A Sentiment Analysis Approach using Effective Feature Reduction Method

Arash Mazidi
Islamic Azad University of Aliabad Katool Branch
Aliabad Katool, Iran

Elham Damghanijazi
Islamic Azad University of Aliabad Katool Branch
Aliabad Katool, Iran

ABSTRACT
Sentiment analysis has attracted researchers in recent years which aims to present an automatic method for analyzing comments, assessments, opinions and sentiments of a text. In this paper, Ngram feature vector and POS (Part of Speech) are extracted from text and it is tried to find a proper combination of feature vectors so that texts can be classified as positive and negative opinions. In order to choose the most useful features, information gain ratio is used, then machine learning algorithms are used to investigate the effect of different features on sentiment analysis. In this paper, 4 groups of data are studied including film evaluation, products' evaluation (including book, DVD and electronics). Classification results are studied for three types of feature vectors: Ngram feature vector, POS feature vector and combination feature vector of Ngram with POS. results show that combinational feature vector performs better in sentiment analysis. By combining features, Boolean Multinominal Naïve Bayes (BMNB) results are improved compared to support vector machine classification algorithm.

Keywords
Sentiment analysis, Feature selection method, machine learning, support vector machine, information gain, information gain ratio.

1. INTRODUCTION
Web pages contain a lot of texts which include comments, opinions and evaluations of users about a product or service. Information which can be obtained from these data are useful and necessary for manufacturers and organizations which offer services. Additionally, it gives useful information for users who wish to select a specific product or service. Consider a person who wants to buy a cell-phone or a digital camera, who would undoubtedly look for information about quality of the products and quality of services. But today, with the ever increasing growth of internet, one can obtain information about a specific product using online comments and experiences of hundreds of people who have used that product. But there are some problems including large data volume and contradiction of some data. For example, 300000 daily messages are recorded on Twitter page of Justin Bieber [1]. Thus, it is necessary to automate this process and deliver the final results to the user. Sentiment analysis feature which is called opinion analysis also, is a research study which studies opinions, sentiments, evaluations, appraisals and attitudes of human about an entity like products, services, organizations, people, events and their features [2]. Before 2000, few researches were done in this field due to lack of data resources. But, after 2000, with growth of internet and social networks, stability of data volume, this area became an active field in natural language processing [2]. Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine [3].

Text documents are normally represented as a feature-document matrix in sentiment analysis. Features can be single words from the text document or more complex pairs extracted by different schemes that adds information in order to enrich the feature-document matrix representation. Having diverse feature types however creates a problem of high dimensionality due to the vast number of features and relations they hold. Thus, feature selection helps in ensuring that effective and efficient sentiment analysis applications can be developed by selecting features that are relevant and informative to assist classifiers to perform better and to reduce the processing load by narrowing down the feature set [4].

In opinion mining, feature extraction plays a very important role in summarizing reviews. Opinion mining is basically concerned with extracting the opinions of the users and analyzing them to draw a meaningful conclusion from their respective ideas in terms of reviews and feedbacks. Also, the goal of a feature selection method is to score the features by order of some measure of relevance, estimated from some training set [5].

The remainder of the paper is organized as follows. In Section 2, related works on sentiment analysis and feature selection for sentiment analysis are reviewed. Section 3 defines the proposed algorithm in this paper. In Section 4, presents the performance evaluation of our approach for feature selection. Finally, some conclusive remarks are given in Section 5.

2. RELATED WORKS
There are many heuristic algorithms that can be used to solve the problems [6][7]. Data mining approaches can be applied in different problems. Take, for instance, cloud computing environements contains hundreds of servers and thousands of Virtual Machines (VMs) [8]. Hence, there resources and countless requests made by customers causes big-data in the environments [9]. There are several data mining appproaches such as classification techniques that can be applied in autonomic management of these data[10]. Pang and Lee established feature vector by extracting Unigram, Bigram and adjectives from text. In this study, Naïve Bayes algorithm, support vector machine and maximum Entropy were used to analyze sentiment of movie review data sets. They showed that in feature vector representation, using term presence gives better results compared to using term frequency. Moreover, support vector machine showed higher precision in classification compared to other methods [11]. In some researches, Unigram, Bigram, Trigram or a combination of them is used, and the results indicated that combination of these features improves classification operation [12]. Some papers have used POS to form feature vectors and have

reported favorable results [13], [14]. Using POS label along with the word itself is also used as feature vector which can reduce sematic ambiguity, thus it increases accuracy of the machine learning algorithm [15].

Ngram model of characters has also been used in studies. For example, Bigram of "Like" would be "Li ik ke" [16]. But in this mode, large number of features causes problems and using feature selection algorithms created time complexity problem.

One of the main problems in the process of sentiment analysis is the large number of features. Some of these features are redundant and some others are not related to intent classes and these problems reduce precision and speed of classification algorithm. Using feature selection algorithms might solve the aforementioned problems and improve speed and precision of the machine learning algorithm [17].

There are two types of feature selection method: single-variable and multi-variable. Single variable method considers each feature alone, evaluates the feature and ranks it. Like Chai-square, Log likelihood and information gain. Although these methods are fast, but because they evaluate each feature alone and do not consider its relation with other features, their precision is lower. These algorithms are appropriate for data sets with large feature vectors [18].

Some studies have used log likelihood ration to select useful features [12], [13]. On the other hand, using information gain algorithm is also common and gives favorable results for classifying data in sentiment analysis procedure [12], [13].

Some studies have used multi-variable methods to select useful features including Genetic Algorithm. Problem of this algorithm is its time complexity [16]. Abbasi et.al. in 2011, presented a complete set of Ngram features for sentiment analysis and used feature relation network to increase accuracy of sentiment analysis on dataset [18].

Agrawal and Meital in 2013 studied using information gain. Minimum redundancy and maximum dependency for feature selection. They also employed Unigram, Bigram and a selection of POS words to model the text. The proposed classification method was applied to movie review and gave better results compared to those obtained by Abbasi et.al. They showed that minimum redundancy and maximum dependency selection methods perform better compared to information gain method [19].

In order to model texts, different feature vectors are presented and some studies have used a limited number of feature vectors. Some studies have also provided a complete set of Ngram feature vectors and have used a feature selection method to extract most useful features. In this paper, it has been tried to select feature vectors such that favorable results are obtained using a simple feature selection method, such that time complexity is eliminated. Set of features used in this research is combination of Ngram features and POS word. Positive effect of Ngram features in sentiment analysis is proved. Feature vector of POS word is also selected to reduce semantic ambiguity of the selected words such that texts are classified into positive and negative classes with higher accuracy.

3. PROPOSED ALGORITHM

Web pages and social networks comprise a large volume of texts. Some of such texts include comments and opinions of used about a specific entity. In this study, a method is proposed which classifies data sets to positive and negative opinions about a specific entity.

Principles of the proposed method are shown in Figure 1. As mentioned before, in this paper, it has been tried to present a vector of important and effective features in sentiment analysis and their proper combination. For this purpose, text is pre-processed. First, words which have been repeated less than three times are eliminated, then text is converted to normal processing framework. For example, words like I'm, Don't are converted to I am and Do not. Stopwords are eliminated. Finally, negation words like Never, Not and No are managed such that each negation word is eliminated and two words before the negation words and two words after that are converter for negative form by adding a Not. Anyway, if dot is met, negation is stopped. For example, "I like hamid, but I don't like javad" is converted to "I like hamid not like javad". In the next step, and negation words are manages as Like-NOT, Like Hamid, Javad_Not. In this example, before negation word, NOT, there is only one comma. Thus upon reaching to the first punctuation mark, here comma, negation is stopped and not word before not is converted to the agreed from, but two words after not are converted to agreed negative form.

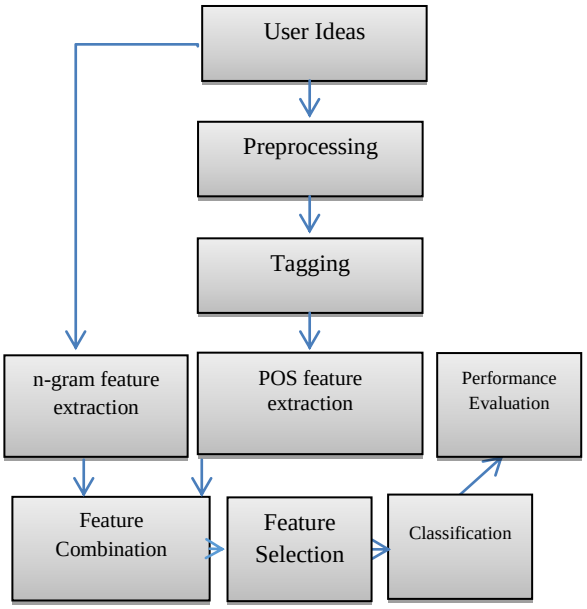

Fig 1. The Framwork of Proposed Method

3.1. Tagging Words' Role

For tagging words' role. POS Tagger software from Stanford University is used. This software has been developed by natural language processing team of Stanford University. In most scientific researches in the field of natural language processing, this software is used. Figure 2 shows an example of using this software.

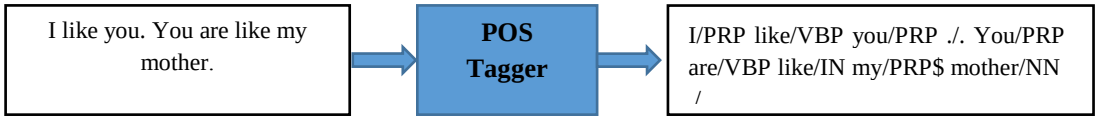

Fig 2. An Example of POS Tagger

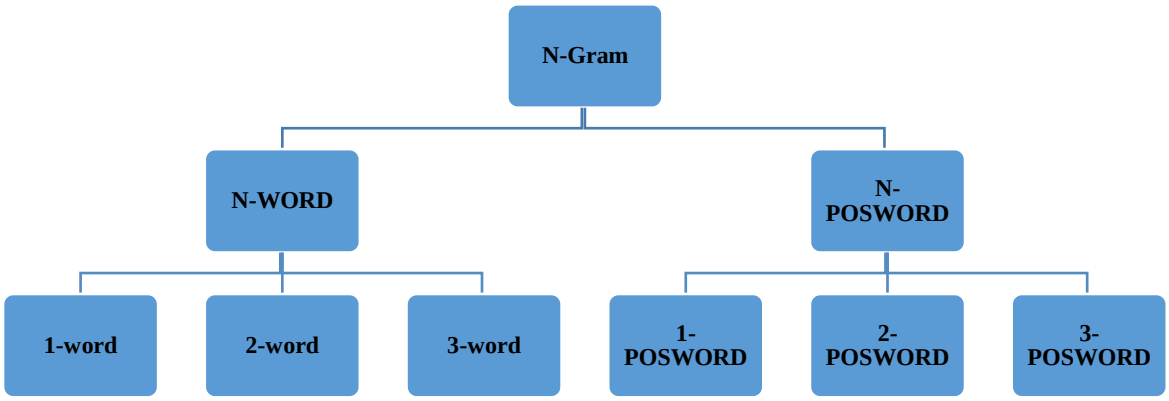

Fig 3. The proposed features for sentiment analysis

3.2. Feature Extraction

Main purpose of this study if to extract desired features for modeling texts. Figure 3, shows this set of features, after performing preprocessings and tagging, features are extracted from the text. Table 1 shows examples of these features.

Studies performed previously have widely used unigram and bigram. Effectiveness of unigram features in sentiment analysis has been proved. In this study, unigram+bigram+trigram features and 1POSWord+2POSWord+3POSWord are used. None of these features are studied individually. In the following, Ngram is used instead of unigram+bigram+trigram and POSWord is used instead of 1POSWord+2POSWod+3POSWord for simplicity.

Table 1. The proposed feature sets and examples for each feature

I go home.		Example
I, go, home	Unigram	N-gram Features
I go, go home	Bigram	
I go home	Trigram	
I/FW, go/VBP, home/NN	1-posword	POSWord-Features
I/FW go/VBP, go/VBP home/NN	2-posword	
I/FW go/VBP home/NN	3-posword	

3.3. Feature Selection Algorithm

In order to select the most useful features, information gain ratio algorithm is used which is a single variable method. Information gain ratio algorithm is the modified version if information gain algorithm. Information gain ratio assigns a non-zero weight to each feature which has useful information for classification. And this weight is obtained using the equation 1.

$$InfoGain(C,w) = -\sum_{j=1}^{k} p(C_j) log\ (p(C_j)) + p(w) \sum_{j=1}^{k} P(C_j|w) \log\left(p(C_j|w)\right) + p(\overline{w}) \sum_{j=1}^{k} p(C_j|\overline{w}) \log p(C_j|\overline{w}) \quad (1)$$

p(Ci) is a fraction of documents which belong to class C_i. p(w) is a fraction of documents in which word w exists. $p(C_i/w)$ is a fraction of documents which belong to class C_i and word w exists in them.

For normalizing information gain, Equation (2) is used.

$$IntrinsinInfo(C,w) = -\sum_{j=1}^{k} \frac{|C_j|}{|C|} \log \frac{|C_j|}{|C|} \quad (2)$$

In Equation (2), $|C_j|$ is the number of documents which belong to class j and |C| is the total number of documents.

Using Equation (3), information gain ratio is calculated.

$$InfoGainRatio(C,w) = \frac{InfoGain(C,w)}{IntrinsinInfo(C,w)} \quad (3)$$

Normalizing information gain allows us to discriminate features with higher accuracy, thus it is preferred to use information gain ratio algorithm instead of information gain algorithm. Features with higher informationGainRation are selected for classification.

3.4. Evaluation Measure

In order to evaluate classification efficiency of sentiment analysis, recall, precision and F-measure are used [19] (Equations 4,5,6).

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (4)$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (5)$$

In this research, F-measure which is a combination of recall and precision is used (Equation 6).

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \quad (6)$$

4. IMPLEMENTATION AND RESULTS

4.1 Dataset

In order to evaluate the proposed method, 4 data sets are used. Movie review data set was proposed by Lee et.al [11]. In 2006, Blitzer et.al. presented a dataset including reviews on Amazon's products [20]. There are two versions of this data set: preprocessed version and raw version. In the first version, a series of useful preprocesses are applied to the text and texts are ready for researchers. But in this study, raw version of this data set is used because of using POS Tagger and requirement to raw texts. In order to evaluate stability of results, the proposed method is tested on review data sets of Amazon products including books, DVD and electronic equipment.

4.2 Machine Learning Algorithm

Speed and accuracy of BMNB classification algorithm for sentiment analysis is higher compared to other algorithms including Naïve Bayes and support vector machine [20]. Thus for finding most useful features, BMNB algorithm is used. As mentioned before, accuracy of SVM algorithm is higher compared to Naïve Bayes and Maximum Entropy [1]. In order to compare the improvements with previous studies, SVM method is used along with BMNB. SVM and BMNB algorithms of weka 3.6 with default settings were used [21].

4.3 Discussion

Table (2) shows results obtained for features considered in this study and their combination. As can be seen, accuracy of results obtained by applying BMNB is higher than SVM. Thus it can be emphasized that accuracy of SVM for sentiment analysis is higher. Moreover, results presented in Table (2) show that accuracy of Ngram features is lower than POSWord in most cases, because POSWord POSWord features not only include the words but also include their roles and resolve ambiguity in some cases. For example, world like means being interested and similar. Therefore it has ambiguity. Word Like with verb tag can help the machine learning algorithm to classify texts because it describes positive attitude of the writer towards a specific entity. In POSWord feature vector, this ambiguity is re solved by tagging word's role.

Using a combination of Ngram and POSWord features improves classification accuracy significantly. But in other data sets it has resulted in less accuracy. Because type of texts and description of users is movie review set is different from Amazon's Products review sets in terms of structure. In movie review data set, number of features before and after pre-processing and applying primary filters has been more than fifty thousand features which indicated large volume of this set. While, other data sets had a maximum of twenty four thousand features.

As can be seen in Table (2), accuracy of SVM for book and electronic data set for all three types of feature vectors has not improved much which is due to redundant and unrelated features. Thus, in these cases, a more favorable feature selection algorithm can be used to discriminate redundant and non-useful features. Information gain ratio method assigns weight to each feature which described importance of each feature and does not discriminate redundant and non-useful features directly. As mentioned before, redundant features and non-useful features reduce effectiveness of other features and decrease accuracy and speed of the classification algorithm.

Table (3) compares results of the current study with previous studies including Agrawal and Meital studies. In this table, best results of both studies are compared.

Agrawal and Meital in 2013, employed minimum redundancy and maximum dependency [19]. This method is able to find redundant and non-useful features better. But in the current study, purpose is to present a set of rich features which can be used in other data sets. As can be seen, movie review and dvd data sets have improved more significantly. And for book and electronic data set, BMNB algorithm has improved significantly. But for SVM, Agrawal and Meital obtained higher accuracy for book and electronic data set by employing a more favorable feature selection method.

Table 2: F-measure for Ngram, POSWord and Ngram+POSWord feature vectors

Electronic		Book		Dvd		Movie		Dataset
BMNB	SVM	BMNB	SVM	BMNB	SVM	BMNB	SVM	
92.2	87.8	93.8	87.5	95.1	87.7	98.4	91.4	Ngram
92.4	88.6	94.0	87.9	94.0	87.2	97.6	91.6	POSWord
93.7	88.4	94.7	87.8	96.3	88.4	99.4	96.7	Ngram+POSWord

Table 3: Comparing results of the proposed method with method presented by Agrawal and Meital. In this table, best results are compared

Electronic		book		Dvd		Movie		Datasets
BMNB	SVM	BMNB	SVM	BMNB	SVM	BMNB	SVM	
91.8	89.0	92.5	88.3	91.5	88.0	91.1	90.2	Agrawal,Mittal method [19]
93.7	88.4	94.7	87.7	96.3	88.4	99.4	96.7	Proposed Ngram+POSWord method

5. CONCLUSIONS

In this study, linguistic features of a feature set for texts are classified into positive and negative sets. Using a set of linguistic features for sentiment analysis is not new. In this study, Ngram and POSWord are employed to present a proper model of text. And SVM and BMNB method are used for classifying these algorithms. It is shown that accuracy of BNMB is higher than SVM. Accuracy of POSWord features is higher than Ngram geatures, because they have better information to resolve ambiguity. Moreover, combining POSWord and Ngram features has helped resolving ambiguities and has improved classification accuracy.

Using a proper method for selecting useful features can be studied in future studied. Redundant and unrelated features reduce effectiveness of other features and decrease accuracy of the classification algorithm. Thus it is better to use a favorable feature selection algorithm with proper speed which can find most useful features among a large volume of features.

6. REFERENCES

[1] Ghiassi, M., Skinner, J. and Zimbra, D., 2013. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, *40*(16), pp.6266-6282.

[2] Cambria, E., 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, *31*(2), pp.102-107.

[3] Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), pp.1093-1113.

[4] Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), pp.1093-1113.

[5] Asghar, M.Z., Khan, A., Ahmad, S. and Kundi, F.M., 2014. A review of feature extraction in sentiment analysis. *Journal of Basic and Applied Scientific Research*, *4*(3), pp.181-186.

[6] Mazidi, A., Fakhrahmad, M., & Sadreddini, M. 2016. A meta-heuristic approach to CVRP problem: local search optimization based on GA and ant colony. *Journal of Advances in Computer Research*, *7*(1), 1-22.

[7] Damghanijazi, E., & Mazidi, A., 2017. Meta-Heuristic Approaches for Solving Travelling Salesman Problem. International Journal of Advanced Research in Computer Science. 8(4).

[8] Horri, A., Rahmanian, A., & Dastghaibyfard, G. H. 2015. Energy and performance-aware virtual machine consolidation in Cloud computing a two dimensional approach. Turkish Journal of Engineering, 1, 20–35.

[9] Rahmanian, A., Dastghaibyfard, G., & Tahayori, H. 2017. Penalty-aware and cost-efficient resource management in cloud data centers. International Journal of Communication Systems, 30(8), e3179. http://doi.org/10.1002/dac.3179

[10] Ghobaei-Arani, M., Shamsi, M., & Rahmanian, A. A. 2017. An efficient approach for improving virtual machine placement in cloud computing environment. Journal of Experimental & Theoretical Artificial Intelligence, 1–23. http://doi.org/10.1080/0952813X. 2017.1310308 .

[11] Pang, B., Lee, L. and Vaithyanathan, S., 2002, July. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86). Association for Computational Linguistics.

[12] g, V., Dasgupta, S. and Arifin, S.M., 2006, July. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions* (pp. 611-618). Association for Computational Linguistics.

[13] Gamon, M., 2004, August. 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). Association for Computational Linguistics.

[14] Fei, Z., Liu, J. and Wu, G., 2004, September. Sentiment classification using phrase patterns. In *Computer and Information Technology, 2004. CIT'04. The Fourth International Conference on* (pp. 1147-1152). IEEE.

[15] Wiebe, J., Wilson, T., Bruce, R., Bell, M. and Martin, M., 2004. Learning subjective language. *Computational linguistics*, *30*(3), pp.277-308.

[16] Abbasi, A., Chen, H., Thoms, S. and Fu, T., 2008. Affect analysis of web forums and blogs using correlation ensembles. *IEEE Transactions on Knowledge and Data Engineering*, *20*(9), pp.1168-1180.

[17] Hall, M.A. and Smith, L.A., 1997. Feature subset selection: a correlation based filter approach.

[18] Abbasi, A., France, S., Zhang, Z. and Chen, H., 2011. Selecting attributes for sentiment classification using feature relation networks. *IEEE Transactions on Knowledge and Data Engineering*, *23*(3), pp.447-462.

[19] Agarwal, B. and Mittal, N., 2013, March. Optimal feature selection for sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 13-24). Springer Berlin Heidelberg.

[20] Blitzer, J., Dredze, M. and Pereira, F., 2007, June. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL* (Vol. 7, pp. 440-447).

[21] WEKA. Open Source Machine Learning Software Weka, http://www.cs.waikato.ac.nz/ml/weka/