Investigation of Content-based Features in Sentiment Analysis Performance

Nurdhiya Hazwani Binti Helmee
Intelligent Systems Research Group,
Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia,
Malaysia
d33yaawani3@gmail.com

Nurfadhlina Binti Mohd Sharef
Intelligent Systems Research Group,
Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia,
Malaysia
nurfadhlina@upm.edu.my

***Abstract*— This study focuses on the evaluation of semi-supervised learning on machine learning-based classification approaches (Naïve Bayes and SVM) against the unsupervised learning based on feature extraction of the syntactic and lexicon frequency for sentiment analysis (SA) of product review dataset. This study utilizes the Content-Free and Content-Specific feature which involve the Syntactic and Lexical feature type comprising of Unigram, Bigram, Adjective and Adverb features. Existing data sets on product review for the training and testing of the model are used. The Naïve Bayes and SVM classifiers are used. The results demonstrate that with suitable selection of features the Multinomial Naive Bayes algorithm performs reasonably well based on F-Measure and accuracy and at times matches the popularly believed superior performance of SVM in SA task. It is discovered that the best feature is Unigram JJ+POS combined with all the lexical resources in this study which is MPQA, Bing Liu and General Inquirer and applied in Multinomial Naïve Bayes algorithm.**

Keywords—Sentiment Analysis; Content based Features; Machine Learning

I. INTRODUCTION

There are several formal review sites which allow users to drop their point of view for certain products and services experienced such as CNet, Amazon, CitySearch, Merchant Circle, Yelp, Expedia, Google Places, TripAdvisor, Social Mentions and others. Additionally, blogs and online social network such as Facebook and Twitter also have become a medium to disclose the reviews. However, when faced with tremendous amounts of online information from the Internet, information seekers usually find it very difficult to yield accurate information that is useful to them.

In fact, the customer's reviews contain opinions which are very useful for quality improvement of products or services and may lead to good profit of the products or service providers. The review and mining of opinions involves classifying sentiment as positive or negative and is typically described as sentiment analysis [1]. Sentiment analysis techniques can be classified into several categories which are (i) reviews orientation (known as sentiment classification and requires machine learning approaches), (ii) determining reviews subjectivity (known as subjectivity classification and requires language processing), and (iii) strength of review orientation (which is typically expressed as a score according to the sentiment class) [2].

Machine learning approach treats the sentiment classification problem as a topic-based text classification (Liu, 2007). In previous sentiment classification research, few studies have adopted both Content-Free and Content-Specific approach. The content-free approach involves syntactic and lexical features while the content-specific approach represents word n-gram such as unigram and bigram (Dang et. al, 2010). To the best of our knowledge, there is no consensus whether combination of content-free and content-specific approach is better from the content-free feature or content-specific approach alone. Therefore, this study investigates the performance of the features for sentiment classification to identify and understand the impact of the feature utilization.

The rest of the paper is organized as follows. The second section describes the features of sentiment analysis while the third part of the paper details the steps performed in the sentiment analysis. The results are explained in section four followed by the conclusion in section five.

II. SENTIMENT ANALYSIS FEATURES

In sentiment analysis studies four types of explicit features have been used, namely syntactic, lexical, link-based, and stylistic features. Syntactic attributes are the most common set of features for sentiment analysis. Syntactic attributes contain word n-grams [3], [4], part of-speech (POS) tags [5], and punctuation.

Since reviews typically contain phrase patterns, POS tag is combined with the n-gram pattern [6], [7]. Phrase patterns like 'n+aj' (noun followed by positive adjective) usually denote positive sentiment orientation while 'n+dj' (noun followed by negative adjective) often express negative sentiment [6].

The identification of the adjectives orientation is by incorporating lexical resources such as SentiWordNet [8], MPQA [9], Bing Liu [10] and General Inquirer (GI) [11]. Typically the frequency of each feature will be utilized as the inputs for the supervised machine learning models such as Naïve Bayes (NB), Maximum Entropy and Support Vector Machine (SVM) [3], [12].

Link/citation analysis is applied in link-based features for detecting sentiment for web and documents. [13] demonstrated that opinion web pages had linking to each other. Link-based features have been used in limited studies. So the effectiveness of them for sentiment analysis is unclear.

The stylistic features contain structural and lexical attributes which are used in many previous stylometric/authorship works [14]. Lexical and structural style markers have been used in limited sentiment analysis study. [15] applied hapaxlegomena (unique/once occurring words) for subjectivity and opinion perception. They found that the presence of unique words in subjective text is higher than objective document. [16] applied lexical features like length of sentence for classification of feedback surveys. Lexical style markers (words per message, and words per sentence) were used in [17] for analysing of web blog. Previous studies [18]–[20] have been shown style markers to be highly common in web discourse.

III. Reviews Feature Extraction For Sentiment Classification

The sentiment classification task (as shown in **Fig. 1**) consists of several steps where the main activity is towards extraction of the features from the reviews. The first step is the tokenization which is applied to segment text the by splitting it by spaces to form a bag of words. Short forms such as "don't", "I'll", "she'd" will remain as one word.

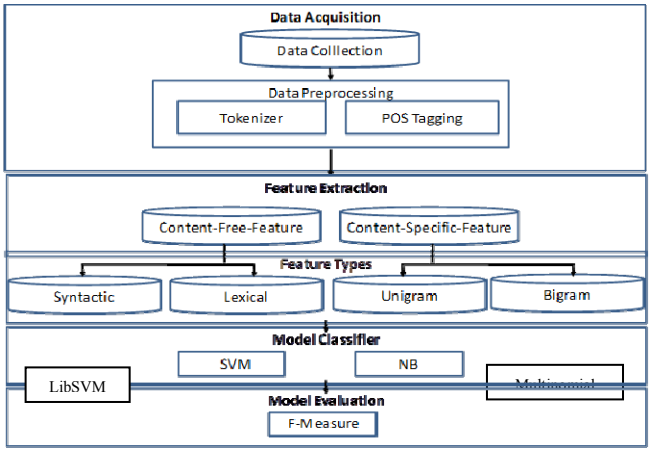

Fig. 1 Content-based Sentiment Analysis Framework

The second step is the extraction of the content-free and content-specific features. The Weka tool is used to extract the unigram and bigram values as the lexical feature. A unigram can be thought of as a window placed over a text, such that we only look at one word at a time. A bigram can be thought of as a window that shows two words at a time. The example of bigrams are 'be amazed', 'be happier', 'beautiful design' and 'completely takes'.

The Part of Speech (POS) tagging incorporated to extract the values for syntactic. The POS is described as the assignment of appropriate grammatical classes (e.g., JJ for adjective, RB for adverb, N for noun) to each word in natural language processing. The Stanford POS tagger [21] is adopted to obtain the POS tag for each word. The POS is also used to filter the adjectives which are then matched to the lexical resources. TABLE 1 and TABLE 2 show the extracted unigram and bigram respectively.

TABLE 1. Example of Extracted Unigram

JJ Positive	JJ Negative	RB Positive	RB Negative
excellent	Damaged	quickly	not
decent	Hard	Widely	never
intuitive	Loose	relatively	
creative	Dusty	suprisingly	

TABLE 2. Example of Extracted Bigram

RB Positive + JJ Positive	RB Negative +JJ Positive	RB Positive + JJ Negative	RB Negative + JJ Negative
Relatively intuitive	Not creative	Widely decent	Never damaged

The frequencies of the features are then fed as the input to train the classification model. Several features are used such as the number of the characters, and the frequency of adjectives and adverbs according to the orientation of sentiments as listed in three lexicons which are MPQA, BingLiu and General Inquiror. In this study the SVM, LibSVM, Multinomial and NB are used as the classifier and this process is performed using Weka. Finally the performance of the features as implemented in the classifiers is monitored to compare the accuracy of the chosen classifiers.

IV. Evaluation of the Performance of the Content Based Approaches

The dataset used for the evaluation is sourced from the Amazon's Apex and Canon G3 product review. The evaluation of the models is based on the standard classification performance metrics which are precision, recall and F-measure. These are used in the previous information retrieval and text classification studies (Dang et. al, (2010)) . The formula of evaluation metrics as described below where Class 1 as Positive and Class 2 as Negative. For each threshold, we evaluate the model with the Accuracy, Precision, Recall and F-Measure to ensure that this process can provide legitimate contribution to an information seeing system and meet the user satisfaction. F measure is also a commonly used performance assessment. The F measure is usually the most important performance evaluation tool in the text mining area. Therefore, in this study, we focus on F-measure to compare the percentage of each feature set.

$$\text{precision}(i) = \frac{\#\text{ of correctly identified instances for class } i}{\text{total }\#\text{ of instances identified as class } i}$$

$$\text{recall}(i) = \frac{\#\text{ of correctly identified instances for class } i}{\text{total }\#\text{ of instances in class } i}$$

$$F\text{-measure}(i) = \frac{2 \times \text{precision}(i) \times \text{recall}(i)}{\text{precision}(i) + \text{recall}(i)}$$

TABLE 3 provides the result of Content-Free and Content-Specific features and details on features of Syntactic, Lexical, Unigram and Bigram where the bold values indicate the best result for the investigated feature. Based on the result, the F1 feature yielded the best compare to F2 for NB, MNB and SMO. However, this result show that F1 which is only Content-Free feature which include the Syntactic and Lexical features obtained high result rather than F2 for the combination of Content-Free and Content-Specific feature of Syntactic, Lexical, Unigram and Bigram features for Naïve Bayes, Multinomial Naïve Bayes and SMO but LibSVM depicts that F2 is better than F1. Multinomial NB (MNB) yielded the best result compare to other classifier with 87.79% for Accuracy and 87.70% F-Measure for F1 compare to other classifier. For F2, Multinomial NB also obtained highest result of 66.93% Accuracy and 66.63% F-Measure compare to other classifier. MNB works well for data that can easily be counted. In this study the frequency is focused on Adjective (JJ) and Adverb (RB) whilst NB has strong feature independent assumption which suitable for large dataset (Hall , 2009). The MNB is simple and can be trivially scaled for large numbers of classes, unlike discriminative classifiers. It is generally robust even when its assumptions are violated. Being a probabilistic model, it is very easy to extend for structured modeling tasks, such as multi-field documents and multi-label classes. The main limit to the use of MNB is its strong modelling assumptions. The same simplifying assumptions that make it efficient and reliable also result in a severe performance gap when compared to discriminative classifiers such as SVM.

TABLE 3. **Results**

Set	**Feature Type**	**Feature**		**Classifier (10-folds cv)**							
				NB		**Multinomial NB**		**LibSVM**		**SVM**	
				F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy
F1	Content Free Feature	Syntactic and Lexical		72.60	72.64	87.70	87.79	33.50	50.14	64.90	64.93
F2	Content Free-Feature + Content Specific Feature	General Inquirer	Unigram + POS	67.30	67.29	67.10	67.29	33.50	50.14	64.79	64.79
			Unigram(RB) + POS	66.50	66.57	**67.40**	67.57	33.50	50.14	64.79	64.79
			Unigram(JJ) + POS	67.30	67.29	67.10	67.29	33.50	50.14	64.90	64.86
			Bigram + POS	59.20	59.29	62.40	62.79	33.50	50.14	59.80	59.79
			Bigram(RB, JJ) + POS	**67.30**	**67.29**	67.10	67.29	33.50	50.14	64.80	64.79
		MPQA	Unigram + POS	68.30	**68.36**	68.00	68.21	34.10	50.43	65.40	65.36
			Unigram(RB) + POS	66.60	66.64	67.40	67.64	33.50	50.14	65.10	65.07
			Unigram(JJ) + POS	67.90	67.93	**68.10**	68.29	34.10	50.43	65.40	65.36
			Bigram + POS	62.90	63.00	62.00	62.79	34.80	50.71	61.20	61.21
			Bigram(RB, JJ) + POS	**68.30**	65.40	68.00	68.21	34.10	50.43	65.40	65.36
		BingLiu	Unigram + POS	67.20	67.21	67.70	67.93	33.70	50.21	65.40	65.36
			Unigram(RB) + POS	66.30	66.29	66.90	67.07	33.50	50.14	65.40	65.43
			Unigram(JJ) + POS	67.40	67.36	**68.10**	68.29	33.70	50.21	65.60	65.57
			Bigram + POS	63.30	63.36	63.60	64.29	34.40	50.57	62.10	62.14
			Bigram(RB, JJ) + POS	**67.20**	67.21	67.70	67.93	33.70	50.21	65.40	65.36
		GI+MPQA+BingLiu	Unigram + POS	67.20	67.29	68.00	68.21	37.80	51.86	65.40	65.36
			Unigram(RB) + POS	67.20	67.21	67.00	67.21	33.50	50.14	65.60	65.64
			Unigram(JJ) + POS	66.40	66.50	**68.40**	**68.57**	37.80	51.86	65.10	65.07
			Bigram + POS	63.60	63.86	62.60	63.43	41.30	53.57	61.50	61.50
			Bigram(RB, JJ) + POS	**67.20**	67.29	68.00	68.21	37.80	51.86	65.40	65.36

The highest lexical resources is Unigram JJ+POS feature with 68.40% for the combination of lexical resources (MPQA+GI+Bing Liu) with Multinomial Naïve Bayes and Figure 6 (Accuracy), the highest lexical resources is Unigram JJ+POS feature with 68.57% for combination of lexical resources (MPQA+GI+Bing Liu) for Multinomial Naïve Bayes.

V. Conclusion

This study focuses on the investigation of the performance between the content free and content specific feature for sentiment analysis of product reviews. The feature types investigated are Syntactic, Lexical, Unigram and Bigram which are applied in the Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers. For the future, the effectiveness of Unigram + Bigram feature will be investigated since using only bigrams the feature space is very sparse. The strength of opinions identification could also be explored in the future study. Besides, the performance of other lexical resources such as SentiWordnet 3.0 to identify the polarity and strength for each lexicon

References

[1] N. M. Sharef and M. Y. Shafazand, "An Improved Deep Learning Appproach for Sentiment Mining," in *Fourth World Congress on Information and Communication Technologies (WICT 2014)*, 2014.

[2] N. M. Sharef and F. Haghanikhameneh, "Content-Based Analysis Method for Sentiment Scoring in Microblogging Mining," in *Frontiers in Artificial Intelligence and Applications: New Trends in Software Methodologies, Tools and Techniques*, 2014, pp. 398–414.

[3] B. Pang, L. Lee, H. Rd, and S. Jose, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in *Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 79–86.

[4] B. Pang and L. Lee, "A sentimental education," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, 2004, p. 271–es.

[5] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings Of The Tenth ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*, 2004, pp. 168–177.

[6] Z. Fei, J. Liu, and G. Wu, "Sentiment classification using phrase patterns," in *In Proceedings of the 4 th IEEE International Conference on Computer Information Technology*, 2004, pp. 1147–1152.

[7] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques," in *Third IEEE International Conference on Data Mining*, 2003, pp. 427–434.

[8] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*, 2010, vol. 0, pp. 2200–2204.

[9] "MPQA Opinion Corpus | MPQA." .

[10] B. Liu, "Sentiment Analysis and Opinion Mining," *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, May 2012.

[11] V. Milea, N. M. Sharef, T. Martin, R. J. Almeida, U. Kaymak, and F. Frazincar, "Prediction of the MSCI Euro Index Based on Fuzzy Grammar Fragments Extracted from European Central Bank Statements," in *International Conference on Soft Computing and Pattern Recognition 2010*, 2010, pp. 231–236.

[12] J. Khairnar and M. Kinikar, "Machine Learning Algorithms for Opinion Mining and Sentiment Classification," *Int. J. Sci. Res. Publ.*, vol. 3, no. 6, pp. 1–6, 2013.

[13] M. Efron, "Cultural orientation: Classifying subjective documents by cociation analysis," in *In Proceedings of the AAAI Fall Symposium Series on Style and Meaning in Language, Art, Music, and Design*, 2004, pp. 41–48.

[14] O. De Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," *ACM SIGMOD Rec.*, vol. 30, no. 4, p. 55, Dec. 2001.

[15] E. Riloff, J. Wiebe, and T. Wilson, "Learning subjective nouns using extraction pattern bootstrapping," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* -, 2003, vol. 4, pp. 25–32.

[16] M. Gamon, "Sentiment classification on customer feedback data," in *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*, 2004, p. 841–es.

[17] G. Mishne, "Experiments with mood classification in blog posts," in *Proceedings of the 1st Workshop on Stylistic Analysis of Text for Information Access*, 2005.

[18] A. Abbasi, "Applying Authorship Analysis to Extremist-Group Web Forum Messages," *IEEE Intell. Syst.*, vol. 20, no. 5, pp. 67–75, Sep. 2005.

[19] J. Schler, M. Koppel, S. Argamon, and J. Pennebaker, "Effects of Age and Gender on Blogging," in *In Proceedings of the AAAI Spring Symposium Computational Approaches to Analyzing Weblogs.Menlo Park, CA*, 2006, pp. 191–197.

[20] Y. Hu, H. Li, Y. Cao, L. Teng, D. Meyerzon, and Q. Zheng, "Automatic extraction of titles from general documents using machine learning," *Inf. Process. Manag.*, vol. 42, no. 5, pp. 1276–1293, Sep. 2006.

[21] K. Toutanova and C. D. Manning, "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger," in *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, 2000, pp. 63–70.