Pertanika J. Sci. & Technol. 25 (S): 125 - 132 (2017)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Empirical Investigation of Feature Sets Effectiveness in Product Review Sentiment Classification

Nurfadhlina Mohd Sharef* and Rozilah Rosli

Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia

ABSTRACT

Sentiment analysis classification has been typically performed by combining features that represent the dataset at hand. Existing works have employed various features individually such as the syntactical, lexical and machine learning, and some have hybridized to reach optimistic results. Since the debate on the best combination is still unresolved this paper addresses the empirical investigation of the combination of features for product review classification. Results indicate the Support Vector Machine classification model combined with any of the observed lexicon namely MPQA, BingLiu and General Inquirer and either the unigram or inte-gration of unigram and bigram features is the top performer.

Keywords: Product review, sentiment classification, sentiment features

INTRODUCTION

There are two approaches to extract sentiment polarity automatically (Pang, & Lee, 2008); semantic-based approach (Turney, 2002) and machine learning-based approach (Pang, Lee, & Vaithyanathan, 2002). Semantic orientation provides better generality while machine learning yields maximum accuracy. Both approaches may apply different features and methods, for example semantic features (e.g.: lexical term) and syntactic features (e.g.: NGram) (Ghiassi, Skinner, & Zimbra, 2013). Support Vector Machine (SVM) and Naïve Bayes (NB) are the commonly applied machine learning-based classifiers (Pang, Lee, & Vaithyanathan, 2002; Cristianini, & Shawe-Taylor, 2000) for sentiment polarity determination.

NGram method is one of the most applied feature representation technique for sentiment analysis as it provides the probability of a term appearance in a context. Besides NGram, the Part-of-Speech (POS) is also popular. POS

ARTICLE INFO

Article history:
Received: 15 August 2016
Accepted: 18 May 2017

E-mail addresses:
nurfadhlina@upm.edu.my (Nurfadhlina Mohd Sharef),
rozilahrosli@gmail.com (Rozilah Rosli)
*Corresponding Author

ISSN: 0128-7680 © 2017 Universiti Putra Malaysia Press.

tagging is the task of identifying and assigning POS to each word or token such as nouns, verbs, adjectives and adverbs in a sentence according to the context. Besides the distributional-based approach as in NGram and POS, there are also approaches that focus on the sentiment orientation of the online review through the usage of lexicon.

The performance of the features in review datasets is still being investigated with some suggesting that the combination of unigram and bigram is the best whilst others claiming that unigram alone is sufficient (Sidorov, Miranda-jiménez, Viveros-jiménez, Díaz-rangel, & Suárez-guerra, 2012). Instead of independently carrying out lexicon-based approach or learning-based approach, there were also techniques that combined both approaches (He, Wu, Yan, Akula, & Shen, 2015). Nevertheless, none of the earlier studies have claimed their work to be the best model to identify the effective features with regards to the comparison of NGram features (unigram, bigram and combination of unigram and bigram) and lexicon approaches for the domain of product review (Ravi, & Ravi, 2015). This paper addresses this gap through an empirical investigation of the feature combinations performance in four machine learning al-gorithms namely NB, Multinomial Naïve Bayes (MNB), Sequential Minimal Optimization (SMO) or Library for Support Vector Machine (SVM) and reports the finding which is tested on product review.

This paper is composed of four sections. The first part briefly introduces the approaches for product review sentiment analysis. The second section further describes the approaches followed by the combination of the features and approaches in section three. The fourth section details the results and the paper is concluded in section five.

RELATED WORKS

Feature-based opinion extraction from product reviews has been studied by a few researchers (Cruz, Troyano, Enríquez, Ortega, & Vallejo, 2010) and various kinds of features sets and joint features have been exploited while ignoring an efficient integration of different types of features to improve the sentiment classification performance (Xia, Zong, & Li, 2011).

Several researchers have used syntactic features and NGram approach in their studies (Cruz, et al. 2010; Dang, Zhang, & Chen, 2010; Ghiassi, Skinner, & Zimbra, 2013; Kang, Yoo, & Han, 2012; Sarvabhotla, Pingali, & Varma, 2011). Go, Bhayani, and Huang (2009) contributed interesting work on a variety of topics that do not target a specific domain with the main idea of using tweets with emoticons for distance supervised learning. They used unigram, bigram, combination unigram and bigram and combination unigram with POS tags as the features. The result of their work shows that the accuracy of the classifiers improved with the use of a combination feature of unigram + bigram. Kang, Yoo, and Han (2012) also found that unigram +bigram to be an effective feature for restaurant reviews sentiment analysis.

In contrast, results from (Sarvabhotla, Pingali, & Varma, 2011) show that unigram outperforms bigram and its combination. Dang, Zhang, and Chen (2010) drew the conclusion that adding sentiment features significantly improved sentiment classification performance. Yoon, Elhadad, and Bakken, (2013) used the combination of unigram, bigram and trigram in their work, but unfortunately did not report on their effectiveness.

FEATURE REPRESENTATIONS FOR PRODUCT REVIEW SENTIMENT CLASSIFICATION

In this paper the three settings of features observed are Syntactic (Raw), Syntactic (POS) and Syntactic+Lexical, as shown in Figure 1. In Syntactic (Raw), the features are generated by the machine learning two attributes (text and class). In Syntactic (POS), the NGram features are replaced with the POS while the lexical based approach depends on the employed sentiment lexicon.

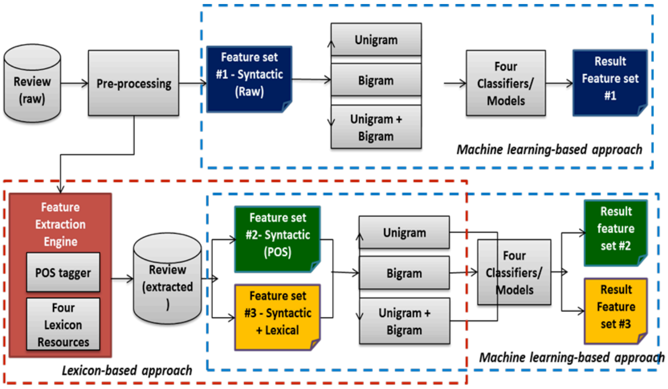

Figure 1. Feature Representation Setting

The N-Grams Features

There are three settings of NGram used here namely Unigram, Bigram and Uni-gram+Bigram. The frequency of each feature is observed and stored based on the N-Gram setting. Unigram is a type of NGram which refers to a pattern made up of one word (Kang, Yoo, & Han, 2012); in this paper a unigram for POS is represented by the adjective (JJ) label based on the POS tags. This is because adjective words express sentiment and indicates polarity either positive or negative, such as “Good”, “Excellent”, “Satisfied” and “Poor”.

Bigrams refer to the patterns that are made up of two words (Kang, Yoo, & Han, 2012). In this paper adjective phrases (AdjPs) are utilized to represent bigrams. AdjPs are built around adjectives, which indicate properties of nouns and may have adverb phrases (AdvPs) that are built around adverbs which indicate qualities of verbs (VB), adverbs (RB), and adjective (JJ). Hence, in specific the part of speech that are extracted as AdjP are RB+JJ and RB+VB such as “Very Nice” and “Looks Bad”.

Unigrams+Bigrams value is represented in terms of the frequency of JJ and AdjP. The example representations of unigram+bigram are “Good”, “Excellent”, “Satisfied”, “Poor”, “Very Nice”, “Looks Bad” which are the combination of unigram words and bigram phrases.

Lexicon

Three lexicon types used here are Bing Liu, Multi-Perspective Question Answering (MPQA)[1] and General Inquirer. Bing Liu lexicon[2] has 6789 opinion words in it word lists including 2006 positive words and 4783 negative words. The MPQA Subjectivity lexicon contains 7628 clues/words comprising of 2718 positive and 4910 negative words. Harvard General Inquirer[3] simply known as Inquirer or GI provides dictionary of entry word/word sense in spreadsheet format comprising of 4206 terms altogether, where 2291 is positive and 1915 is negative.

Syntactic (POS)

This feature considers the frequency of words that have gone through POS tagging. For uni-gram, the attributes are *Review, freqJJ* and *@@class@@*. While attributes for bigram are *Review, freqADJP* and *@@class@@* and for unigram+bigram its attributes are combination of unigram and bigram which are *Review, freqJJ, freqADJP* and *@@class@@*. The classification model setting is similar to the steps done for Syntactic (raw) except for the last part, where NGram tokenizer was not selected and the default remain for WordTokenizer.

Syntactic +Lexical (POS+Lexicon)

This refers to is the extension of Syntactic+Lexical (POS). It involves the identification of word frequency in three lexicon sources; BingLiu, GI and MPQA. So, the NGram setting is same as Syntactic+Lexical (POS) but the different is only on the attributes in the dataset. The attributes observed are:

- Unigram - *Review, freqJJ, freqJJPos<LexiconSource>, freqJJNeg<LexiconSource>* and *@@class@@*.
- Bigram - *Review, freqADJP, freqJJPosBingLiu, freqJJNeg<LexiconSource>, freqRBPos<LexiconSource>, freqRBNeg<LexiconSource>* and *@@class@@*
- Unigram+bigram - its attributes are combination of unigram and bigram which are *Review, freqJJ, freqADJP, freqJJPos<LexiconSource>, freqJJNeg<LexiconSource>, freqRBPos<LexiconSource>, freqRBNeg<LexiconSource>* and *@@class@@*.

EXPERIMENT AND RESULTS

The product review dataset[4] contains 2200 reviews (in sentences) including 11 product names. These different products consist of Canon-G3, Canon-PowerShot-SD500, Canon-S100, Creative-Labs-Nomad-Jukebox-Zen-Xtra-40GB, Creative-Labs-Nomad-Jukebox-Zen-Xtra-40GB, iPod, Linksys-Router, Micro-MP3, Nikon-Coolpix-4300, Nokia-6610 and Nokia-6600. There are 200 reviews available for each product.

[1] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

[2] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon.

[3] http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm

[4] Originates from http://www.cs.cmu.edu/~maheshj/data/acl-ijcnlp2009.html which was also used by Mahesh Joshi and Carolyn Penstein-Rosé to conduct their experiment as in [42+3].

The set of reviews that has gone through pre-processing will be the input to feature the extraction process. The features that are extracted will be processed once more and used as the training dataset during the training phase. This resulted in the generation of training models according to the classifier algorithms namely NB, MNB, SMO, LibSVM. These classifiers are chosen as they are the best, according to the literature. Each algorithm is produced based on a combination of feature type, lexicon and NGram type. There are 60 experimental cases based on the combination of the classifier, feature type, lexicon and NGram type that have been written as [Classifier]-[Feature_Type]-[Lexicon]-[NGgram_Type]. The ratio of the training and test set size is 1400:1000.

The performance of the features is evaluated according to the accuracy value. Accuracy is the proportion of the total number of predictions that were correct and calculated according to (TP +TN)/(TP+FP+TN+FN), where the variables are observed based on the confusion matrix as shown in Table 1.The metric is chosen due to it suitability for binary classification task as performed in this research.

Table 1
Confusion Matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	Total True Positive examples (TP)	Total False Negative examples (FN)
	Negative	Total False Positive examples (FP)	Total True Negative examples (TN)

Table 2 shows the comparison of the accuracy of feature representation between the Syntactic (Raw) and Syntactic (POS) observation. For Syntactic (Raw) the unigram feature performed best compared to bigram and unigram+bigram. Specifically, unigram through NB, MNB and LibSVM achieved highest accuracy compared to bigram and unigram+bigram for the same algorithm. Unlike NB, MNB and LibSVM, SMO gained the best results in unigram+bigram. Since we used accuracy to measure the effective feature, in this context we have found that unigram+bigram using SMO is the most accurate (93.5%) among all in Syntactic (raw).

Meanwhile, for Syntactic (POS), the performances of unigram, bigram and unigram+bigram are not distinct. Unigram using LibSVM has the highest accuracy compared with bigram and unigram+bigram. To find the effective feature the highest value of accuracy recorded is 92.60%, unigram using algorithm SMO is the best feature in the context of syntactic (POS) feature.

Table 2
Comparison of feature representation accuracy in Syntactic (Raw) and Syntactic (POS)

Algorithm	Syntactic (Raw)			Syntactic (POS)		
	Uni-gram	Bigram	Uni-gram + Bigram	Uni-gram	Bigram	Uni-gram + Bigram
NB	71.30	67.30	69.40	70.50	71.20	70.90
MNB	83.20	81.60	81.60	83.30	83.10	83.10
SMO	92.50	90.30	93.50	92.60	92.50	92.40
LibSVM	80.90	72.10	79.30	79.00	77.30	77.80

Table 3 shows the comparison of accuracy across the NGram and lexicon source setting in each classifier model. The discussion for the performance of NGram by this feature can be broken down into specific lexicon resources. For Bing Liu lexicon, we observe that unigram and bigram performed well in most algorithms, specifically for NB, MNB and LibSVM compared to unigram+bigram. This situation contrasts with SMO where unigram+bigram has performed better than unigram and bigram. The accuracy for unigram+bigram recorded the best. Therefore, the effective feature based on accuracy metric for Syntactic+Lexical (Bing Liu) is unigram+bigram using SMO with 92.9% accuracy.

Table 3
Accuracy in combination of NGrams, Syntactic+Lexical feature and Lexicon resources

Classifier	Unigram			Bigram			Unigram + Bigram		
	Bing Liu	GI	MPQA	Bing Liu	GI	MPQA	Bing Liu	GI	MPQA
NB	71.70	72.00	71.10	71.50	72.20	71.40	70.70	71.80	71.10
MNB	83.10	83.30	83.00	83.20	83.20	83.10	83.00	83.00	83.10
SMO	92.80	92.50	92.90	92.70	92.30	92.90	92.90	92.70	92.90
LibSVM	79.10	79.40	79.30	78.60	77.50	77.90	78.60	77.70	78.00

The General Inquirer's performance pattern has not much difference from Bing Liu. In most cases unigram or bigram recorded better results compared to unigram+bigram. This happened to the features using NB, MNB and LibSVM but not for SMO where unigram+bigram performed better than unigram and bigram. The highest accuracy obtained is 92.7%. Unigram+bigram using SMO is the best feature for Syntactic+Lexical (General Inquirer).

The performances of unigram, bigram and unigram+bigram are almost at the same level when NB, MNB and SMO are used. Unlike LibSVM, unigram performed better than bigram and unigram+bigram. However, for Syntactic+Lexical (MPQA), unigram, bigram and unigram+bigram have exactly the same accuracies (92.9%) by using SMO algorithm. This value is also the highest accuracy recorded for Syntactic+Lexical (MPQA) feature. Therefore, the best features in this context is a combination of all NGram (unigram, bigram and unigram+bigram) using SMO.

The performance level of unigram in Syntactic+Lexical feature is similar to each other regardless of the type of algorithm applied. However, with the percentage of accuracy 92.9% when applied SMO as a classifier, in the context of unigram, MPQA become the best lexicon resource. Bigram has the same performance in all the lexicon resources. However, bigram using MPQA and SMO has gained the highest accuracy, which is 92.9%. Meanwhile, the performances of unigram+bigram are not much different across the lexicon resources. Overall, the best feature of unigram+bigram for Syntactic+Lexical technique are when the Bing Liu and MPQA lexicons are used with SMO as the classifier, which has achieved the highest accuracy at 92.9%.

The results reflect that the features' strength is very similar indicating that there is no huge difference in exploiting the probability of the word distribution and the POS constituents of the words. This means that the importance of each word is almost equal, in terms of the sentiment polarity representation. Although the lexical based features logically emphasize the orientation of the polarity, the results obtained did not reflect this, as the accuracy from the best combination in Syntactic (Raw) (with accuracy 93.50% by Unigram+Bigram+SMO) and Syntactic (POS) (with accuracy 92.60% by Unigram+SMO) and Syntactic+Lexical (with accuracy 92.90% by Unigram+Bigram+SMO+BingLiu or accuracy 92.90% by Unigram+Bigram+SMO+MPQA) are very thinly different. This may be due to the ignorance of the sentiment intensity (which is beyond the scope of this paper). The performance of the algorithms is also closely linked to the features used.

CONCLUSION

This paper addresses the identification of the effective feature combination for product review. Although there are several works that have used syntactic, lexical and its combination for product review, the best combination has not been explored yet.

We found the best feature for Syntactic (Raw) is Unigram +Bigram with SMO, the best feature for Syntactic (POS) is Unigram with SMO while all NGram types with MPQA and SMO and Unigram +Bigram with Bing Liu are considered the best features for Syntactic+Lexical. Thus, Unigram + Bigram can be seen as an effective feature for product review sentiment classification. SMO was also found to return the best performance for our product review dataset with all feature sets, outperforming NB, MNB and LibSVM. Meanwhile, MPQA and Bing Liu are better lexicon resources for product review compared to General Inquirer, most likely because both Bing Liu and MPQA have more sentiment words in their dictionary (where MPQA has 7628 terms, Bing Liu has 6789 sentiment terms) compared to General Inquirer (has 4206 terms.)

Future work should investigate if the same performance applies for different languages and domain.

REFERENCES

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machine and other Kernel-based learning method. *Cambridge University Press*, March 2000.

Cruz, F. L., Troyano, J. A., Enríquez, F., Ortega, F. J., & Vallejo, C. G. (2010). A knowledge-rich approach to feature-based opinion extraction from product reviews. *In Proceedings of the 2nd international workshop on Search and mining user-generated contents, ACM*, (pp. 13-20).

Dang, Y., Zhang, Y., & Chen, H. (2010). A lexicon enhanced method for sentiment classification: An experiment on online product reviews. *Intelligent Systems, IEEE, 25*(4), 46-53.

Ghiassi, M., J. Skinner., & Zimbra, D. (2013). Twitter Brand Sentiment Analysis: A hybrid system using n-gram analysis and Dynamic Artificial Neural Network. *Expert Systems with Applications*.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1-12.

He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Inf. Manag, 52*, 801–812.

Kang, H., Yoo, S. J., & Han, D. (2012). Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications, 39*(5), 6000-6010.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

Ravi K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Syst.*, Jun. 2015.

Sarvabhotla, K., Pingali, P., & Varma, V. (2011). Sentiment classification: A lexical similarity based approach for extracting subjectivity in documents. *Information Retrieval,14*(3), 337-353.

Sharma, A., & Dey, S. (2012). A document-level sentiment analysis approach using artificial neural network and sentiment lexicons. *ACM SIGAPP Applied Computing Review, 12*(4), 67-75.

Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., ... & Gordon, J. (2012, October). Empirical study of machine learning based approach for opinion mining in tweets. In *Mexican international conference on Artificial intelligence* (pp. 1-14). Springer, Berlin, Heidelberg.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th annual meeting on as-sociation for computational linguistics.* Association for Computational Linguistics, (pp. 417-424).

Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences, 181*(6), 1138-1152.

Yoon, S., Elhadad, N., & Bakken, S. (2013). A practical approach for content mining of Tweets. *American Journal of Preventive Medicine, 45*(1), 122-129.