Sentiment Analysis Tool using Cosine and Jaccard Implementation

Shraddha Deshpande
Department of Computer Engineering
Smt. Kashibai Navale College of Engineering, UoP,
Pune, India.

Mrunmayee Shinde
Department of Computer Engineering
Smt. Kashibai Navale College of Engineering, UoP
Pune, India

Jonika Rathi
Department of Computer Engineering
Smt. Kashibai Navale College of Engineering, UoP,
Pune, India.

Shanu Gandhi
Department of ComputerEngineering
Smt. Kashibai Navale College of Engineering, UoP,
Pune, India.

Vaishali Deshmukh
Department of Computer Engineering
Smt. Kashibai Navale College of Engineering, UoP
Pune, India

ABSTRACT
With the evolution of web technology, huge chunk of data is available over the web for users. The users not only explore the resources present on web, but also provide feedback, thus generating additional useful information. Sentiment analysis also known as Opinion mining deals with automating the task of classifying a textual review expressed in natural language as either positive or negative. In general, supervised methods consist of two stages. Firstly is extraction of information followed by its classification. The proposed approach consists of two major algorithms jaccard and cosine to implement a tool that will carry out the process of extraction of sentiments from a textual review and rate it accordingly. In this paper, the shortcomings of the existing approach will be discussed, along with the future & directions for research.

General Terms
Cosine Similarity, Jaccard Implementation, Sentiment Analyzing tool

Keywords
Sentiment Analysis, Opinion mining, Jaccard, Cosine, Term Frequency (TF), Inverse Document Frequency (IDF)

1. INTRODUCTION
Nowadays, People refer to the reviews of a product, reviews of a movie etc before making a purchase. A product with negative reviews will be less preferred over the product having positive review. Reviews are expressed in Natural language.

The main aim of opinion mining is to analyze views of the people and use them in decision making. The World Wide Web is a huge repository of massive data that can be structured or unstructured [3,5]. Opinion Mining or Sentiment analysis tool involves building a system to explore user's opinions made in certain documents like, comments, reviews or tweets, about a product or a

topic. It aims to determine the attitude of a user about some topic. With increasing popularity of opinion-base websites and other resources new challenges has been arrived in opinion mining. It is now becoming evident that the views expressed on the web can be influential to readers in forming their opinions on some topic. Also these reviews are taken into consideration by the vendors and policy makers, providing them with scope to improve.

There are several challenges in the field of sentiment analysis. The most common challenges are given here. Firstly, Word Sense Disambiguation (WSD), a classical NLP problem is often observed to occur. For example, *"amazingly unpredictable plot in the movie"* is a positive phrase, while *"an unpredictable steering wheel"* is a negative one. The opinion word *unpredictable* is used in different senses. Secondly, addressing the problem of sudden deviation from positive to negative polarity, as in "*The movie has a great collection of songs, superb storyline and spectacular photography; the director has managed to make a mess of the whole plot".* Thirdly, unless the negations are handled properly data can completely mislead. "*Not only do I not approve dophy 1220, but also hesitate to call it a phone"* has a positive polarity word *approve*; but its effect is negated by many negations [5].

2. RELATED WORK
From time to time, many extraction systems have been developed. Human Computer Interaction can be related to study of emotion and informatics [2]. Experiments show that a person's behaviour is affected by his/her emotional state. As per paper [3] sentiment analysis has mainly three levels, Document level analysis, Sentence level analysis and Feature level analysis. In Document level analysis and Sentence level analysis one cannot identify reviewer's likes or dislikes on specific feature of thatobject. It has been found that document level and sentence level classification are not enough to identify each and every one detail about sentiments expressed in a document as sentiments may be expressed with respect to different features. In Feature level method algorithm with parts of speech tags is used to improve the accuracy on the benchmark dataset. It is fine- grained analysis process which takes every feature of object into consideration. The feature level method includes State Vector Machines (SVM) but SVM does not provide accurate results and hence must be calculated using Jaccard and Cosine implementation.

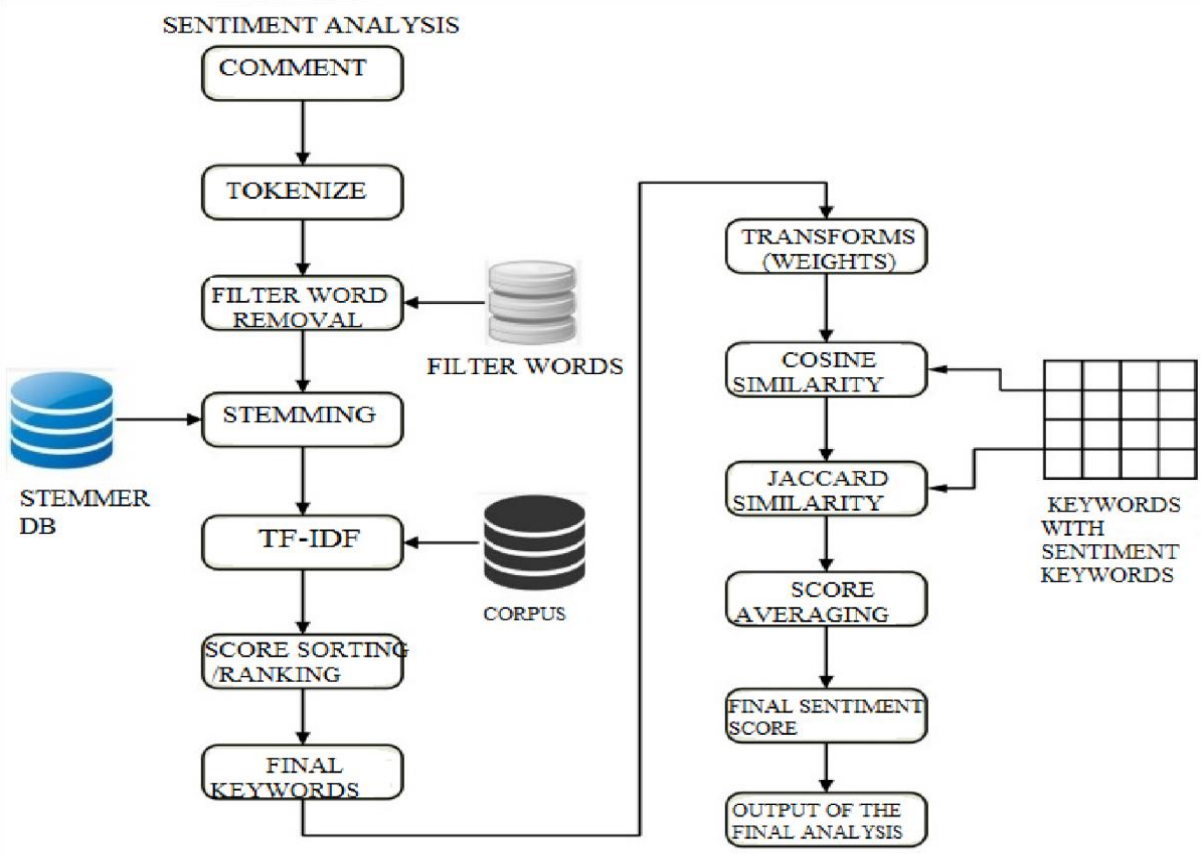

Fig 1: System Architecture

Many researchers have been addressing the problem of sentiment classification on textual reviews. Some datasets are available and have been used by many researchers in order to compare the results and the dataset of movie reviews is the most popular benchmark dataset in the literature. Since the focus of our study is on the overall opinion (positive or negative) expressed in the review [1].

3. PROPOSED WORK

The proposed framework presents an approach with similarity measure to give better accuracy and efficiency than SVM.

Similarity measure consists of three methods namely Jaccard & Dice and Cosine, any of which can be used for. It is based on the analysis of comments in a given review.

The proposed architecture for sentimental analysis is shown in figure 1. The Sentiment expressed in the textual review is analysed. The blocks represent the informative part. Firstly, comments to be analysed are tokenised by the process of tokenisation i.e. words which do not contribute in opinion extraction are removed. These include it, the, is, was and so on. Stem or root word is extracted from the tokenised words by the method of stemming. The term frequency and inverse document frequency are calculated. And on the basis of their score, the words are usually sorted. Later, the cosine similarity as well as jaccard similarity equation is applied on the score of the words to obtain the average score and produce the final review rating.

4. ALGORITHMIC APPROACH

4.1 Jaccard Similarity

a. It starts with finding important keywords in documents and removing irrelevant words.

b. A TF-IDF approach is used initially.

c. Given a document collection D, a word w, and an individual fw, D document dϵD, we calculate where, fw, d equals the number of times w appears in d.

d. |D| is the size of the corpus (data collection from twitter, blogs etc.) fw, D equals the number of documents in which w appears in D.

e. Once we get important terms in documents then similarity measure is applied.

$$JS(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

f. Finally we get a feature set depending on its similarity i.e. positive and negative.

$$wd(fw, d) * \log\left(\frac{|D|}{\quad}\right)$$

4.2 Cosine Similarity

a. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them vectors cannot be greater than 90°.

b. In Information Retrieval and text mining, each term is notionally assigned a different dimension and a document is characterized by a vector where the value of each dimension corresponds to the number of times that term appears in the document.

c. It gives a measure of how similar two documents are likely to be in terms of their subject matter. The cosine of two vectors can be derived by using the Euclidean dot product formula:

$$a.b = ||a||b||\cos\Theta$$

d. In the case of information retrieval, the cosine similarity of two documents will range from 0 to 1, since the term frequencies (TF-IDF weights) cannot be negative.

5. IMPLEMENTATION DETAILS

The proposed work will include the following different modules.

1) Collecting dataset.
2) Tokenizing
3) Pre-processing and storing domain specific keyword
4) Calculating TF-IDF.
5) Similarity measure.
6) Feature Extraction.
7) Classification and Analysis.

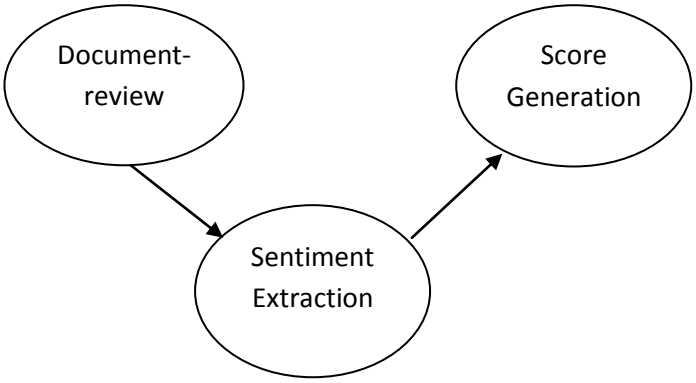

Fig 2: Abstract Flow

6. PERFORMANCE MEASUREMENT

The classification performance will be evaluated in three terms accuracy, recall and precision as defined below. A confusion matrix is used for this.

	Machine says yes	Machine says no
Human says yes	True positive	False negative
Human says no	False positive	True negative

Table 1. Confusion Matrix Table

$$\text{Accuracy} = \frac{\text{True positivereviews + True Negative reviews}}{\text{Total number of documents}}$$

$$\text{Recall} = \frac{\text{True positive review}}{\text{True positive reviews + false negative reviews}}$$

$$\text{Precision} = \frac{\text{True positive review}}{\text{True positive review+ false positive review}}$$

7. FUTURE WORK AND CONCLUSION

This paper summarizes and proposes a framework to systematically unfurl the sentiments expressed in the document collection. Based on the study on sentimental analysis with respect to SVM and similarity measure we concluded following points:

1) Methodologies under similarity measure lead to better accuracy and should be implemented in less computational complexity than SVM.
2) It should stand for maximum number of features and samples.

One of the major roadblocks in this project is dependency over the data set. Data mining deals probabilistic information filtering. A stronger data set can be designed and used so decrease the probability of errors and achieve higher accuracy. The results so obtained show that the user input parameters determine the sentiments, a score is generated suggesting the sentiment level.

8. REFERENCES

[1] Malmaz Roshanaei, Shivakant Mishra,"An Analysis of Positivity and Negativity Attributes of Users in Twitter,"IEEE /ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014) page 365-370, July 2014.

[2] Rodrigo Moraes, Joao Francisco Valiati, Wilson P. Gaviao Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," Expert Systems with Applications vol. 40, page 621-633, August 2013.

[3] Neha S. Joshi, "A Feature Dependent Method for Sentiment Analysis to understand User Context," cPGCON 2014, Third Post Graduate and Research Scholar Symposium, 2014.

[4] Sowmya Kamath S, Anusha Bagalkotkar, Ashesh Khandelwal, Shivam Pandey, Kumari Poornima, "Sentiment Analysis Based Approaches for Understanding User Context in Web Content", 978-0-76954958-3/13, 2013 IEEE.

[5] S Chandrakala, C Sindhu, "Opinion Mining And Sentiment Analysis Classification: A Survey," ICTACT Journal on Soft Computing, vol. 03, issue: 01, October 2012,

[6] G. Marreiros, R. Santos, C. Ramos, 1. Neves, "Context Aware Emotional Model for Group Decision Making," IEEE Trans. of Intelligent Sys, vol. 25, no. 2, pp. 31-39, January 2010

[7] Samad Hasan Basaria, Burairah Hussina, I. Gede Pramudya Anantaa, Junta Zeniarjab, "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machineand Particle Swarm Optimization", Procedia

[8] Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, Ma