Work Integrated Learning Programmes

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Information Retrieval (SS ZG537)

Assignment

Second Semester 2018-2019

Enhancement Bag-of-Words Model for Solving the Challenges of Sentiment Analysis

Submitted By:

Nilesh D. Ghodekar (2018ht12544)

Problem statement

Sentiment Analysis is a fundamental task in Natural Language Processing (NLP). Its uses are many: from analysing political sentiment on social media, gathering insight from user-generated product reviews or even for financial purposes, such as developing trading strategies based on market sentiment . The goal of most sentiment classification tasks is to identify the overall sentiment polarity of the documents in question, i.e. is the sentiment of the document positive or negative?.

Sentiment analysis has challenges such as Subjectivity and Tone, Comparisons, Emojis etc.

Sentiment analysis involves identification of sentiment meaning, expressions, Polarity and expressions strength, and their relationship to the subject. The volume of linguistics resources is enormous. The bag-of-words (BOW) models evaluation uses many techniques such as Naive Bayes (NB), Support Vector Machine (SVM) and Maximum Entropy (ME) classifiers that have been exhibited to go well in the binary positive negative sentiment classification tasks on document-level datasets like movie reviews.

For sentiment analysis Bag-Of-Words model is most widely used. The bag-of-words model is very simple to understand and implement and offers a lot of flexibility for customization on your specific text data. But bag-of-words has two major limitations: using a manual evaluation for a lexicon in determining the evaluation of words and analysing sentiments with low accuracy because of neglecting the language grammar effects of the words and ignore semantics of the words.

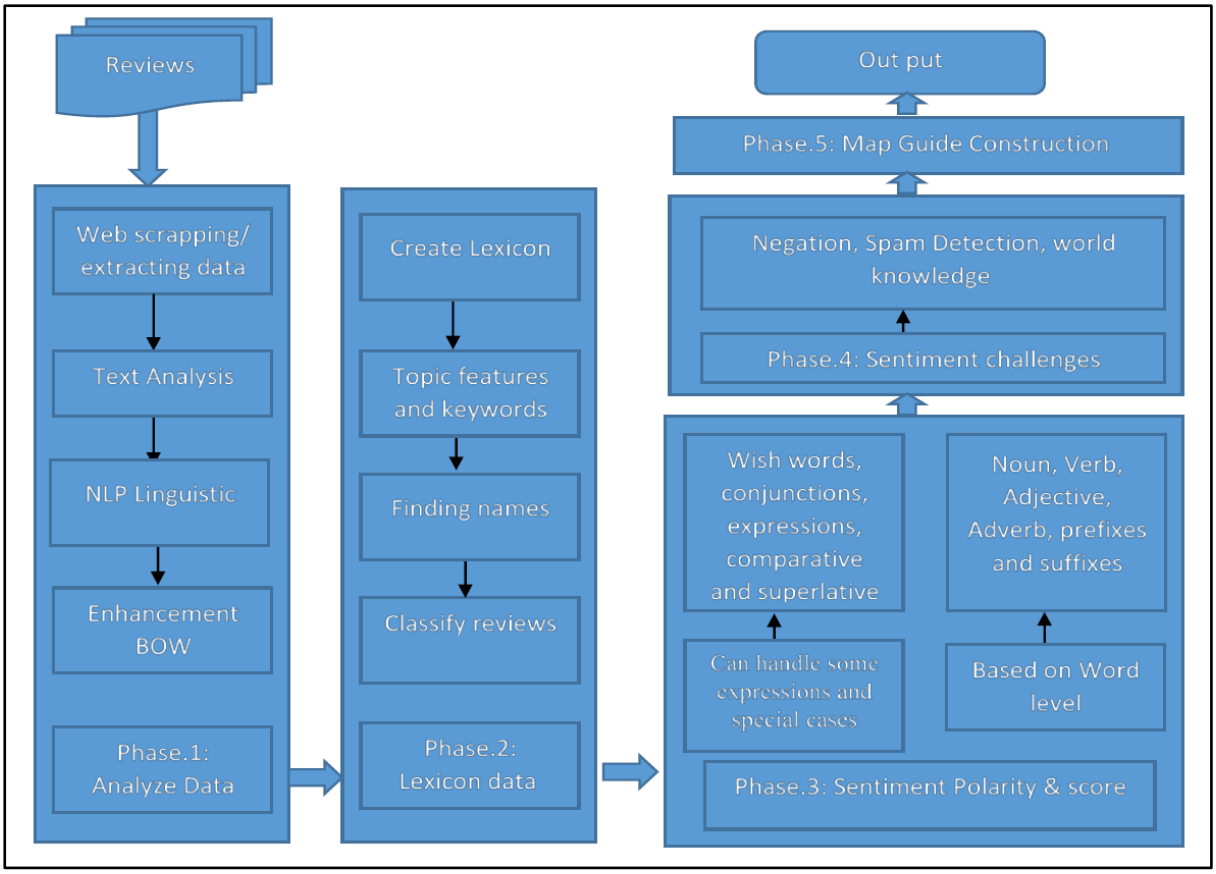
Solution approach:

A new technique to evaluate online sentiments in one topic domain and produce a solution for some significant sentiment analysis challenges that improves the accuracy of sentiment analysis performed called Sentiment analysis of online papers "SAOOP" is proposed. The proposed technique relies on the enhancement bag-of-words model for evaluating sentiment polarity and score automatically by using the words weight instead of term frequency. This technique also can classify the reviews based on features and keywords of the scientific topic domain.

The proposed lexicon is constructed automatically which is based on hierarchical database model to give the correct scores with respect a topic features and keywords. The new lexical approach uses for saving time and ease searching process for each word.

Although SAOOP aims at the evaluation of words but it can handle some cases of expressions and phrases with respect the order of each word. SAOOP also computes the total score of each review by calculating the aggregate score of review words. For measure accuracy, the comparison between our proposed enhancement BOW technique and the standard BOW model on the scientific domain is made.

Architecture: *SAOOP Architecture*



The architecture has five phases for reaching to sentiment score. The input is online reviews.

1. The phase one called "Analyze Data" which includes some functions as web scraping and extracting data, text analysis, NLP linguistics, and Enhancement BOW.
2. The phase two called "Lexicon data" illustrates that creating a lexicon, topic features and keywords, finding names, and classify sentiment reviews. This phase explains that classify reviews from one or more class in assuming five classes (Topic, citation number, the publishing date of paper, authors, and place of publications). In the second function, extract the features and keywords of scientific domain as the names of authors, the names or shortcuts of conferences and journals. Finding names declares that how to recognize some names of each class.
3. In the third phase entitled "sentiment analysis score and polarity", the proposed technique can detect the polarity based on one of sentiment classification (very negative, negative, neutral, positive, and very positive).
4. The fourth phase is the solutions of sentiment challenges. This depends on the word level, it contains proposed solutions to deal with some challenges "Spam and fake detections", "Implicit and Explicit Negation", and "World knowledge" based on topic features.
5. Last phase is creating a map guide which is based on the sentiment scores related to the most related papers according to keywords and fields classification. The output declares in the total sentiment score and polarity of each paper

Results:

Enhancement BOW model in the proposed SAOOP technique has been shown to extremely effective, since it captures more contextual meaning based on word weight, resulting a classification accuracy of 83.5%.

This is a clear indication of the effectiveness of incorporating the impact of world knowledge, spam detection, and negation, by interesting the topic domain features and keywords and constructing the newly miniature lexicon. Although the proposed technique is based on the word-by-word model, it can understand some phrases as do not directly through caring with the classification of reviews.

Algorithm	Standard BOW	Enhancement BOW
Goal	Text analysis and give polarity for words in text	Evaluate sentiment score for reviews
Sentiment classification	2 or three classes	5 classes
Input type	Documents, text or images	Reviews
Data size	Small number of texts or review	Large number of reviews
Data set	Any scope, refer topic domain to minimize dictionary	Topic domain is the best to minimize dictionary and can extraction features and entities
Clarity	No	Yes
Efficiency	No , less accuracy and manually dictionaries	Yes, high accuracy
Memorability	No	Yes
Simplicity	Yes	Yes

The above table describes the comparison between standard BOW and enhancement BOW.

Conclusion:

The enhancement of Bag-of-Words model on online scientific papers reviews and the incorporate contextual polarity and effect of sentiment analysis challenges to improve the sentiment accuracy. SAOOP aims at evaluating for reviews of scientific papers and from scientific papers is called CiteULike website, analyzes and classifies the textual content of the sentiment reviews of each paper. The proposed SAOOP can classify sentiment reviews and visualize the relationships between them based on extract features and keywords of scientific domain.

The efficiency of the proposed algorithm improves over standard BOW algorithm.

Limitations of the paper:

This solution not suits for phrases. Also There are numerous weaknesses with the bag of words model, especially when applied to natural language processing tasks are not addressed in the paper.

Improvements Suggested:

Using TextRank above problems can be addressed. TextRank is able to incorporate word sequence information. TextRank algorithm has ability to identify multi-word phrases and summarize text.

The TextRank algorithm was introduced in 2004 by Rada Mihalcea and Paul Tarau. TextRank is a graph ranking algorithm - this simply means that nodes in a graph can be scored using information from the global graph. A well-known graph ranking algorithm is Google's PageRank. In the context of text, words are nodes/vertices and the cooccurrence of words together forms a link/relationship between the words (i.e., an edge). Mihalcea and Tarau introduce a vertex ranking algorithm that takes into consideration edge weights. TextRank can be used for keyword extraction and text summarization.