Sentiment Analysis using Averaged Histogram

Subarno Pal
Dept. Of Computer Science and Engineering
Academy Of Technology
Hooghly, India

Soumadip Ghosh
Dept. Of Computer Science and Engineering
Academy Of Technology
Hooghly, India

ABSTRACT

Sentiment analysis or opinion mining is a process of categorizing and identifying the sentiment expressed in a particular text. The need of automatic sentiment retrieval of the text is quite high as amount of reviews obtained from the Internet are huge in number. Reviews on various ‘E-commerce websites’, ‘social networks’, and ‘movie review websites’ come up huge in number regularly. These reviews on popular products help in determining the public opinion towards the product. An averaged histogram model is proposed in the process that deals with text classification in continuous variable approach. After data cleaning and feature extraction from the reviews, average histograms are constructed for every class, containing a generalized feature representation in that particular class. Histograms of every test elements are then matched with the averaged histograms of every class using k-Nearest Neighbor and Naïve Bayesian Classifier. Results showed on 3000 reviews a steady classification accuracy of 79-80% with the Naïve Bayesian Classifier with very little cost of computation, and increase in the number of training dataset k-Nearest Neighbor can give up to a high accuracy of 85%. This work proposed here is language independent, neither include any dictionary nor depend on the meaning of any word.

General Terms

Sentiment Analysis, Text classification, k-Nearest Neighbor, Naïve Bayesian Classifier

Keywords

Averaged Histogram

1. INTRODUCTION

Commercial websites like Amazon, Yelp & IMDb reviews are a major platform today where people express their opinion towards any particular event or subject. Numerous tweets come up during political events that clearly show the public trend towards any political party or issue. People express reviews of the latest movies they have watched on IMDb like websites. Product and service reviews from Amazon, eBay also help us to decide which product to buy and which service to avail. Moreover the reviews obtained from personal blogging websites are mostly unbiased and contains personal experience towards a particular product or service. According to some recent statistics in 2016, the micro-blogging website Twitter averaged at 313 million monthly active users, while almost 244 million users avail Amazon, and 164 million eBay worldwide, who regularly share their reviews on these websites. Micro-blogging websites users also vary across the globe and consist of people from different age groups and socioeconomic backgrounds.

Text classification is a technique that classifies a text on the basis of matching patterns of words or phrases present in it. The main challenge of text classification is to find out the exact characteristics that tells the opinion or sentiment of the writer towards the object. Sentiment of a text is mainly classified into two types, 'positive' and 'negative' classes. The Positive class determines the reviews in favor, and negative class determines the reviews that are against the subject. Based on the classification of every single review, a cumulative inference could be drawn from a recent collected dataset of standard reviews on that particular subject that would show us the present scenario of public sentiment.

General supervised machine learning [13] approach is usually followed up for sentiment analysis, which involves training the machine with a part of the labeled dataset, followed by testing of the remaining elements. We used this machine learning approach for classification of the sentiment from the obtained dataset.

The main purpose of sentiment analysis is to make a market study or research on a particular event or item using machine learning techniques. Based on the result of sentiment analysis further steps are taken by authorities dealing with the products or events. Our study involves determining the sentiment of a small text review or micro-blog with general classification algorithms modified to some extent.

Text classification is generally done with discrete variables, but this works is done on a continuous variables using average histogram approach for every class. It doesn’t include any dictionary and does not depend on the meaning of any word.

Section 2 of this paper contains related work and section 3 proposed the method for sentiment analysis in detail including the feature extraction and classification. Section 4 contains the experimental results obtained and it’s in depth discussion.

2. RELATED WORKS

Sentiment analysis sometimes called opinion mining is a general classification problem and has involved many researchers in recent times. Analysis of text involves extraction of opinion or sentiment of the writer writing the review. A huge amount of such work has been focused on the document level, sentence level and the Entity/Subject level of the text.

Determining the sentiment associated with the product review Mehto A. et. al. [1] proposed a ‘Lexicon based approach for Sentiment Analysis’ based on an aspect catalog. The keywords present in aspect catalog identified in those sentences in which features of any product are mentioned. Aspect catalog is referred again to find degree of importance corresponding to the feature with respect to the product/subject. Based on these sentences weighted features from the aspect catalog are summed up to find the sentiment of the text.

In a recent work [2] using Onto-Fuzzy Logic on hash-tagged words of twitter are given special preference for determining sentiment of the text. Hash-tags are categorized in many types such as topic hash tags, sentiment hash-tags and the last type is sentiment-topic hash tags varying on different parts of speech and the polarity of the text.

Chikersal P. et. al. [3] developed by Rule-based Classifier combining with Supervised Learning, used the rule-based classifier which is based on rules that are dependent on the occurrences of featured keywords and polarity in tweets. Whereas, the Support Vector Machine (SVM) is trained on semantic, dependency, and sentiment lexicon based features. The tweets are classified as positive, negative or unknown by the rule-based classifier, and as positive, negative or neutral by the SVM.

Kumar A. et. al. [4] proposed Sentiment Analysis on Twitter (Sentence Level) to develop mixed model of corpus based and dictionary based method to determine the semantic orientation of the opinion words of tweets. Based on the corpus method of dividing the sentence into different parts of speech the dictionary method is applied. The sentiment is calculated using linear equation by pre-processing the tweet into simple sentence and including emotion intensifiers.

Wang et. al. [5] illustrates three types of useful information that are sentiment polarity of tweets containing the hash-tag; hash-tags co-occurrence relationship and the literal meaning of hash-tags. In order to incorporate the first two types of information into a classification frame-work where hash-tags can be classified collectively, they proposed a novel graph model and investigate three approximate collective classification algorithms for inference. They also showed that the performance can be remarkably improved using an enhanced boosting classification setting in which they employed the literal meaning of hash-tags as a semi-supervised information.

Nizam and AkÕn in their unsupervised learning for sentiment classification in Turkish [18] used tweet words as features and tweet data were clustered in positive, negative and neutral labeled classes. Then, this dataset is used to detect classification accuracy with NB, DT and KNN algorithms.

He and Zhou used semi-supervised learning on movie review dataset [21] in which they obtained an initial classifier by including previous information extracted from an existing sentiment lexicon. The extracted information used as training data for classifier and unlabeled features were labeled by extracted information, showed a much better result than other similar experiments.

3. METHODOLOGY

The dataset consisted of micro blogs and reviews from different web sources preferably IMDB, Amazon, Yelp etc. The micro blogs and reviews are small in size and emphasize on the quality of any product or item showing the sentiment of the writer towards the object. The dataset consisted of two main classes 'positive' and 'negative'.

3.1 Data preprocessing and cleaning:

This process includes the removal of punctuation from the text sets and tokenization of the text into single words in a single case (Upper or Lower case). Each token consists of a single word without any punctuation in the texts. Repetitions of the tokens are removed. Detailed example of this process is shown on the table:

Table 1. Data Preprocessing Example

Original Text Review	Tokens Obtained
'A very, very, very slow-moving, aimless, movie about a distressed, drifting young man.'	'a', 'very', 'slow-moving', 'aimless', 'movie', 'about', 'distressed', 'drifting', 'young' and 'man'.

3.2 Data Selection

Tokens obtained after preprocessing consists of words from the text that might not be responsible for sentiment of the text, those tokens decrease the efficiency of the system. So a set of tokens are to be selected from the entire token set for a better computational efficiency.

Negative Keywords: The words or parts of speech that do not take any part in expressing the sentiment of the text such as the prepositions, articles and conjunctions are listed from the dictionary as negative keywords, and it is completely independent of the dataset used. The tokens consisting of the Negative Keywords are eliminated from the set.

Positive Keywords: The tokens that are obtained after removal of the Negative Keywords are taken into consideration and from that set the most frequent 1000 keywords are selected as positive keywords and those are used for determining the sentiment of the text.

Table 2. Data Selection Example

Previously Obtained Tokens	Selected Tokens
'a', 'very', 'slow-moving', 'aimless', 'movie', 'about', 'distressed', 'drifting', 'young' and 'man'.	'very', 'slow-moving', 'aimless', 'movie', 'distressed', 'drifting', 'young' and 'man'

3.3 Feature Extraction

A feature vector is a histogram or N-grams [12] consisting of 1000 features, the set of positive keywords arranged in the decreasing order of their frequency. A histogram is a set of values of the frequency of every positive keyword divided by the total number of positive keywords occurring in the text. This gives the relative frequency of the features in a text.

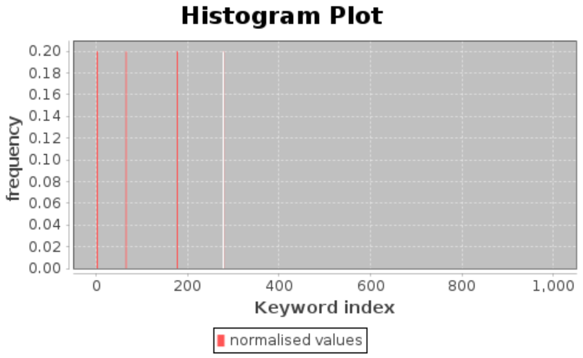

Fig 1: Histogram obtained from a feature vector

3.4 Training

Dataset is divided into separate Training and Testing part. Histograms for every the training element is computed.

Construction of Average Histogram: Histograms of a particular class is added up and divided by the number of training elements of that class in the set. This computation gives Average Histograms of a particular class. This Average Histogram of every class represents the probability of relative occurrence of every feature in the class.

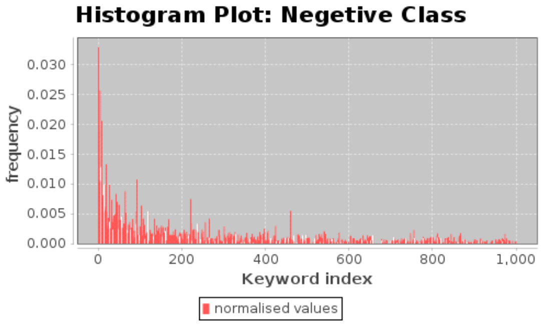

Fig 2: Average histogram of Negative class obtained from the IMDb dataset

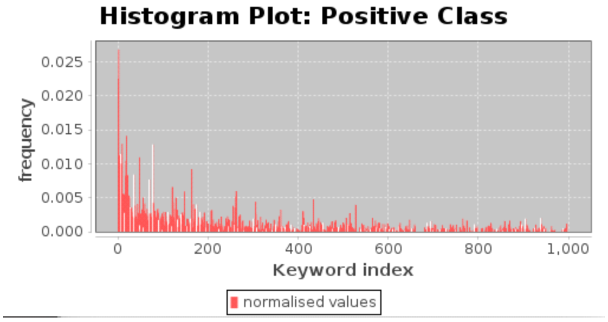

Fig 3: Average histogram of Negative class obtained from the IMDb dataset

3.5 Classification

5. Classification: The remaining dataset are used for testing of the model. Histogram for every data element is constructed and is matched with the average histogram of every class using two different classifiers described below. The performance of classifiers is extremely sensitive to the quality of training data [20].

3.5.1 k-Nearest Neighbor Classifier (kNN):

The Nearest Neighbor classifier is based upon learning by analogy, computing the distance between the feature vectors of test tuple and all the training tuples, k minimum distances for the testing tuple and the training set, the majority vote for the class present in the set is to be predicted the class of the tuple[7], [18].

As we are using one average histogram for every class so best match in distance will be considered as the predicted class of the text, therefore the default value of k will be 1.

The Euclidean distance is used to measure the distance between the feature vectors, as it gives better results on continuous variables. The Euclidean distance between two point p and q is the length of the line segment connecting them (pq).In Cartesian coordinates, if p = (p1, p2,..., pn) and q = (q1, q2,..., qn) are two points in Euclidean n-space, then the distance (d) from p to q, or from q to p is given by the Pythagorean formula:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \qquad (1)$$

The block diagram of the process in shown in Fig. 4

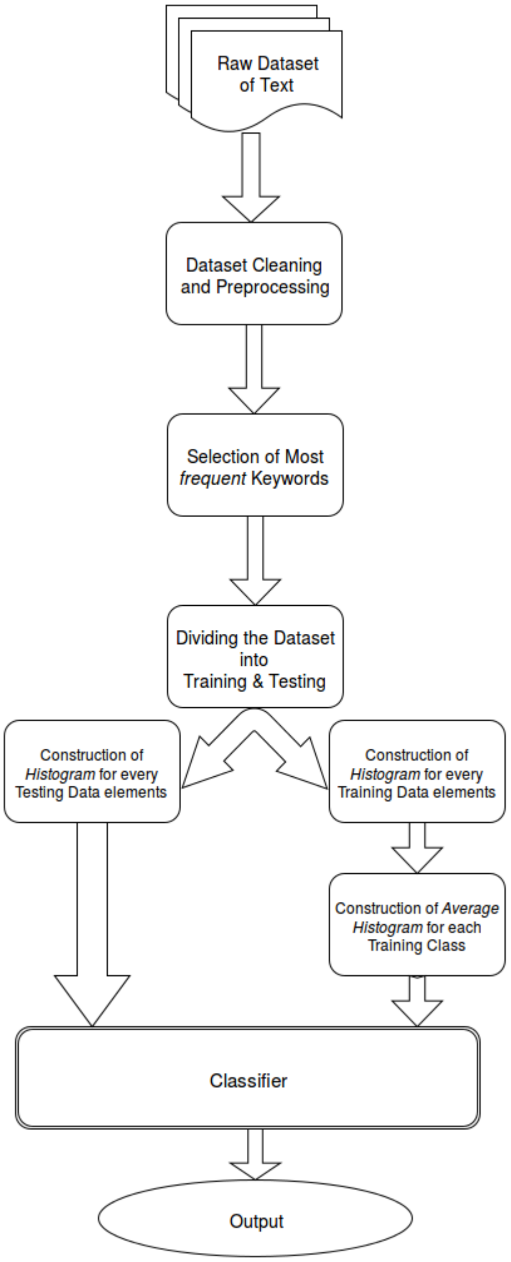
Raw Dataset of Text
Dataset Cleaning and Preprocessing
Selection of Most *frequent* Keywords
Dividing the Dataset into Training & Testing
Construction of *Histogram* for every Testing Data elements
Construction of *Histogram* for every Training Data elements
Construction of *Average Histogram* for each Training Class
Classifier
Output

Fig 4: Proposed classification method based on averaged histogram.

3.5.2 Naive Bayesian Classifier:

It is a probabilistic model used for classification [14]-[19]. Each tuple represented by n-dimensional attribute vector, X = (x1, x2, ... , xn) of m classes C1, C2,..., Cm.

The average histogram represents the probability of every feature in a text of that class. And the obtained histogram from testing test is the probability of every feature in that testing text.

Thus by Bayes theorem probability of a tuple X, the probability of it belonging to class C_i is [7].

Now by maximizing we get the predictive class of the text.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (2)$$

4. RESULTS AND DISCUSSION

Dataset: The dataset [6] used in the experiment consisted of reviews from amazon [9], Yelp [10] and IMDb [8] websites. Dataset consisted of 1000 reviews of items or product labeled into positive and negative classes individually from three sources.

Results: Two different classifiers ended up two different results that are discussed below:

4.1 k-Nearest Neighbor Classifier Result:

For kNN classifier we could see a steady growth in the accuracy of the machine as the number elements in the training sets were increased. kNN classifier performs well on averaged histograms for a highly trained system and the accuracy can go up to 87.85% (training with 900 elements). The dataset consisted of micro blogs and reviews from different web sources preferably IMDB, Amazon, Yelp etc. The micro blogs and reviews are small in size and emphasize on the quality of any product or item showing the sentiment of the writer. The dataset is mainly classified into two classes 'positive' and 'negative'.

Table 3. Accuracy of k-Nearest Neighbor classifier expressed in percentage

No. of Training Data elements	Accuracy (%)			Average Accuracy (%)
	Yelp dataset	Amazon dataset	IMDb dataset	
650	71.26	71.83	70.40	71.16
700	72.48	71.81	71.47	71.92
750	70.75	71.56	72.77	71.67
800	78.28	70.20	75.75	74.74

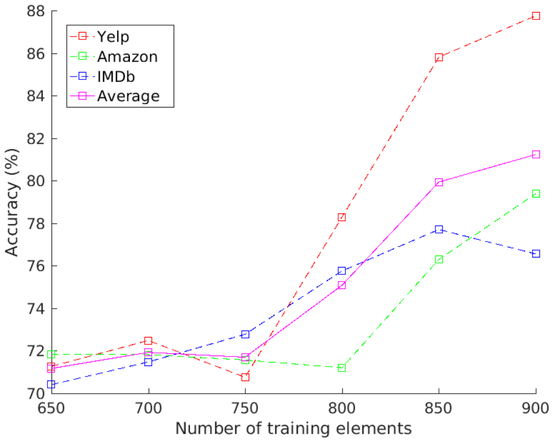

Fig 5: Accuracy of k-NN of k-Nearest Neighbor classifier

4.2 Naive Bayesian Classifier Result:

It showed a standard output of 78-80% accuracy varying on the different dataset with 700 to 900 training elements. The Naive Bayesian Classifier is quite steady in output and didn't have high peaks or low drenches in the graph.

Table 4. Accuracy of Naïve Bayesian Classifier expressed in percentage

No. of Training Data elements	Accuracy (%)			Average Accuracy (%)
	Yelp dataset	Amazon dataset	IMDb dataset	
650	77.01	82.47	76.72	78.73
700	77.18	81.54	76.17	78.29
750	78.62	80.24	77.82	78.89
800	80.80	79.19	76.76	78.91
850	81.75	79.05	77.67	79.49
900	82.75	78.67	78.65	80.02

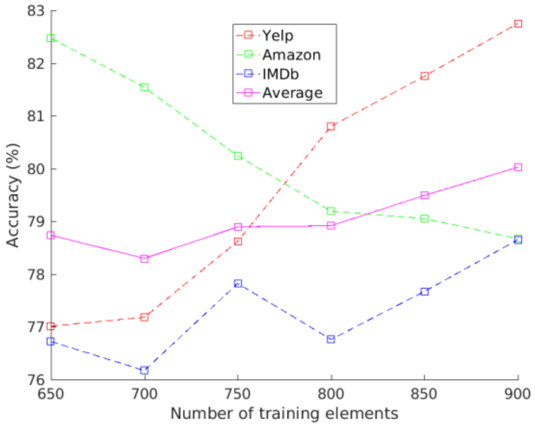

Fig 6: Accuracy of Naive Bayesian classifier

5. EXPERIMENTAL ANALYSIS

From the experimental results that we have got KNN classifier showed increasing results with increase in training data, but the accuracy, but the Naive Bayesian model showed steady results throughout. The Naive Bayesian model showed much reliable and constant classification accuracy than the KNN classifier which only gave good result when highly trained. For better accuracy in both the models training set to be increased as much as possible, then we can achieve a very good steady accuracy.

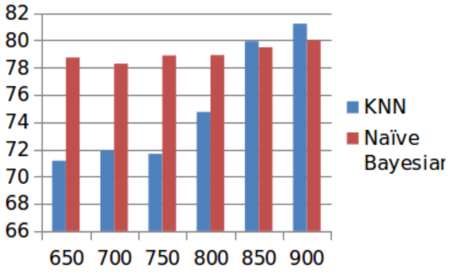

Fig 7: Comparison of Accuracy between kNN and Naive Bayesian classifier

6. CONCLUSION

A sentiment analysis technique using average histogram model has been proposed. Text classification is generally taken as discrete variables classification but this concept builds a continuous variable approach where we could achieve a decent accuracy with minimum cost of computation. Two classifiers, kNN and Naïve Bayesian classifier that give decent accuracy with minimum computation cost are tested and results are shown. The steadiness of the Naïve Bayesian Classifier is quite better than the kNN. No dictionary used in the system, that means the model and the results are Language Independent and do not depend on the meaning of any particular word. Just by changing the Negative Keywords set for any language the system may be easily implemented. The proposed work demands very low computational operations achieving a steady results, so this can be implemented on systems with low computational power. This work might be extended on labeled dataset increasing the number of classes in the system. This work might be also extended on large text dataset containing document level data from social media or news data. Prediction models featuring different emotional lexicon and features might further improve the accuracy of the system.

7. REFERENCES

[1] A. Mehto and K. Indras. "Data Mining through Sentiment Analysis: Lexicon based Sentiment Analysis Model using Aspect Catalogue", IEEE, 2016 Symposium on Colossal Data Analysis and Networking (CDAN).

[2] S. Joshi, S. Mehta, P. Mestry and A. Save. "A New Approach to Target Dependent Sentiment Analysis withOnto-Fuzzy Logic", 2 nd IEEE International Conference on Engineering and Technology (ICETECH), 17 th & 18 th March 2016.

[3] P. Chikersal, S. Poria and E. Cambria. "SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning", Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp.647–651,

[4] A. Kumar and T. Mary Sebastian. "Sentiment Analysis on Twitter", International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012, pp.372-378.

[5] X. Wang, F. Wei, X. Liu, M. Zhou, M. Zhang. "Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach", Microsoft Research Asia, Beijing, China.

[6] Kotzias et. al. "From Group to Individual labels using Deep Features", KDD 2015.

[7] Han J. and Kamber M. "Data Mining: Concepts and Techniques".

[8] imdb: Maas et. al., 2011 'Learning word vectors for sentiment analysis'

[9] amazon: McAuley et. al., 2013 'Hidden factors and hidden topics: Understanding rating dimensions with review text'

[10] Yelp: Yelp dataset challenge http://www.Yelp.com/dataset_challenge

[11] Kouloumpis E. et. al. "Twitter Sentiment Analysis: The Good the Bad and the OMG!", Fifth International AAAI Conference on Weblogs and Social Media.

[12] Warintarawej, P., Laurent, A., Pompidor, P. and Laurent, B. (2010) 'Classification of brand names based on n-grams', International Conference of Soft Computing and Pattern Recognition, 2010.

[13] H.-X. Shi and X.-J. Li, 'A sentiment analysis model for hotel reviews based on supervised learning', 2011 International Conference on Machine Learning and Cybernetics, Jan. 2011.

[14] Maqbool Al-Maimani ,Naomie Salim, Ahmed M. Al-Naamany, "Semantic And Fuzzy Aspects Of Opinion Mining",Journal Of Theoretical And Applied Information Technology Vol. 63 No.2, 20th May 2014, pp.330-342.

[15] H. Kang, S. J. Yoo, D. Han, "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews", Expert Systems with Applications, vol. 39, no. 5, pp. 6000–6010, 2012.

[16] M. Çetin and M. F. Amasyali, "Active learning for Turkish sentiment analysis," In IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA), pp. 1–4, 2013.

[17] R. Moraes, J. F. Valiati, and W. P. Gavião Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN", Expert Systems with Applications, vol. 40, no. 2, pp. 621–633,2013.

[18] H. Nizam and S. S. AkÕn, "Sosyal Medyada Makine Ö÷renmesiile Duygu Analizinde Dengeli ve Dengesiz Veri Setlerinin PerformanslarÕnÕnKarúÕlaútÕrÕlmasÕ", In 19. Türkiye'de ønternet KonferansÕ, øzmir, 2014.

[19] M. Meral and B. Diri, "Sentiment analysis on Twitter", Signal Processing and Communications Applications Conference (SIU), pp. 690–693, 2014.

[20] X. Zhu, X. Wu, and Y. Yang. Effective classification of noisy data streams with attribute-oriented dynamic classifier selection. Knowledge and Information Systems, 9(3):339–363, 2006.

[21] Y. He, D. Zhou, "Self-training from labeled features for sentiment analysis", Information Processing & Management, vol. 47, no. 4, pp. 606–616, 2011.