2nd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)
8-10 December 2016, Rajshahi-6204, Bangladesh

Sentiment Mining from Bangla Data using Mutual Information

Animesh Kumar Paul*, Pintu Chandra Shill
Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna, Bangladesh
*animesh10kuet@gmail.com

Abstract— **Due to the explosion of social networking sites, blogs and review sites (for example, Amazon, Twitter, and Facebook, etc.) it provides an overwhelming amount of textual information. We need to organize, explore, analyze the information for making a better decision from the side of customers and companies. Thus, sentiment analysis is the best way in which it determines the author's feelings expressed in reviews as positive or negative opinions by analyzing an enormous number of documents. In this work, we used Mutual Information (MI) for the feature selection process and also used Multinomial Naive Bayes (MNB) for the classification of Bangla and English reviews. The experimental results demonstrate that the system can achieve satisfactory accuracy for Bangla dataset compare to English dataset where Bangla dataset is generated from Amazon's Watches English dataset.**

Keywords—Sentiment Analysis, Sentiment levels, Feature Selection, Mutual Information, Multinomial Naive Bayes (MNB).

I. INTRODUCTION

With the rapid wideness of e-commerce, many companies are placing their product into online market for selling, and a huge number of user now buy a product from online. Millions of people express their shopping experience and product features and their nuances. Due to the explosion of social networking sites, blogs and review sites (for example, Amazon, Twitter, Facebook, and Consumer Reports), it provides an overwhelming amount of textual information about the present condition of the product as an active feedback from the user. From these active feedback information, manufacturing companies [1] want to know what is the feeling of their customers on their or other company's products which helps them to maintain their online reputation [2-4] and at the same time, user easily can get ideas who want to know which product will be good for them with compare to others product.

Due to the increasing amount of user-generated contents (opinion, reviews, comments, feedbacks, suggestions), analyzing all of this online information is necessary for making an effective decision from the customers and company. Explore, analyze and organize these kinds of extensive information manually is time-consuming. Automatic sentiment analysis for these user-generated information refers to the field of natural language processing (NLP), computational linguistics (CL), and text mining. Sentiment analysis [5] is the task of retrieve the opinions about the product and classify the given contextual information into different polarity (positive or negative opinion). This text classification task is also referred to as polarity classification.

The research in the field of sentiment analysis started much earlier. In [6], here it used several machine learning systems (Naive Bayes, Maximum Entropy, and SVM techniques) to classify a large corpus of movie reviews for unigrams and bigrams. The idea of the difficulties in the opinion classification was to get from [6]. In [7], it used an unsupervised method to classify the reviews as thumbs up (recommended) or thumbs down (not recommended). It uses document level opinion classification. In [8], here it used separate subjective sentences from the rest of the text, and it achieved the best result using a SVM method based. In [9], it used WordNet syntactic relations together with topic relevance to calculate the subjectivity scores for words. In [10], it is as [6], taking comments off of social networking sites about movie reviews. In [11], it summarized the opinions expressed in reviews about the different parts of a product which are distinct from the classical document summarization. In [12], it used part-of-speech information for minion the opinions because adjective has a correlation with subjectivity [13]. In [14], sentiment analysis is done at the phrase and sentence level except for the document level. In [15], it worked on Bangla sentiment analysis and it identify the sentiment information in each report, aggregates them and represents the summary information in text.

In this paper, it investigates a small part (positive and negative attitudes towards products.) of the significant problems in sentiment analysis for Bangla data. The goal is to determine the polarity of Bangla language texts using Multinomial Naive Bayes based on Feature selection method.

The association of this article is as follows: the details process of sentiment analysis is described in section II. After that, Section III describes the experimental results and analysis. Finally, some concluding remarks are presented in section IV.

II. Mutual Information based Multinomial Naive Bayes Model

The overall process of opinion mining is shown in Fig.1. It contains several steps: i) Collect the dataset, ii) Process the training dataset before passing it to Feature extraction method,

978-1-5090-5785-6/16/$31.00 ©2016 IEEE

iii) Select the necessary features using Mutual information method, iv) Train the multinomial naive bayes based on only the selected features, v) Test the model using the Testing dataset.

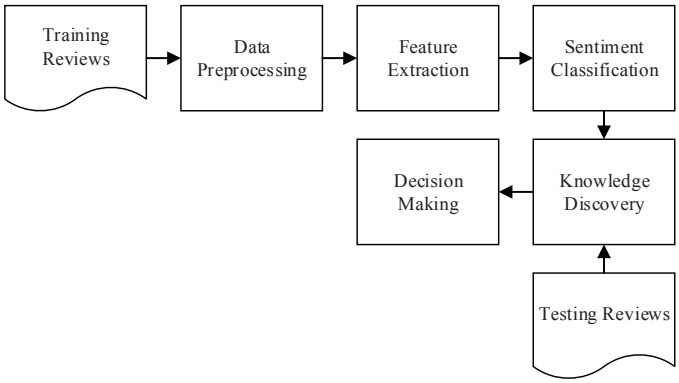

Fig. 1. Flow chart of Text Classification Process (Training + Testing)

A. Dataset

For this research purpose, we needed to use English and Bangla dataset. A set of product reviews collected from Web data: Amazon Reviews [16] are used as English corpus which contains over 68356 reviews of watch-products [16]. All the reviews are given ratings to be ranged from 1-star to 5-star such as (1, 2, 3, 4, 5). For Bangla corpus, Bangla dataset is generated from this Amazon dataset.

B. Data Preprocessing

1) English Corpus Preprocess

From the English corpus, we took only the reviews as a positive (score = 1) which have rating 4 or 5 and a negative (score = 0) which have rating 1 or 2. The same number of positive and negative reviews for English and Bangla corpus are used on training and testing phases.

a) Stopping Words

Before passing the training data to feature extraction algorithm, the training data are filtered by the stopping word means to remove all unnecessary words from the training data for getting better features as output from the feature extraction algorithm. If there will be not at least 3 characters in a word, then it will be considered as a stopping word. And also some stopping words such as away, awfully, b, back, be, became, didn't, different, do, does, doesn't, doing, don't, done, etc. are taken into consideration which is collected from different websites. All punctuation except periods, apostrophes, and hyphens, additional white spaces, URLs, repeating letters, non-English alphabetic words is removed.

2) Bangla Corpus Preprocess

For sentiment analysis, it's hard to find a standard Bangla dataset, and for this reason, a Bangla dataset is generated from the processed English dataset by translating each word one by one. Non-Bangla alphabetic words are removed.

3) Negation Handling

In the sentiment analysis, negation handling is one of the contributing factors for the classifier. Negation is one of the most common linguistic constructions [17], and the polarity of a sentence can be changed for this. For this reason, we need to consider the negation in the sentiment analysis [18, 19]. For English corpus, we consider "*no", "not", "never", "n't"* as negation term which affects the polarity of a word. For the presence of negation word in a sentence, all the words will not indicate the negative meanings [19]. For English corpus, we just transformed all words into ("not_" + word) which are after the negation word. If the given line is "I don't like this watch." after negation the given line will be "I do not not_like not_this not_watch.".

For the Bangla corpus, negation words affect the sentence polarity like English. The negation terms in Bangla are "না", "নয়", and "নি" which will change the contextual polarity. It's hard to understand which portions of the sentence will be affected by the negation word in Bangla corpus and from this thinking, we transformed all words into ("না_" + word) of a sentence. If the given line is "আমি করি না পছন্দ এই ঘড়ি।" after negation the given line will be "না_আমি না_করি না না_পছন্দ না_এই না_ঘড়ি।".

C. Feature Extraction Using Mutual Information

In the machine learning algorithms like SVM, Neural Networks, etc., feature extraction used as an integrated module in the different system. In a larger picture, it is used to identify a subset of features for further text classification stage. Feature extraction method is used to remove redundant features (decrease the dimensionality of the space of feature) which contain high disambiguation capabilities, avoid system failure, handling the skewed datasets, reduce the computational cost, and minimize the overfitting of the learning system. In the feature extraction process, it identifies the portions of a given document which will affect the contextual polarity (positive or negative sentiment) and combines those parts of the documents so that it will increase the probability of the document falling into one of these two polarities (0, 1). If we use higher dimensional features like bigrams, trigrams, these will increase the number of features in which it contains many redundant and noisy features, and those would affect the accuracy of the system. There are different feature selection approaches such as Mutual Information, $\chi 2$ Feature selection, Frequency-based feature selection, etc. Here, for feature extraction, we used mutual information for filtering the features from the training dataset both for Bangla and English.

Mutual Information (MI) [20] measures how much information the presence/absence of a term contributes to making the correct classification decision on c. Formally, the mutual information of two random variables can be shown as:

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t) P(C = e_c)},$$

Where U is a random variable that takes values $e_t = 1$ (the document contains term t) and $e_t = 0$ (the document does not contain t), and C is a random variable that takes values $e_c = 1$ (the document is in class c) and $e_c = 0$ (the document is not in class c). We write U_t and C_c if it is not clear from context which term t and class c we are referring to.

For Maximum Likelihood Estimates (MLEs) of the probabilities, above equation is equivalent to Equation:

$$I(U;C) = \frac{N_{11}}{N} \log_2 \frac{N\, N_{11}}{N_{1.} N_{.1}} + \frac{N_{01}}{N} \log_2 \frac{N\, N_{01}}{N_{0.} N_{.1}} + \frac{N_{10}}{N} \log_2 \frac{N\, N_{10}}{N_{1.} N_{.0}} + \frac{N_{00}}{N} \log_2 \frac{N\, N_{00}}{N_{0.} N_{.0}}$$

Where the Ns are counts of documents that have the values of e_tc and e_c that are indicated by the two subscripts. For example, N10 is the number of documents that contain t (e_t = 1) and are not an equal number (e_c = 0). N1. = N10 + N11 is the number of documents that contain t (e_t = 1) and we count documents independent of class membership ($e_c \in \{0, 1\}$). N = N00 + N01 + N10 + N11 is the total number of documents.

D. Text Classification using Multinomial Naive Bayes

Multinomial Naive Bayes (MNB) [21] is a probabilistic learning classifier which is based on Bayes' decision rule with robust and naive independence assumptions. It is a modified version of Naive Bayes that is designed more for text documents. Original Naive Bayes [22] only considers the presence and absence of particular words in a document whereas MNB uses multinomial distribution for all pairs where it uses the word counts and rectify the underlying calculations to act within. MNB shows desire efficiency with the accuracy.

In the multinomial naive Bayes for a given document, the probability of a document d being in class c is computed as

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c)$$

Where $P(t_k|c)$ is the conditional probability of term t_k occurring in a document of class c. $P(t_k|c)$ denotes a measure of how much evidence t_k contributes that c is the correct class. P(c) is the prior probability of a document occurring in class c. Choose the document that has a higher prior probability when document's terms do not provide clear evidence for one class versus another.

In text classification, our goal is to find the best appropriate class for the document. The best class in NB classification is the most likely or maximum a posteriori (MAP) class c_{map}:

$$c_{map} = arg\max \hat{P}(c|d) = argmax\, \hat{P}(c) \prod_{1 \le k \le n_d} \hat{P}(t_k|c)$$

We write $\hat{P}$ for P because we do not know the true values of the parameters P(c) and $P(t_k|c)$, but estimate them from the training set.

The above equation can be written as:

$$c_{map} = argmax\, [\log \hat{P}(c) + \prod_{1 \le k \le n_d} log\hat{P}(t_k|c)]$$

Each conditional parameter $log\hat{P}(t_k|c)$ is a weight that indicates how good an indicator t_k is for c. Similarly, the prior $\log \hat{P}(c)$is a weight that indicates the relative frequency of c.

For the priors,

$$\hat{P}(c) = \frac{N_c}{N'}$$

Where N_c is the number of documents in class c and N is the total number of documents.

We estimate the conditional probability $\hat{P}(t|c)$ as the relative frequency of term t in documents belonging to class c:

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

Where T_{ct} is the number of occurrences of t in training documents from class c, including multiple occurrences of a term in a document. T_{ct} is a count of occurrences in all positions k in the documents in the training set.s

We use the add-on or Laplace smoothing, which simply adds one to each count because some words may not exist for a particular class c.

$$\hat{P}(t|c) = \frac{T_{ct}}{\sum_{t' \in V}(T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B'}$$

Where B = |V| is the number of terms in the vocabulary. Add-one smoothing can be interpreted as a uniform prior.

III. EXPERIMENTAL RESULT

We use Amazon's Watches [16] dataset which contains 68356 reviews, but in the empirical analysis, we used only 16000 reviews with 8000 positive and 8000 negative level. We used 13000 reviews which include 6500 positive and 6500 negative reviews for the training purpose and 3000 reviews for testing purpose which contains the equal number of positive and negative reviews.

After processing the chosen dataset, mutual information method is used to extract the optimal number of features from the training dataset. Accuracy, sensitivity, specificity depends on the number of selected features. Tiny number/a large number of selected features are unable to give better performance in distinguishing the polarity into the document. Fig.2 shown that a small number and a vast number of selected features can't be able to provide better accuracy and we need to choose an optimal number of features. We selected the top 10000 features with maximum mutual information for English and Bangla dataset.

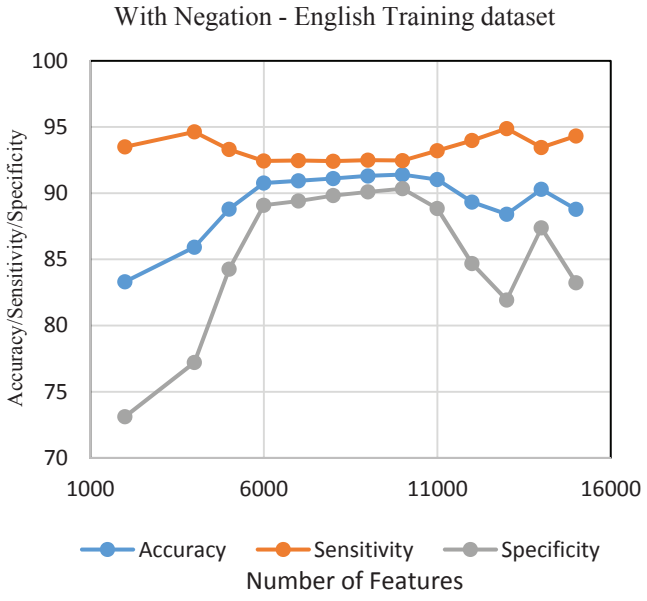

Fig. 2. System Accuracy, Sensitivity, Specificity Based on Different Number of Feature Selection.

For English [Fig.3(a)], in the testing phase using English training data for testing we get 91.1% accuracy without using negation and get 91.4% accuracy with negation. In testing phase, using English testing data, we get 85.1% accuracy without using negation and get 85.8% accuracy with negation.

For Bangla [Fig.3(b)], in testing phase using Bangla training data for testing we get 88.54% accuracy without using negation and get 87.79% accuracy with negation. In the testing phase, using Bangla testing data, we get 84.78% accuracy without using negation and get 83.77% accuracy with negation. The accuracy for Bangla dataset is close to English dataset's accuracy where Bangla and English dataset both are identical.

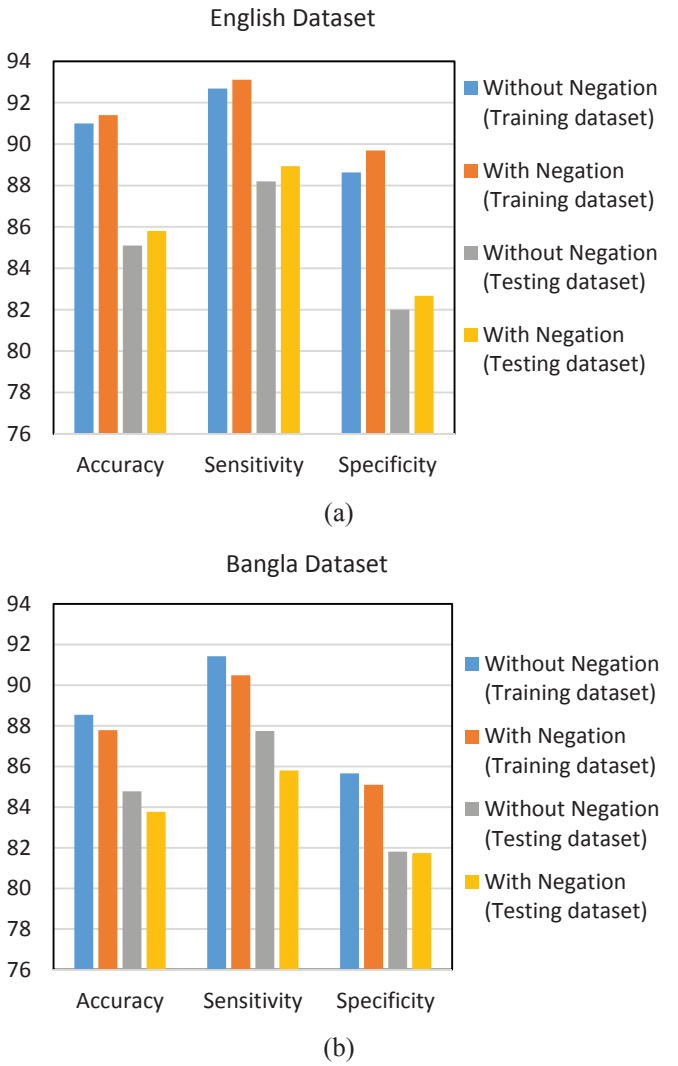

Fig. 3. Evaluation of the given system on different dataset (a) English Dataset (b) Bangla Dataset.

IV. CONCLUSION

For analyzing the reviews and classifying the sentiments, we used the Multinomial Naive Bayes (MNB) based on feature selection method (Mutual Information) for both Bangla and English datasets. In the feature selection stage, preprocessed data are passed to the Mutual Information (MI) method, and then MI selects an optimal set of features. Based on these features, MNB gives the prediction for each document to classify into positive or negative opinions. From the experimental result, the accuracy for the Bangla dataset slightly differs from the accuracy for the English dataset of the proposed system. For this method, we got a satisfactory level of accuracy for Bangla dataset compare to English dataset where Bangla dataset is generated from Amazon's Watches dataset.

REFERENCES

[1] Ann, and Khurshid Ahmad. "Sentiment polarity identification in financial news: Acohesion-based approach." ANNUAL MEETING-ASSOCIATION FORCOMPUTATIONAL LINGUISTICS. Vol. 45. No. 1. 2007.

[2] Das, Sanjiv and Mike Chen, "Yahoo! for Amazon: Extracting market sentiment from stockmessage boards", Proceedings of APFA-2001.

[3] Das, Sanjiv and Mike Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the web, Management Science", 53(9): pp.1375-1388, 2007

[4] Yubo Chen and JinhongXie, "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix, Management Science", vol 54, no 3, pp.477-491, 2008.

[5] Bing Liu, "Sentiment analysis and opinion mining, Morgan and Claypool publishers", 2012.

[6] Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?:sentiment classification using machine learning techniques", in Proceedings of the ACL-02 conference on Empirical methods in natural language processingVolume 10, pp.79–86, 2002

[7] P.D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of the Association for Computational Linguistics (ACL), pp.417–424, 2002.

[8] Pang, B., & Lee, L., "A sentimental education: Sentiment analysis using subjectivitysummarization based on Minimum Cuts". in Proceedings of the 42nd Annual Meeting ofthe Association of Computational Linguistics. 2004.

[9] Mullen, T., & Collier, N., "Sentiment analysis using support vector machines withdiverse information sources". In Proceedings EMNLP'04, pp.412-418, 2004.

[10] K. Yessenov & S. Misailovic, "Sentiment analysis of movie review comments. Methodology", pp.1–17,2009.

[11] M. Hu & B. Liu, "Mining and summarizing customerreviews", in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp.168–177. ACM, 2004.

[12] M.Hu, & B. Liu, "Mining opinion features in customer reviews", Proceedings of 19th National Conference on Artificial Intelligence, pp. 755-760, 2004

[13] Wiebe, J., Bruce, R., Bell, M., Martin, M., & Wilson, T., "A corpus study of evaluative and speculative language", Proceedings of 2nd ACL SIGdial Workshop on Discourse and Dialogue. Aalborg, Denmark, 2001

[14] T. Wilson,"Recognizing contextual polarity in phrase-level sentimentanalysis", In Proceedings of HLT-EMNLP, pp. 347–354, 2005.

[15] Das, A. & Bandyopadhyay, S.,"SentiWordNet for Bangla." In Knowledge Sharing Event-4: Task 2: BuildingElectronic Dictionary", Mysore, 2010.

[16] Web data: Amazon reviews, https://snap.stanford.edu/data/web-Amazon.html

[17] Maral Dadvar, Claudia Hauff, Franciska de Jong, ''Scope of Negation Detection in Sentiment Analysis''

[18] Jia, L., C. Yu and W. Meng, "The effect of negation on sentiment analysis and retrieval effectiveness", in 8th International Conference on Information and Knowledge Management, 2009.

[19] Wiegand, M., et al. A survey on the role of negation insentiment analysis. in '10 Proceedings of the Workshop onNegation and Speculation in Natural Language Processing 2010: Association for Computational Linguistics.

[20] Huawen Liu, Jigui Sun, Lei Liu, Huijie Zhang, "Feature selection with dynamic mutual information", Pattern Recognition 42, pp.1330 – 1339, 2009.

[21] Daniel Jurafsky & James H. Martin , "Speech and Language Processing", Chapter 7, 2015.

[22] V. Muralidharana, V. Sugumaranc, "A comparative study of Naive Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis", Applied Soft Computing Volume 12, Issue 8, pp.2023–2029, 2012.