ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/105144, November 2016*

Utility of Corpus based Approach in the Recognition of Opinionated Text

Hina Gupta* and Nitasha Hasteer

Department of Information Technology, Amity University, Noida - 201313, Uttar Pradesh, India;
guptahina189@gmail.com, nhaster@amity.edu

Abstract

Objectives: The proposed work focuses on mining opinion word catalogue by using corpus based approach. The motive is to use different parts of speech to improve the classification of the sentiments. **Methods/Statistical Analysis**: The methodology involved in the proposed work incorporates both the sentiment orientation approach and machine learning approach. The various features like content – specific, content-free and other sentiment features have been used to classify the sentiments. The previous works in the field involved only some specific parts of speech, which have been replaced by the usage of nouns, adjectives, verbs and adverbs. In this approach an algorithm for calculation of sentiment feature has been proposed. **Findings**: The algorithm proposed in this work is more efficient in comparison to other existing work. In this work, since we have developed a corpus based approach amalgamating both the machine learning and semantic orientation approaches into a common skeleton, it improvises the classification method. Our projected method also incorporates the content-specific and content free features involved in the existing approaches. It also utilizes the infrequent and sentiment features in the semantic orientation approach. The proposed technique can be classified into three main modules: Acquiring of data, generation of features, followed by classification and evaluation. **Application/ Improvements:** The researches to be done in future can deal with other feature generation methods. Moreover the method can also be improved by making the modifications so that the feature classification can be done on quite large data sets. The method can further be implemented for multilingual languages to build a multilingual sentiment-based lexicon..

Keywords: Corpus–based Approach, Opinionated Text, Sentiment Analysis, Sentiment Classification, Sentiment Features

1. Introduction

In today's era of social networking, there has been a huge demand of the content generated by user. Micro blogging sites have millions of people contributing their thoughts and judgment frequently because of its distinguishing feature of short and simple manner of expression. People are getting involved in and exchanging opinions and ideas through online social activities. They do so by posting their views in the form of tweets, blogs and using various other web forums. Thus the amount of such data is increasing on the web. This data is of utmost importance for both the product and service provider. The users use this data for making their decision regarding the purchase of the product.

1.1 Problem Description

Various approaches for sentiment classification can be involved in studying the information about the opinion and sentiment on the web. This can be done by finding and analyzing the data consisting of opinions and emo-

*Author for correspondence

tions to conclude whether a text is objective or subjective or it contains positive or negative sentiments[1]. The various methodologies that have been employed in previous works, to accumulate or gather the word list of opinion words are: Dictionary based or corpus-based approaches[2]. WORDNET's synsets are used to gather opinion words in Dictionary based approach. Although this succeeds in having a good collection of words but fails to identify opinion words which are context dependent[3]. On the other hand, corpus-based approach, finds the context dependent opinion words by depending on the syntactic or co-occurrence patterns in large corpora[4]. Features can be classified into frequent features and infrequent features. Frequent features are the most talked about and refers to the features from any entity's review, context-specific and content free features[5]. Some studies focus on features which are not common and generally talked about. Such features are termed as infrequent features[6]. These features are of interest only to some potential customers[7]. The method proposed in this work amalgamates both the dictionary-based and corpus-based techniques involving both the frequent and infrequent features into a single methodology to get better on sentiment classification and analysis[8].

1.2 Sentiment Classification Approaches

Generally sentiment classification is a directional based approach on emotions and opinions about oneself. The major motive of the sentiment analysis is to classify the sentiment as positive or negative. The problem can further be generalised on classifying the sentiment as subjective/opinionated or factual/objective. Various approaches have been utilized earlier such as semantic orientation, machine learning, Naive Bayesian algorithm and Support Vector Machine are amongst the most common algorithms that have been utilized for the purpose of classification. In comparison to the machine learning approach the semantic orientation approach does not require any prior training. It performs classification based on the positive and the negative phrases. There are many different techniques used in semantic orientation approach. The two mostly used techniques are 1. Dictionary based approach and 2. Corpus based technique. The corpus based method focuses on finding the co-occurrence patterns of phrases to extract their sentiment. Diverse methodologies have evolved to find out sentiments. For instance, in[9] evaluated the semantic orientation of a phrase as the communal rank among the phrase and the word "excellent" (optimistic polarity) minus the communal rank among the phrase and the word "poor" (pessimistic polarity). In[10] utilized a process of bootstrapping to acquire linguistically affluent patterns of prejudiced expressions, to classify amongst subjective and objective expressions. Initializing a group of objective patterns extracted from preceding literature, the process utilized an algorithm of pattern mining to find out probable subjective patterns. The gathered patterns were then employed to make a decision on whether an expression was subjective or in contrast to it. The other technique based on Dictionary, operated on antonyms, synonyms and other hierarchies in Word Net or other lexicon to resolve word sentiments[11].

2. Features Involved in Sentiment Classification

The features used in sentiment classification can be bifurcated into the following three categories namely: 1. Sentiment features, 2. Content-free features and 3. Content-specific features. Amongst the above stated features the content-free features include structural features, lexical features and syntactic features[11,12]. The lexical features include lexical features based on character, measurement of richness of vocabulary and lexical features based on words. It is basically a measurement of the variation of character based or word based lexical. Syntactic features generally deal with the patterns used for sentence formation. It includes various parts of speech, words and punctuations. The organization of text and other layout architecture are covered under the structural features. The use of file extension, sizes, different fonts, patterns, colours are included in the technical feature set of the structural features. The important keywords and phrases come under the roof of content specific features, for example word n-grams. These features are cooperative in the improvisation of classification of text. On considering it can be seen that small group of words or phrases

Table 1. Different part of speech used by different researchers

Reference	Part of Speech Used
(13)	Adjective
(14)	Adjective
(4)	Adjective and adverb
(10)	Adverb, adjective, noun, verbs

have been used in the determination of the sentiments. The Table 1 shows different researchers and the part of speech used by them for determining the sentiment. The stated features are not sufficient to learn about the concept. Therefore, the process of feature selection can be applied. It focuses on the identification of minimal-sized features subset. The method of feature selection results in the generation of different attributes from the feature domain and their evaluation on the basis of some assessment criteria. This results in the identification of the best subset of feature.

3. Motivation

The two major approaches of sentiment classification are machine learning and sentiment orientation. Each of them has its own pros and cons. The machine learning approach is better in accuracy due to its training set which makes it domain dependent than the semantic orientation approach. The main drawback of this method is that training has to be done in case of new data set. On the contrary, no prior training is required in case of semantic orientation approach making it domain independent. This turns semantic orientation approach into a generalized approach but lacks behind in accuracy. The corpus based techniques for semantic orientation is dependent on a huge corpus to evaluate the statistical inference needed to calculate the semantic orientation for each phrase. Opinion words dependent on context are obtained using corpus based approach. On the other hand a fine or superior lexicon is essential for dictionary based technique. Some studies conducted earlier dealt with the improvement of performance of sentiment classification using both the approaches. The motivation for our work came from the above studies done in the field. In this work we have developed a corpus based approach amalgamating both the machine learning and semantic orientation approaches into a common skeleton. Our projected method incorporates the content-specific and content free features involved in the existing approaches. It also utilizes the infrequent and sentiment features in the semantic orientation approach.

4. Design and Implementation

The design of our projected method is based on corpus based approach consists of the following three modules: 1. Acquiring or gathering of data, 2. Generation of features and 3. Classification and assessment.

4.1 Acquiring or Gathering of Data

This module deals with the collection of datasets. These datasets are then parsed and stored in a database. Our application domain is online reviews of products. We selected this as our domain because of its increasing

importance in today's scenario. These reviews influence a person's purchasing decision regarding a product. At this point we have just proposed an algorithm.

4.2 Generation of Features

This section focuses on three types of features that are used in our proposed methodology. It includes Sentiment Features (SF), Content-Specific Features (CSF) and Content Free Features (CFF). From the three features, two features CSF and CFF are from machine learning approach and SF is from sentiment oriented approach. The Table 2 depicts the various feature utilized by CSF and CFF features.

- Now, for each synset the score of average polarity is calculated. This calculation is done separately for each synsets using the prior-polarity formula.

Score (word = POS) I = (k∈SentiWordNet (word_POS&polarity))(SentiWordNet_Score(k)i)

(mod (synsets (word_POS)))

Where, POS € {noun, adjective, verb, adverb} and i € {positivity, negativity, objectivity} k € synonym set of a word in a given intellect.

Table 2. The features utilized by the CSF and CFF features

Feature	Includes
Content Specific Features [CSF]	Unigrams and bigrams
Content Free Features [CFF]	Lexical features, structure features and syntactic features

4.3 Extraction of Features and Calculation of Sentiment Score

- Each sentence should be parsed to yield the part-of-speech tag of each word (POS). This would specify whether the word is a noun, adjective, verb or adverb etc.
- Sentiment score should be calculated for the extracted part-of-speech tag in the obtained in the above step. For this calculation SENTI-WORDNET can be used. There are sets of synonyms of the all the words used in the language. These are called synsets. For each synset, the SENTI-WORDNET provides three sentiment values, 1. Positivity, 2. Negativity and 3. Objectivity.

In the remaining section we will consider score as 'sc', word as 'wd', objective as 'O', positivity as 'P' and negativity as 'N'

4.4 Strategy for Sentiment Feature-Calculation

In order to sort out uncommon subjective words, use mid score on the scale of 0-1 score scale, which turns out to be 0.5

The following algorithm can be employed to acknowledge to which category of "PON" the word belongs to. The algorithm is as follows:

```
If (sc (wd=POS) O > 0.5) then
        {
            word = " objective"
        }
          Else
            {
            If ((sc (wd=POS) P) > (sc (wd=POS)N))then
              {
                word = " positive"
              }
             Else
              {
                word = " negative
              }
           }
```

Rule out the words for which the negative and the positive scores are equal as they lack to illustrate comprehensible polarity inclination.

4.5 Classification and Assessment

To scrutinize the classification based on corpus-based approach, we can use SVM to contrast the performances of diverse feature sets. For every assessment set, arbitrarily select 90% of data for training and the residual 10% as assessment data. Recapitulate the performance values under the attribute names of overall accuracy, average recall, average precision and average F-measure for all the given experimental data. Amongst the different features: CFF and SF is domain independent and CSF is domain dependent.

To evaluate the performance of these features, we craft three varied feature sets in increasing way.

- CFF alone consists of content-free features.
- Feature set CFF+CSF consists of content-free and content-specific features.
- SF+CSF+CFF consist of sentiment features, content-specific features, and content-free.

Here CFF, SF are domain independent, therefore the feature sets CFF, SF+CFF are domain independent and the feature sets CFF+CSF, SF+CSS+CSF are domain dependent. At the time when the quantity of feature is big, feature selection can be improvised by choosing the most favourable subset of features. Hence by constructing three feature sets as CFF, CFF+CSF, CFF+CSF+SF the effectiveness of the proposed sentiment classification can be done. Given below is the flow diagram that depicts the projected corpus based approach Figure 1.

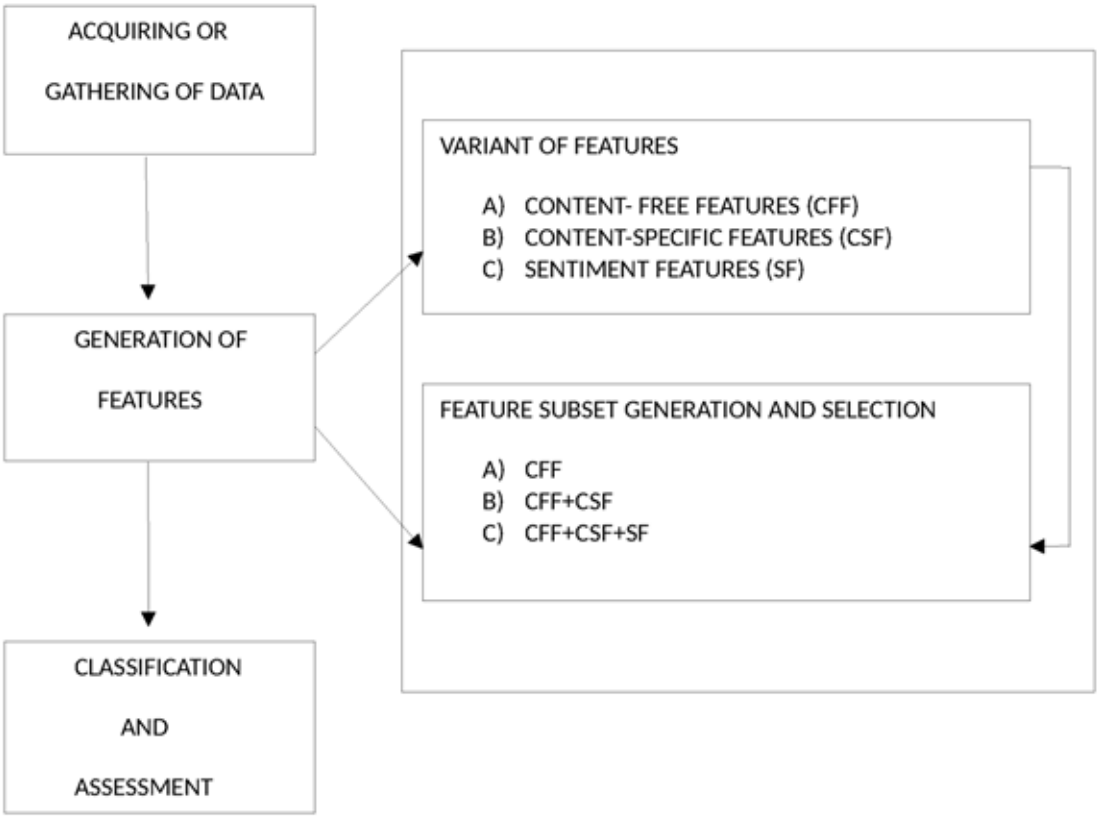

Figure 1. Flow diagram of corpus based approach for sentiment classification.

5. Conclusion

The methodology and algorithm proposed in this work employed a corpus-based approach to produce sentiment features. This also included infrequent features to improvise the effectiveness of classification methods. The researches to be done in future can deal with other feature generation methods. Moreover the method can also be improved by making the modifications so that the feature classification can be done on quite large data sets. Regarding the future scope of this work, the algorithm will be applied by us on real world data set available through amazon.com, epinoins.com, dpreview.com, etc. The method can further be implemented for multilingual languages to build a multilingual sentiment-based lexicon in future.

6. References

1. Dash M, Liu H. Feature selection for classification. Intelligent Data Analysis. 1997 Mar; 1:131–56.
2. Kim SM, Hovy E. Determining the sentiment of opinions. Proceedings of the 20th International Conference on Computational Linguistics (COLING); USA. 2004 Jan. p. 1–8.
3. Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), USA. 2002 Jul; 10:79–86.
4. Turney PD. Thumbs Up or Thumbs Down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meetings of the Association of Computational Linguistics; USA. 2002 Dec. p. 417–24.

5. Riloff E, Wiebe J. Learning extraction patterns for subjective expressions. Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2003 Jul. p. 105–12.
6. Turney PD, Littman LM. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transaction Information Systems. 2003 Oct; 21(4):315–46.
7. Razavi AH, Inkpen D, Matwin S, Uritsky S. Offensive language detection using multi-level classification. Springer Berlin Heidelberg; 2010 May-Jun. p. 16–27.
8. Wiebe J, Wilson T, Bruce R, Bell M, Martin M. Learning subjective language. Computational Linguistics. 2004 Sep; 30(3):277–308.
9. Abbasi A, Chen H. Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems. 2005 Sep; 20(5):67–75.
10. Using SentiWord Net for Multilingual Sentiment Analysis 2008. Available from: http://ieeexplore.ieee.org/document/4498370/
11. Esuli A, Sebastiani F. Senti-WordNet: A publicly available lexical resource for opinion mining. Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-06); 2006. p. 417–22.
12. Devitt A, Ahmad K. Sentiment polarity identification in financial news: A cohesion-based approach. Proceedings of 45th Annual Meeting of the Association of Computational Linguistics; 2007 Jun. p. 984–91.
13. Hatzivassiloglou V, Wiebe J. Effects of adjective orientation and gradability on sentence subjectivity. Proceedings of the 18th International Conference on Computational Linguistics, USA. 2000; 1:299–305.
14. Hu M, Liu B. Mining and summarizing customer reviews. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge discovery and Data Mining; USA. 2004 Aug. p. 168–77.