Preprocessing and Feature Selection Approach for Efficient Sentiment Analysis on Product Reviews

Monalisa Ghosh and Gautam Sanyal

Abstract In the recent years opinion mining plays an important role by business analyst before launching a product. Opinion mining mainly concerns about detecting and extracting the feature from various opinion rich resources like review sites, discussion forum, blogs and news corpora so on. The data obtained from those are highly unstructured in nature and very large in volume, therefore data preprocessing plays an essential role in sentiment analysis. Researchers are trying to develop newer algorithm. This research paper attempts to develop a better opinion mining algorithm and the performance has been worked out.

Keywords Information retrieval · Web data analysis · Preprocessing · Opinion mining · Feature selection · N-gram model

1 Introduction

Sentiment analysis also known as opinion mining is the process of determining the emotional tones behind a series of words, in recent years it has been receiving a lot of attention from researchers. This field has many interrelated sub problems rather than a single problem to solve, which makes this field more challenging. Sentiment classification task is performed for several reasons like to indicate the ups and downs of overall attitude of a brand or product, to pull out customer feedback on some topic or brand, to compare the attitude of one product with another. Sentiment analysis performed not for marketing purpose only, it has been useful in other areas also politics; law/policy making; social media monitoring etc. Many online sites including epinions.com, amazon.com, ebay.com, are highly depends on customers

M. Ghosh (✉) · G. Sanyal
Department of Computer Science and Engineering, National Institute
of Technology, Durgapur, West Bengal, India
e-mail: monalisa_05mca@yahoo.com

G. Sanyal
e-mail: nitgsanyal@gmail.com

S.C. Satapathy et al. (eds.), *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, Advances in Intelligent Systems and Computing 515, DOI 10.1007/978-981-10-3153-3_72

feedback for product evaluation. Properly identified reviews present a baseline of information that indicates ideal levels and supports the business intelligence [1].

This research paper discuss the pre processing [2, 3] technique which applied on structured as well as unstructured data to perform sentiment classification task in a significant way. Our work aims to explain how to get preprocess the online reviews in order to identify the sentiment and determine the sentiment polarity [4, 5] whether it is positive or negative [5–7].

The rest of the paper is structured as follows. Section 2 presents the existing works which can relate with our approach. Then Sect. 3 describes the various pre-processing techniques which required to process huge volume of user generated content. Section 4 explains in detail the feature selection techniques that are found to be suitable. Sections 5 and 6 deals with the classification technique. Section 7 presents details about experimental setup and elaborate the results also. Section 8 concludes the proposed method by providing a summary of the work.

2 Related Work

Sentiment Analysis is the field where many studies have been carried out on Sentiment-based Classification. Sentiment classification technique can help researchers to identify first whether a text is subjective or objective and then to determine whether the subjective text contain positive or negative sentiments. There are mainly two approaches considered from previous sentiment classification studies, *machine learning approach* and *lexicon based approach* or *semantic orientation approach* [6, 8, 9]. Both approaches have their own advantages and disadvantages and here we discuss some related works with these methods as well as their combined approaches.

The task of automatic word classification according to their polarity was may be the first major attempt by Hatzivassiloglou and Mc Keown (1997); the Wall Street Journal corpus was used instead of internet to determine whether a word was positive or negative [1].

Jiang et al. [10] work on a novel approach by using WordNet. All features for polarity mining extracted and stored in seed list, then checked with extracted opinion word from reviews. With the help of WordNet, whenever a synonym is matched it stored in seed list with same polarity while with antonym it is stored with opposite polarity and consistently seed list keeps on increasing and updated.

It means that one can estimate the polarity of a word according to its connected conjunction and adjective whose polarity is known by using WordNet or Senti-WordNet [11]. But now, no lexicon could cover the whole words and its semantic orientation, because different words have different semantic orientation in different contexts.

Therefore, machine learning approach was a great deal for sentiment classification problem. Machine learning method of sentiment classification can build more features and effectively change the dataset [12–14].

Pang Bo et al. [12] applied machine learning technique with statistical feature selection methods for the data of various field like product review and movie review. They claimed that Machine learning algorithm performed very well on sentiment classification. In Zha et al. [15] categorized the positive and negative opinions on the aspects from the pros and cons of the reviews. To determine the customer opinion in the context of free text reviews, they trained the classifier by pros and cons reviews. They evaluated the performances of three supervised classifiers SVM, NB, ME using the evaluation matrices F1-Measure, which defined as the harmonic mean of precision and recall. SentiView tool [3] used as an interactive visualization system and it focuses on analysis of public sentiments for popular topics on the Internet. Uncertainty modeling and model-driven adjustment is combined in SentiView, it mines and models the changes of the sentiment on public topics, by searching and correlating frequent words in text data.

Furthermore, the statistical method TF-IDF has been successfully applied in text classification. It can evaluate how the word is important for a file set. Kim et al. [7] proposed a novel approach to use only the term frequency part of TF-IDF, as an unsupervised weighting scheme which assigned an adjusted score to each term. The comparison is done against traditional TF-IDF weighting scheme on multiple benchmark and the proposed method gives better results.

Govindarajan [13] found that there are the disadvantages for the SVM classifier that performed not up to the mark for small dataset. Then proposed to combine both classifier SVM and NB which known as Hybrid approach to perform sentiment classification tasks, with the aim of efficiently integrating the advantages of the NB and SVM.

Dang et al. [9] proposed a lexicon enhanced method for sentiment classification by incorporating both approaches Machine learning and semantic orientation into single framework. Specifically, they combined sentiment features with content- free and content-specific features used in the existing machine-learning approach.

Therefore, our previous work [1] was based on unsupervised linguistic method for classifying sentiment of online product reviews at sentence level. SentiWordNet used to calculate the overall sentiment score of each sentence, and after summing up all opinion score the review can be classified either positive review or negative review.

A large number of research papers [16–18] published in the field of sentiment analysis with novel ideas as well as new techniques. As we discussed about previous research work it's clear that most of the work mainly focuses on identifying the sentiment orientation of the text, where very few consider data pre-processing and feature selection as to improve the accuracy. This work proposed an efficient data pre-processing and feature selection technique to get better accuracy.

3 Methodology

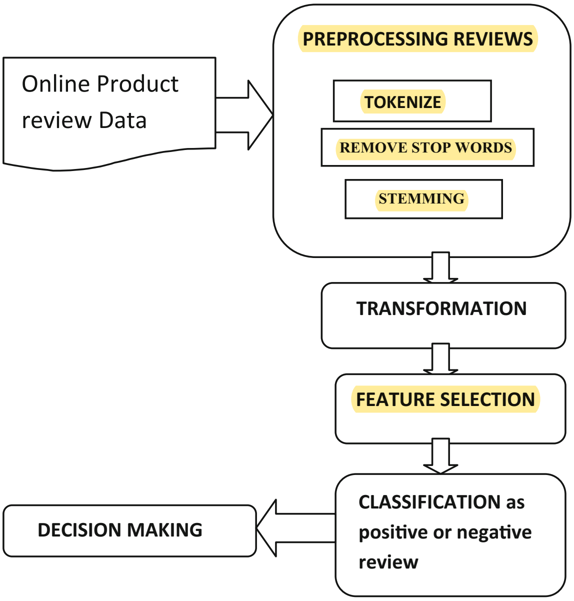
The reviews of a specific product collected from online sources usually those online data contains a lot of noisy text as well as uninformative data. Then we consider the dataset to prepare for sentiment classification. The whole process can be summarized into few steps: online text cleaning, white space removing, tokenization, removal of stop words, stemming, negation handling, noun phrase selection and then feature selection. This section presents the architecture and functional details of our proposed study. Figure 1 show the architecture of our proposed system, which consists of different functional components.

3.1 Preprocessing Phase

3.1.1 Tokenization or Segmentation

In the pre-processing phase, tokenization can be done by splitting documents (crawled reviews) into a list of words. Reviews are scanned to extract tokens consisting of words as well as numbers and after that documents are ready to be used for further processing.

Fig. 1 Architecture of proposed framework for sentiment classification



3.1.2 Removal of Stop Words

Stop words are very common and high frequency words that must be filtered to enhance performance of feature selection algorithm. The stop words removal method reduces dimensionality of the data sets, then the rest of the key words in the review corpus can be easily identified by the automatic feature extraction techniques. Some of the high frequency stop words (prepositions, irrelevant words) like "a", "of", "the", "I", "he", "she", "at", "it", "about", "and" etc. These words don't carry any sentiment information are generally known as 'functional words'. In our experiment we remove stop words for reducing file index size without effecting of user's accuracy level.

3.1.3 Stemming

Stemming is one of the essential parts of preprocessing phase during feature extraction. It is the process converts all the words of the text into their stem, or root form. Stemming is a fast and simple approach which makes the feature extraction process easier. The basic stemming process works like 'automatic,' 'automate,' and 'automation' are each converted into the stem 'automat'. The Porter's stemmer is a popular stemming algorithm for English language. The basic Stemming process can transform the words in following way (Table 1).

4 N-Grams

An n-gram language model becomes popular in researchers because of its simplicity and scalability.

N-gram model mainly works to find a set of n-gram words from a review document. Models those are generally used are 1-gram sequence or unigrams where n = 1, 2-gram sequence or bigrams where n = 2 and 3-gram sequence or trigrams where n = 3 and the sequence can be extended. The following example can define n-gram model in better way

Example— Text Data: "Something is better than nothing."

(n = 1) Unigrams: "something", "is", "better", "than", "nothing".

(n = 2) Bigrams: "something is", "is better", "better than", "than nothing".

(n = 3) Trigrams: "something is better", "is better than", "better than nothing".

Table 1 Transforming word into base form

List of Words	Stem form
Played, plays, playing	*Play*
Argue, argued, argues, arguing.	*argu*

5 Feature Selection

Feature selection is a process where we run through the corpus before the classifier has been trained and remove any features that seem unnecessary.

To perform sentiment classification, at the starting we considered a large no of words, terms or phrases as features that may express opinion. But very few of them actually express positive or negative opinion so we used several methods to filter those features very efficiently to improve accuracy. We applied feature weighting method with certain threshold for the targeted features. The different features weights for a feature set discussed below:

5.1 Feature Presence (FP)

Feature Presence is nothing but to check whether the feature appears in the text or not. Multiple occurrence of the same feature are ignored. We get a vector of binary values like 1 for each feature that presence in the document otherwise the value becomes 0. Feature Presence used by many researches for sentiment classification and Pang et al. [12] were first to use this method.

5.2 Feature Frequency (FF)

This method used in sentiment classification is one of the simplest methods to represent a document with a vector. Feature value is the number of times that feature occurs in the document. For an example, if the word "zoom" appeared in a document 14 times, the associated feature would have a value of 14. Sometimes many high-frequency features are very weak to distinguish the document according to low frequency features.

5.3 Term Frequency Inverse Document Frequency (TFIDF)

The TF-IDF weight is a statistical measure used to evaluate how important a word is to a document in a collection. TF-IDF value mainly consists of two scores one is term frequency and another one is inverse document frequency. Term frequency actually counts how frequent the term has appeared in a given document and inverse document frequency is calculated by dividing the total number of documents by the number of documents that a given term occurs in. Tf-IDF for each feature is calculated by using the following formulae.

$$\mathrm{TF}-\mathrm{IDF}=\ f/n.(-\log_2 \mathrm{N_f}/\mathrm{D}) \tag{1}$$

Where f defines the frequency count of the term in the review document of size n. $\mathrm{N_f}$ is the number of review documents contains the term, and D is the total number of documents in the database.

6 Classification

Supervised learning methods are commonly applied for Sentiment analysis or text classification problem to classify the opinion as positive or negative. Sentiment classification problem is generally of two types one is binary sentiment classification with positive and negative classes and another one is multi-class sentiment classification. We work on binary sentiment classification task with the use of the Naive Bayes classifier and Support Vector Machines for classifying the review documents.

6.1 Naive Bayes (NB)

Naive Bayes classifier one of the simple probabilistic classifier which is based on the applying Bayes theorem. The Naive Bayes classifier uses a feature vector matrix to determine whether a document is under positive classes or negative classes. The probability of a class c given document d is estimated using bayes' rule as follows.

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \tag{2}$$

where P(c) is the prior probability of any document in class c, P(dlc) is the probability of a document d given it is in class c, and P(d) is the probability of the document d.

6.2 Support Vector Machine (SVM)

Support vector machines classifier is a standard non probabilistic binary classifier; it can classify data as either linear or nonlinear. These classifier includes some property like high dimensional feature space, sparse instance vector etc. SVM classifier finds a maximum margin hyperplane which used to separate the d-dimensional data perfectly into its two classes. SVM performs more efficiently to compare with others Naïve Bayes; Maximum Entropy classifier for almost all combination of features.

7 Experiments and Results

In order to evaluate our proposed method for preprocessing dataset and feature selection to perform sentiment analysis. All the experiments are performed based on the following dataset prepared by collecting online reviews of digital camera.

7.1 *Dataset Preparation*

In this section the online customer reviews are crawled from target review site (amajon.com, ebay.com, epinion.com) and store locally after filtering markup language tags. We extract the review of different types of digital cameras like Nikon, Sony, Fuji Film, Canon etc. To evaluate our proposed work we also used some publicly available corpus for product review polarity dataset [19] and a popular corpus on Amazon Product Review Data (more than 5.8 million reviews). We present the whole database it shows the no of items, no of customer reviews, no of sentences per review [1] (Table 2).

7.2 *Performance Evaluation*

We used precision, recall, F1 measures as evaluation matrices [1] to evaluate the experiment. **Precision is** the ratio of true positives among all retrieved instances; **recall is** the ratio of true positives among all positive instances and **F1 measure** is the combination of precision (π) and Recall (ρ).

$$\textbf{Precision}\,(\boldsymbol{\pi}):\ \frac{TP}{TP+FP} \tag{3}$$

$$\textbf{Recall}\,(\boldsymbol{\rho}):\ \frac{TP}{TP+FN} \tag{4}$$

Table 2 Initial review dataset on camera

Product	Reviews	Average line/review	Unique feature
Canon EOS40D	269	9	112
Nikon coolpix 4300	129	4	42
Nikon D3SLR	117	7	56

Table 3 Performance results obtained for feature selection techniques

	TF-IDF		FF		FP	
	$R^{(bp)}$	$R^{(ap)}$	$R^{(bp)}$	$R^{(ap)}$	$R^{(bp)}$	$R^{(ap)}$
Accuracy	63.23	78.49	71.12	78.32	76.337	81.5
Precision	61.29	76.34	68.86	76.66	75.33	80.21
Recall	73.21	78.21	72.21	79.31	79.21	81.66
F-Measure	72.43	79.92	71.77	77.96	77.86	82.72

$$\textbf{F1}: \quad \frac{2TP}{2TP + FP + FN} \tag{5}$$

The above table presents the classification accuracies on both not pre-processed and preprocessed data for each of the features matrices (TF-IDF, FF, FP). The column $R^{(bp)}$ refers to before preprocessing and $R^{(ap)}$ refers to the result after applying preprocessing method (Table 3).

8 Conclusion

Sentiment analysis is one of the most challenging fields which involves with natural language processing. It has a wide range of applications like marketing; politics; news analytics etc. and all these areas are benefited from the result of sentiment analysis. In our research we focused on to applying different preprocessing techniques to remove the noise and irrelevant features from the text for reducing feature space, while at the same time trying to improve the accuracy in sentiment orientation task. We considered two popular classifiers Support Vector Machine and Naïve Bayes in the context of sentiment classification where as the performance of SVM classifier is much better to compare with Naïve Bayes classifier for almost all combination of features, although previous research had already identified the same. The proposed method in this research paper is just an initial step towards the improvement in the techniques for sentiment classification. In future we aim to include more feature selection technique like Information Gain and Chi Square method to find optimal feature subset and the result will make the classifiers more efficient and accurate.

References

1. Ghosh, M., Kar, A.: Unsupervised Linguistic Approach for Sentiment Classification from Online Reviews Using SentiWordNet 3.0. International Journal of Engineering Research and Technology (IJERT). Vol. 2, no. 9 (2013) September (2013).

2. Haddi, E., Liu, X., Shi, Y.: The Role of Text Pre-processing in Sentiment Analysis. Procedia Computer Science. Vol. 17, 26–32 (2013).
3. Wang, C., Xiao, Z., Liu, Y., Xu, Y., Zhou, A., Zhang, K.: SentiView: Sentiment Analysis and Visualization for Internet Popular Topics. Human-Machine Systems. IEEE Transactions. Vol. 43, no. 6 (2013).
4. Hatzivassiloglou, V., McKeown, K.R.: Predicting the semantic orientation of adjectives. Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics. 174–181 (1997).
5. Subrahmanian, V.S., Reforgiato, D.: AVA: Adjective-verb-adverb combinations for sentiment analysis. Intelligent Systems. IEEE 23(4), 43–50 (2008).
6. Turney, P.D.: Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.417–424 (2002).
7. Kim, Y., Zhang, O.: Credibility Adjusted Term Frequency: A Supervised Term Weighting Scheme for Sentiment Analysis and Text Classification. Sentiment and Social Media Analysis. Proc. Of the 5th Workshop on Computational Approaches to Subjectivity. 79–83. Baltimore, Maryland, USA (2014).
8. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: SentiFul: A Lexicon for Sentiment Analysis. IEEE Transactions on Affective Computing. Vol. 2, no.1, 22–36 (2011).
9. Dang, Y., Zhang, Y., Chen, H.: Lexicon Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews. Intelligent Systems. IEEE vol. 25, no. 4, 46–53 (2010).
10. Jiang, P., et al.: An approach based on tree kernels for opinion mining of online product reviews. In Data Mining (ICDM), IEEE 10th International Conference, 256–265 (2010).
11. Esuli, A., Sebastiani, F.: SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Proceedings from International Conference on Language Resources and Evaluation (LREC). Genoa (2006).
12. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification Using Machine Learning Techniques. ACL Press. In. Proceedings of the Conference on Empirical Methods in Natural Language Processing ACL Press. 79–86 (2002).
13. Govindarajan, M.: Sentiment Classification of Movie Reviews Using Hybrid Method. International Journal of Advances in Science Engineering and Technology, Vol. 1, no. 3 (2014).
14. Bollegala, D., Weir. D., Carroll, J.: Cross Domain Sentiment Classification using a Sentiment Sensitive Thesaurus. IEEE Transactions on Knowledge and Data Engineering. Vol. 25, no. 8, 1719–1731 (2012).
15. Zha, Z.J., Yu, J., Tang, J., Wang, M., Chua, T.S.: Product aspect ranking and its applications. IEEE Transaction of Knowledge and Data Engineering. Vol. 26, no. 5 (2014).
16. Fang, X., Zhan, F.: Sentiment Analysis Of Product Review Data. Journal Of Big Data. a Springer Open Journal (2015).
17. Arshad, S., Yaqub, N., Inayat, M.: Sentiment Classification Of Product Reviews At Different Level: A Survey. Proceedings of the 2nd International Conference on Engineering & Emerging Technologies (ICEET). 26–27. Lahore (2015).
18. Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J.: Interpreting the Public Sentiment Variations on Twitter. IEEE Transactions on Knowledge and Data Engineering. vol. 26, no. 5 (2014).
19. http://www.cs.cornell.edu/people/pabo/product-review-data.