ISBN: 978-989-8704-10-8 © 2014

ASSESSMENT OF FEATURE SELECTION METRICS FOR SENTIMENT ANALYSES: TURKISH MOVIE REVIEWS

Fırat Akba, Alaettin Uçan, Ebru Akcapinar Sezer and Hayri Sever
Hacettepe University, Computer Engineering Department, Ankara, Turkey

ABSTRACT

Sentiment analysis systems pursuit the goal of detecting emotions in a given text with machine learning approaches. These texts might include three kinds of emotions such as positive, negative and neutral. Entertainment oriented texts, especially movie reviews, contain huge amount of possible emotional information. In this study, we aimed to represent each movie reviews by using small number of features. For this purpose, information gain, chi-square methods have been implemented to extract features for decreasing costs of calculations and increasing success rate. In experiments, employed corpus includes Turkish movie reviews, support vector machine and naïve bayes had been employed for classification and F_1 score was used for performance evaluation. According to the experimental results, support vector machine achieved 83.9% performance value while classification of movie reviews in two (positive and negative) categories and also we obtained the 63.3% performance value while classification with support vector machine into three categories.

KEYWORDS

Sentiment Analyses, Feature Selection, Support Vector Machine, Naïve Bayes, Turkish Corpus.

1. INTRODUCTION

Sentiment Analysis (SA) can be defined as it is a process that provides extracting people emotions or attitudes by analyzing features on text, images or other things which includes sentiments. SA activities attract enterprises to observe market gaps and new opportunities. For human, emotions and opinions play important role for daily activities or decisions. The field of SA is widely used for commercial and social media fields. There is increasing number of blogs, reviews, forums, social network web pages, related data and information flow which have been created on Internet day by day and this type of huge information flow cannot be processed manually. Therefore, machine learning (ML) methods such as support vector machine (SVM) (Cortes, C and Vapnik, V., 1995) and naïve bayes (NB) (Lewis, 1998) can offer a solution for classifying sentiments on oversized data and information flow without using manpower.

According to our best knowledge, there have been several applications which are developed to satisfy the requirements of SA. These applications are implemented for different native languages, especially for English. In general, supervised ML methods are applied for SA such as SVM, NB, Maximum Entropy (ME) (Berger et al., 1996), K-nearest neighbor (kNN) (Cover, T., and Hart, P., 1967), and N-Gram language model (Suen, C. Y., 1979). There are two types of different SA methods such as lexicon based sentiment and ML based approaches. For the lexicon based analysis, some interesting studies can be summarized. In Turney's (2002) study, calculation of semantic orientation was applied for prior polarity lexicon scores. Esuli et al. (2006) pointed out how to realize translating English DAL (Whissell, 1989) to other languages and calculating polarity scores. This solution was applied by Ghorbel et al. (2011) for French, Valdivia et al. (2012) for Spanish and Denecke (2008) for German language and also other languages had been successfully accomplished. For ML based methods, Pang et al. (2002) presented the method which arranges the traditional topic based text classification for sentiment classification by using classical ML methods such as Naïve Bayes, ME and SVM and also they achieved highest success rate by SVM. When we looked at from the first studies to present ones, sentiment classification success rates are changing between 60% and 86% percent (i.e. Pang et al., 2004; Pak and Paroubek, 2010; Agarwal et al., 2011; Narr et al., 2012; Saif et al., 2012; Becker et al., 2013; Habernal et al., 2013). In these studies, movie reviews, trip advisor, commercial ratings,

social media and SMS data were generally employed as experimental data. Additionally, they applied standard lemmatization and stemming methods and then features were selected by using bag of words method and also weighted features were used in some of them.

Actually, it is easy to reach related studies for English or other languages that belong to same language family. However, there are limited studies in SA for the Turkish. Erogul (2009) accomplished sentiment analysis with implementing SVM and some of natural language processing (NLP) technics by collecting movie review data. He achieved 85% percent of success rate while classifying review data as only positive or negative. Also Boynukalin (2012) tried to improve results by creating a data set and adding new features which were created by using the morphological structure of Turkish Language. Cakmak et al. (2012) showed the relation between root of words, emotion symbols and sentences in Turkish Language, they investigated total 197 words which include emotions by using fuzzy logic technics. Vural et al. (2013) implemented lexicon based SA and they achieved nearly same as much as success rate with ML methods had. Ozsert et al. (2013) studied on SA by using technique of seed word by adding new words for each emotion category. Kaya et al. (2012) got a little bit lower results on SA from the previous researches by using NLP techniques and ML methods on political news data. When all of this studies are considered, SVM can been chosen as most efficient ML method for SA.

In this study, some successful feature selection methods such as information gain and chi-square were employed to select most representative features from the corpus. In other words, effect of feature selection methods on SA is discussed to present comparative results. To construct the corpus Turkish movie reviews were used and, SVM and NB used as classifier.

2. PROPOSED METHODS

2.1 Feature Selections Methods

The major activity in text mining is to classify document into the accurate category. Feature selection methods are widely used for gathering most valuable words for each category in text mining processes. They help to find most distinctive words for each category by calculating some variables on data. There are mostly employed methods such us Chi-Square, Information Gain, Gain Ratio, and Odd Ratio. In this study, two of them (information gain and chi-square) methods were employed because of their simplicity, less computational costs and their efficiency and, their formulas are shown on Eq. (1) and (2), respectively.

$$IG = \frac{a}{N} * \log\frac{a*N}{(a+c)*(a+b)} + \frac{b}{N} * \log\frac{b*N}{(b+d)*(a+b)} + \frac{c}{N} * \log\frac{c*N}{(a+c)*(c+d)} + \frac{d}{N} * \log\frac{d*N}{(b+d)*(c+d)} \quad (1)$$

$$Chi\ Square = N * \frac{(a*d-b*c)^2}{(a+c)+(b+d)+(a+b)+(c+d)} \quad (2)$$

As known in Eq. (1) and (2): *N* is Total number of sample reviews, *a* is number of documents in the positive category which contain this term, *b* is number of documents in the positive category which do not contain this term, *c* is number of documents in the negative category which contain this term, *d* is number of documents in the negative category which do not contain this term.

2.1.1 Implementing SVM and Dataset

SVM is a discriminative classifier formally defined by a separating hyperplane. In other words, it produces an optimal hyperplane which categorizes new examples by using given labeled training data (Cortes, C and Vapnik, V., 1995). SVM gets nearly 80% successes on classifying two categories and 60% for three categories on English (Agarwal, A. et al., 2011). That is why we have chosen SVM as a classifier on this study. There are several kernels and methods which are used in SVM. LibSVM on Weka (Holmes, G., et al., 1994) was used with C-SVC kernel and Linear (u.v′) option in this research.

In this study, movie review corpus was created by crawling the data source (www.beyazperde.com, 2014) and we got labeled data by owner of the reviews. For corpus operations, Ms-SQL 2012 Enterprise was used as the management tool. The system designed for SA can be summarized as it crawls reviews at first and then filtering processes are applied (stemmer and lemmatizer), after that feature selection is implemented and

finally SVM procedures run. There are some methods like character based N-gram or spell check option of lemmatizer for eliminating suffixes and prefixes to find the root of word. However, we needed more effective stemming method in this study. In fact, some reviews include different style of writing like as abbreviations or slang expressions. At this point, Turkish lemmatizer named as "Zemberek" (Akin and Akin, 2007) presents two services which give more chance to control to sensitivity of catching word roots. First option is "fix the word" and it can able to correct mistakes which are made by writers and the second option is that it can offer suitable word if the word cannot fixed.

2.2 Experimental Results and Discussion

Constructed corpus includes total 219198 reviews about 5662 different movies. All of these reviews were labeled by the people who watched film and labels include ratings changing from 0.5 to 5. We assumed that (4.0, 4.5, 5.0) as positive, (2.5, 3.0, 3.5) as neutral and (0.5, 1.0, 1.5, 2.0) as negative comments. We had only 13350 reviews in set of negative category and we assumed its size as a minimum number of reviews which should select from the remaining categories (positive, neutral). As a result total 40050 reviews were collected from three categories and used for testing and training. Thus, we used balanced data set in experiments. Word number in one review is changing from 1 to 587. Average number of words used in a review is 26. After elimination of all special characters and suffixes from words with Zemberek, roots of words were obtained and total 44 comments could not be converted by Zemberek. After completing filtering phase, "Chi-Square" and "Information Gain were implemented on review data."

Since we employ feature selection process on reviews, if any review has no meaningful words according to feature selection methods, vector of this review contain nothing. In other words, sparse vector structure was used for representing reviews in vector format and these sparse vectors were supplied as input to the SVM modeling. In first experiments binary classification (positive or negative) was applied and achieved F_1 scores are given in Table 1. As can be seen from Table 1, SVM and applied feature selection metrics produce consistent results with previous studies. Thus we can conclude that feature selection methods reduce dimension of the input space, decrease complexity and computational cost without losing performance.

Table 1. Positive and Negative Classification Results for Feature Selection Methods

	Chi Square				**Information Gain**			
# of Top Feature Chosen	# of Sparse Vector	Positive F_1	Negative F_1	Average F_1	# of Sparse Vector	Positive F_1	Negative F_1	Average F_1
250	159	0.833	0.844	0.838	174	0.833	0.844	0.838
375	85	**0.834**	**0.845**	**0.839**	86	**0.834**	**0.845**	**0.839**
500	66	0.833	0.843	0.838	69	0.833	0.843	0.838
750	40	0.832	0.841	0.837	45	0.832	0.841	0.836
875	36	0.831	0.840	0.836	37	0.831	0.840	0.835
1000	34	0.829	0.838	0.834	35	0.829	0.838	0.833
1125	33	0.827	0.835	0.831	**33**	0.827	0.836	0.831
1250	**32**	0.825	0.834	0.830	33	0.825	0.834	0.830

The major discussion point about use of feature selection metrics is that how many features should be considered, because number of features affects number of reviews which can have no vector and it means the number of reviews which are not considered in classification process. We illustrated Fig.1 to advise a threshold for the number of feature which should be selected. As can be seen in Fig.1, use of 375 features for selection is suitable for small complex models with minimum ignored reviews. Additionally, use of 1000 feature can be advised to minimize number of ignored comments.

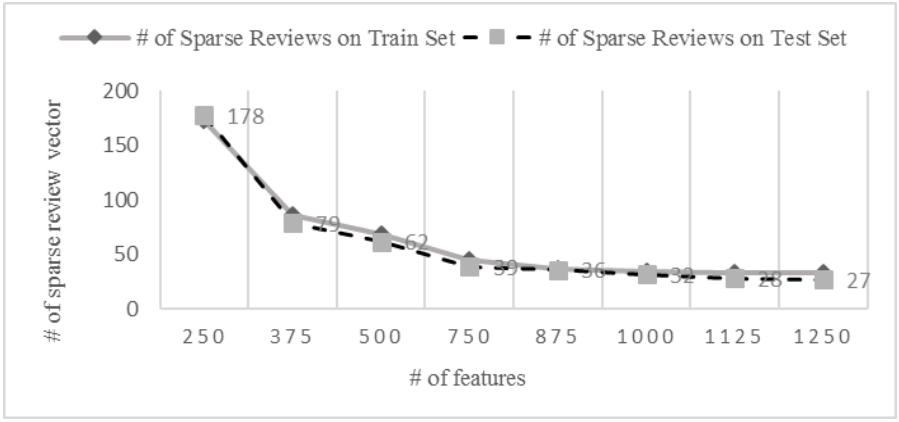

Figure 1. Change in number of sparse reviews against to different feature numbers

In the second experiments, selected features were weighted by using produced scores with employed methods and classification was performed with three classes. Additionally NB was employed in this step with SVM to observe effect of classifier method and obtained results are given in Fig.2. In this step distinct feature number had been increased from 10672 to 11679 after adding neutral reviews in modeling. According to the averaged F_1 scores, classifier has little bit effect on the classification and SVM outperforms than NB in general (Fig.2)

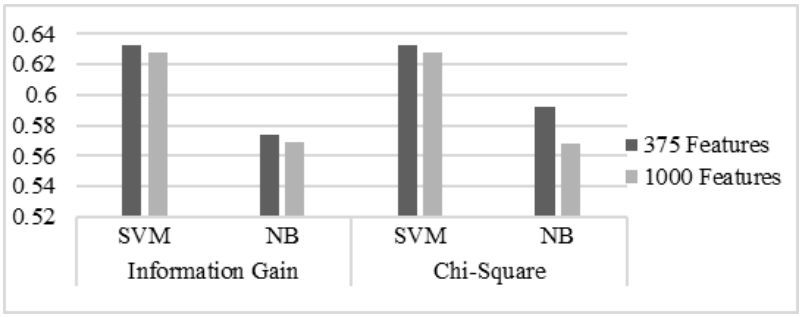

Figure 2. Average F1 score change on Feature Number

As a result, we obtained 83.9% F_1 score while classification of reviews into two categories and 63.3% for classification into three categories. These results are compatible with previous studies given in the first section and we can advise use of feature selection methods to reduce costs sourced from complexity with keeping success rates. In fact, these conclusions are not dependent on the language, because we did not use any language specific property. For this reason, proposed method is implementable for other languages.

3. CONCLUSION AND FUTURE WORKS

Currently, sentiment analysis as known as opinion mining and it is one of hot topics in research issues. In this study SA problem was addressed and use of feature selection metrics were discussed. Because, employed data in SA consists of texts and it means increasing complexity and effort while processing. Commonly used feature selection metrics such as information gain and chi-square were employed here and, SVM and NB were used as classifier. Based on the experimental results, we obtained 83.9% F_1 score for two (positive and negative) and 63.3% F_1 score for three (positive, negative, and neutral) categories. Actually, implementation of experiments on a single corpus constitutes a significant limitation factor. For this reason, we are planning to implement experiments on new Turkish corpuses as an extension of this study. Additionally, other feature selection methods except for employed ones here should be used and their effects on SA problem should be observed.

REFERENCES

Agarwal, A. et al.,2011. Sentiment analysis of twitter data.,*Proceedings of the Workshop on Languages in Social Media*,Association for Computational Linguistics

Akin, AA and Akin, MD, 2007. Zemberek, an open source NLP framework for Turkish Languages Online Available at: https://code.google.com/p/zemberek/

Becker, L., et al., 2013. AVAYA: Sentiment Analysis on Twitter with Self-Training and Polarity Lexicon Expansion.

Berger, AL., et. al., 1996. A maximum entropy approach to natural language processing. *Computational linguistics* 22.1, pp 39-71

BeyazPerde, 2014. Online Available at: http://www.beyazperde.com (Last Accessed 15 January 2014)

Boynukalin, Z.,2012,*Emotion Analysis of Turkish texts by using machine learning methods*.,Master's thesis,Middle East Technical University

Cakmak, O. et al.,2012. Using interval type-2 fuzzy logic to analyze Turkish emotion words.,*Signal & Information Processing Association Annual Summit and Conference* (APSIPA ASC),Asia-Pacific. IEEE

Cortes, C and Vapnik, V., 1995. Support vector machine. *Machine learning* 20.3, pp273-297

Cover, T., and Hart, P., 1967. Nearest neighbor pattern classification. Information Theory, IEEE Transactions on, 13, pp21-27.

Denecke, K.,2008. Using SentiWordNet for multilingual sentiment analysis.,*Data Engineering Workshop. ICDEW 2008. IEEE 24th International Conference on. IEEE*

Erogul, U.,2009,*Sentiment Analysis In Turkish.*,Master's thesis,Middle East Technical University

Esuli, A., and Sebastiani, F.,2006. Sentiwordnet: A publicly available lexical resource for opinion mining. ,*Proceedings of LREC. Vol. 6.*

Ghorbel, H., and David J., 2011, *Sentiment analysis of French movie reviews*. Advances in Distributed Agent-Based Retrieval Tools, Springer Berlin Heidelberg, pp97-108

Habernal, I. et. al., 2013. Sentiment Analysis in Czech Social Media Using Supervised Machine Learning, *WASSA*: 65

Holmes, G., et al., 1994. Weka: A machine learning workbench. In Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on (pp. 357-361). IEEE.

Kaya, M., et. al.,2012. Sentiment Analysis of Turkish Political News.,*Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*-Volume 01,IEEE Computer Society

Lewis, DD., 1998, *Naive (Bayes) at forty: The independence assumption in information retrieval*. Machine learning: ECML-98, Springer Berlin Heidelberg, pp4-15

Narr, S., Hülfenhaus, M., & Albayrak, S. (2012). Language-independent Twitter sentiment analysis. *Knowledge Discovery and Machine Learning (KDML)*, LWA.

Ozsert, CM, and Arzucan O., 2013, *Word polarity detection using a multilingual approach*. Computational Linguistics and Intelligent Text Processing, Springer Berlin Heidelberg, pp75-82

Pak, A., and Paroubek, P.,2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. ,*LREC.*

Pang, B., et. al.,2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.,*Proceedings of the 42nd annual meeting on Association for Computational Linguistics,Association for Computational Linguistics*

Pang, B., et. al.,2002. Thumbs up: sentiment classification using machine learning techniques.,*Proceedings of the ACL-02 conference on Empirical methods in natural language processing*-Volume 10, Association for Computational Linguistics

Saif, H. et. al.,2012. Alleviating data sparsity for twitter sentiment analysis.,*The 2nd Workshop on Making Sense of Microposts*

Suen, C. Y., 1979. N-gram statistics for natural language understanding and text processing. *Pattern Analysis and Machine Intelligence*, IEEE Transactions on, (2), 164-172.

Turney, P. D.,2002. Thumbs up or thumbs down semantic orientation applied to unsupervised classification of reviews.,*Proceedings of the 40th annual meeting on association for computational linguistics*,Association for Computational Linguistics

Whissell, C., 1989. The dictionary of affect in language. *Emotion: Theory, research, and experience* 4, pp113-131

Valdivia, M., et al., 2012. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. , *Expert Systems with Applications*.

Vural, AG, et al., 2013, A *Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish*, Computer and Information Sciences III, Springer London, pp437-445