text as positive or negative on the premise of the positive or negative sentiment (opinion) expressed in the document/sentence towards a topic. A document/sentence with positive or negative sentiment is also said to be of positive or negative polarity respectively [1]. The granularity of polarity can be up to the level of words. That is textual information can be classified as either objective or subjective. Objective (non-polar) sentences and words represent facts, while subjective (polar) sentences and words represent perceptions, perspectives or opinions. It is important to make distinction between subjectivity detection and sentiment analysis as they are two separate task in natural language processing. Sentiment analysis can be dependently or independently done from subjectivity detection. Pang and Lee [2] state that to get better result subjectivity detection performed prior to sentiment analysis.

The task of sentiment analysis is very challenging, not only due to the syntactic and semantic variability of language, but also because it involves the extraction of indirect or implicit assessments of objects, by means of emotions or attitudes. That is why automatic identification of sentiment requires fine grained linguistic analysis techniques and substantial efforts to extract features for machine learning or rule-based approaches.

In this paper, we have used three different machine learning algorithm on tweets for automatic sentiment identification and compare the results with Movie reviews obtained by Pang et al. [3]. We investigated a mixture of features like unigram, bigram, POS and adjectives to search out the effective feature for sentiment analysis. It was observed from our experiment that NB classifier with Laplace smoothing and SVM taking unigram and POS as feature gives better result than other features that we have employed.

2 Related Work

The business potential of sentiment analysis has resulted in an exceedingly important quantity of analysis and Pang [4] provides an overview. Ibrahim et al. [5] presents an in depth survey about different techniques used for opinion mining and sentiment analysis. During this section, we restrict our discussion to the work that is most relevant to our approach. Pang Lee et al. use Naive Bayes, Maximum Entropy and Support Vector Machines for SA of movie reviews considering distinctive features like unigrams, bigrams, combination of both, including parts of speech and position information with unigram, adjectives etc. [2, 3]. It was seen from their experiment that Feature presence is more important than feature frequency. It was also observed that for small feature space Naive Bayes performs better than SVM but when feature space is increased SVM perform better than Naive Bayes classifier.

Our work is to perform sentence-level sentiment identification, where we classified tweets using above three machine learning classifier and compare the results with movie reviews.

3 Machine Learning Methods

Our aim in this work is to find out the effective feature for sentiment classification of tweets being positive sentiment or negative sentiment. We used the standard bag-of-feature frame work for implementing these machine learning algorithms. Let $\{w_1, \ldots, w_m\}$ be the m words that can appear in a tweet/sentence; examples include the unigram word "silent" or the bigram "low price". Let $n_i(t)$ be the number of times w_i occurs in tweet t. Then, each tweet t is represented by the tweet vector $t{:} = (n_1(t), n_2(t), \ldots, n_m(t))$.

3.1 Naive Bayes

Naive Bayes classifier is a simple probabilistic classifier that relies on Bayes theorem. The most likely class according to the Naive Bayes classifier is the class among all classes which maximizes the product of two probabilities prior and likelihood, the word in a tweet given the class i.e. how often that word is expressed in a positive tweets or in a negative tweets

$$C_{NB} = \underset{c_j \in C}{\arg\max}\, p(c_j) \underset{i \in positions}{\Pi} p(w_i|c_j) \tag{1}$$

Research on sentiment analysis tells that word occurrence may matter more than word frequency. As tweets are 140 character length occurrence of a word tell us a lot, but the fact that if it occurs more than once may not tell us much more. So we need to clip all the word count in each tweet at one and remove duplicate words in each tweet to retain a single instance of the word. So for our work we have used another variant of Naive Bayes classifier i.e. binarized (Boolean feature) Multinomial Naive Bayes classifier which assumes the features to be occurrence of count.

Laplace Smoothing. Here we used Laplace smoothing assuming that even if we have not seen a given word in the whole corpus, there is still a chance that our sample of tweets happened to not include that word.

$$\hat{p}(w|c) = \frac{count(w, c) + 1}{count(c) + |V|} \tag{2}$$

3.2 Maximum Entropy

The maximum entropy classifier is a probabilistic classifier which belongs to the class of exponential models that has proven effective during a variety of language process applications. It does not assume that the features are conditionally independent of each other. Here our target is to use the contextual information of the tweets (unigram, bigram, and other characteristics) within the text in order to categorize it to a given class (positive or negative). Maximum entropy estimates takes the following exponential form:

$$P_{ME} = \frac{1}{Z(t)}\exp\left(\sum_{i} \lambda_{i,c} F_{i,c}(t, c)\right) \tag{3}$$

where $Z(t)$ is the size of the training dataset used as a normalization function. $F_{i,c}$ is a indicator function for feature f_i and class c, defined as follows,

$$F_{i,c} = \begin{cases} 1, & n_i > 0 \text{ and } c' = c \\ 0, & \text{otherwise} \end{cases}$$

For estimating the λ parameters we use ten iteration of IIS (improved iterative scaling) algorithm, together with a Gaussian prior to counteract over fitting.

3.3 Support Vector Machine

Support vector machines (SVMs) are widely used for various text categorization in past, usually outperforming Naive Bayes classifier. In case of two-class problem with d dimension, the basic idea is to search a hyperplane, represented by vector $\overleftarrow{w}$, that not just differentiates the tweet vectors in one category from those in alternative, yet for which the separation, or margin, is as large as attainable. Let positive and negative be the correct class of tweet t_j and $c_j \in \{1, -1\}$ refers to the class labels positive and negative, then searching a hyperplane corresponds to a constrained optimization problem; where the solution is described as,

$$\overrightarrow{w} = \sum \alpha_j c_j \overrightarrow{t_j}, \quad \alpha_j \geq 0 \tag{4}$$

where the α_j's are derived by solving a dual optimization problem. The tweet vectors t_j are called support vectors for which α_j is greater than zero, as these tweet vectors contribute to the hyperplane. Classification of tweets includes primarily deciding that facet of $\overleftarrow{w}$'s hyperplane they fall on.

4 Experimental Set-up

4.1 Dataset

For implementation we have used *Niek j. Sanders* data set and *Polarity* data set. *Niek j. Sanders* data set contains 5513 hand classified tweets. The corpus contains tweets about apple, goggle, Microsoft and twitter. Tweets are classified into four classes positive, negative, neutral and irrelevant. Irrelevant tweets are those tweets that are not in English language or not related to the topic. In our experiment we have consider three classes positive, negative and neutral. So we converted all irrelevant class to neutral class. The polarity data set is a set of film review documents available for research in sentiment analysis and opinion mining. The most recent available data set is version 2.0, and it comprises 1000 positive labeled and 1000 negative labeled film reviews extracted from the Internet Movie Database Archive.

4.2 Data Pre-processing

Data pre-processing is done to eliminate the incomplete, noisy and inconsistent data [6]. Data must be pre-processed to apply any of the data mining functionality. We have employed the following pre-processing task before applying Machine learning algorithms.

Replace all URLs with a tag AT_USER, replace targets (e.g. @John) with tag USER; replace all the emoticons with a their sentiment polarity by looking up the emoticon dictionary; replace all negations (e.g. not, no, never, cannot) by tag NOT; replace a sequence of repeated characters by two characters, for example, convert coooooooool to cool; replace the words like what, which, how etc., are not going to contribute to polarity (called stop words);special character like, [],{ },() … should be removed in order to remove discrepancies during the assignment of polarity; stripped hash symbol (*#tomorrow* → *tomorrow*). We have used python regular expression for data pre-processing. We employed python Natural Language Toolkit (*NLTK*3.0) to get unigram, bigram, POS features of tweets.

4.3 Evaluation Metrics

The overall performance of individual classifier is measured by:

$$accuracy = \frac{\#of\ correctly\ labeled\ tweets}{\#of\ all\ the\ tweets\ in\ the\ test\ dataset}$$

Table 1 Accuracy of tweets using different features

	Features	No. of features	Frequency or presence	Naive Bayes		Maximum entropy		Support vector machine	
				Reviews (%)	Tweets (%)	Reviews (%)	Tweets (%)	Reviews (%)	Tweets (%)
1	Unigram	5989	Presence	81.0	81.5	80.4	78.36	82.9	82.5
2	Bigram	19,148	Presence	77.3	78.60	77.4	78.0	77.1	77.8
3	Unigram + bigram	25,748	Presence	80.6	80.92	80.8	79.78	82.7	81.6
4	Unigram + POS	19,061	Presence	81.5	82.0	81.2	80.3	81.9	81.99
5	Adjectives	1197	Presence	77.0	69.48	77.7	76.4	75.1	76.4

5 Results and Discussion

We explore a variety of features that are potent for sentiment analysis. We have used N-gram features like unigrams (n = 1), bigrams (n = 2) that are widely used in different of text classification, including sentiment analysis. In our study we experimented with unigrams and bigrams with boolean features. Each n-gram feature is associated with a boolean value, which is set true if and only if the n-gram is present in the tweet [3]. Table 1 represents the different features we have used and the accuracy results of individual classifier. Here we have performed a comparison between the movie review data set used by Pang Lee et al. and our dataset. From Table 1, it has been observed that when we used NB classifier with Laplace smoothing, the classification accuracies resulting from using unigram as features gives better result in case of tweets than movie reviews, but when we used MaxEnt classifier the accuracy result of Movie reviews are more than the tweets.

We additionally considered usage of bigrams to capture negation words for handling negation and phrases for dealing with Word Sense Disambiguation (WSD). *Line*(2) of results table demonstrates that using bigram as feature does not improve performance of the classifier as that of unigram presence. In our experiment we observed that, although bigram presence does not improve the classification accuracy it is as equally useful a feature as unigram; in reality bigrams are found to be effective features for handling word sense disambiguation. We also experimented considering bigram as single feature but the results were not as good, but combination of unigram and bigram features (*Line*(3) of results table) produces results competitive with those obtained by using unigram.

POS features are verified effective in sentiment analysis. Since adjectives are good indicators of sentiment, they are usually considered as effective feature for sentiment analysis. Our experiment shows (*Line*(5) of results table) that considering only adjectives produces results competitive with those obtained by using unigram and bigram. *Line*(4) of results table shows that all the three classifier produces better result considering unigram and POS as feature. *Line*(1) of results table shows that SVM with unigram as feature produces best result out of all the features we have considered.

6 Conclusion

We have studied the sentiment analysis result for tweets collected from twitter public domain. The results table shows that classification accuracies using unigram presence and POS as feature turned out to be most effective as compared to other alternative features we employed. Though there are many machine learning techniques are available, however no single technique has proven to consistently outperform the other across many domains.

References

1. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Human Lang. Technol. **5**(1), 1–167 (2012)
2. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, p. 271. Association for Computational Linguistics (2004)
3. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
4. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retrieval **2**(1–2), 1–135 (2008)
5. Sadegh, M., Ibrahim, R., Othman, Z.A.: Opinion mining and sentiment analysis: a survey. Int. J. Comput. Technol. **2**(3), 171–178 (2012)
6. Wang, W., Chen, L., Thirunarayan, K., Sheth, A.P.: Harnessing twitter "big data" for automatic emotion identification. In: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom), pp. 587–592. IEEE (2012)
7. Nigam, K., Lafferty, J., McCallum, A.: Using maximum entropy for text classification. In: IJCAI-99 Workshop on Machine Learning for Information Filtering. vol. 1, pp. 61–67 (1999)
8. Turney, P., Littman, M.L.: Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus (2002)
9. Turney, P.D.: 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. Association for Computational Linguistics (2002)
10. Vipul Pandey, C.I.: Sentiment analysis of microblogs. In: Diploma Thesis, CS 229 Project Report, Stanford University