Sentiment Analysis of IMDb Movie Review

Jayshree Sable[1]
Department of Computer engineering
MIT -Academy of engineering
India

Srushti Naikare[2]
Department of Computer Engineering
MIT Academy Of Engineering
Pune , India

Manish Shingare[3]
Department of Computer engineering
MIT Academy Of Engineering Pune ,
Pune , India

Abstract : Sentiment analysis is the process of determining the opinion or feeling from a piece of text. We are living in the "age of customer", where customer knowledge and perception is a key for running a successful business, same is the case with the film industry. In this paper, we have presented our experimental results on the performance evaluation sentiment analysis using lexicon approach. s the lexicon-based framework for sentiment classification, which classifies reviews as a positive, negative, or neutral. The proposed framework also detects and scores the slangs used in the text. The dataset was captured from IMDb movie reviews, in which there are 5000 instances, divided into the 50%-50% of positive and negative ratio for development, cross-validation and final test sets. We have also discussed implementation with the help of flowchart and code snippet.

Keywords : Lexicon ,Sentiment, Sentiment Orientation, Preprocessing, Adjective ,Dictionary

I) INTRODUCTION :

Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to extract and identify subjective information in the whole sentence. It aims to determine the attitude of the speaker or writer with respect to some topic, in our case, with respect to movies. The attitude can beHis or her judgement or evaluation.Affective state (i.e. emotional effect of writer when writing).The intended emotional communication (i.e. emotional effect the writer wishes to have on the reader).

In the last few years there has been a rise in usage of social media such as blogs and social media networks such as Instagram, twitter, etc. which has fueled the interest in sentiment analysis. Online opinion has turned into a kind of virtual currency with rapid increase of reviews, ratings, recommendations and other forms of expressions, for any business that is looking to market their products, identify new opportunities and manage their reputation. In order to automate the process of filtering the noise, understanding the conversations, identifying customers or viewers needs, many are now looking to the field of sentiment analysis. But the main problem of most of the sentiment analysis algorithms is that they use simple terms to express sentiment about a product or service. However, cultural factors, sentence negation, sarcasm, language ambiguity, use of slangs and differing context make it very difficult to turn the string of text into simple pro or con sentiment.

A fundamental task in sentiment analysis is classifying the polarity of given text. It focuses on whether the expressed opinion is positive, negative or neutral. Sometimes it goes beyond polarity and looks at emotional states such as "angry", "sad" and "happy".

Existing approaches to sentiment analysis can be grouped into four main categories. They are keyword spotting, statistical methods, and concept-level techniques. Keyword spotting classifies text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid, and bored. Statistical methods influence elements from machine learning such as latent semantic analysis, support vector machines, bag of words and Semantic Orientation. Concept-level approaches through the analysis of concepts that do not explicitly convey relevant information, but which are implicitly linked to other concepts that do so.

So our project, unlike others, aims to do sentiment analysis using lexical approach. In most of the string of text or sentences expressing the opinion about the movie, their thoughts which contain different writing habits, use of slang words, sarcasm, non-english words etc. so using lexical approach and using different libraries, we can easily counter those problems. Making use of tokenization our system would be easily identifying the main keywords in the sentences.

Then the sentence is examined for subjectivity, only the sentence with subjective expressions are retained and sentences which convey objective expression are discarded. In this manner the filtration and text preparation would do the rest work easy. Then using some common computation techniques sentiment detection would be done. Sentiment classification i.e. identification of the given sentiment is either positive or negative, is done.

II) RELATED WORK :

A lot of researchers , students and have made some remarkable observations and have published national and international papers and journals that vividly explain the importance of the algorithms used in the preprocessing of the dataset .

[1]Walaa Medhat of a School of Electronic Engineering, Canadian International College, Cairo Campus of CBU, Egypt has published (Ain Shams University) a paper based on Sentiment analysis algorithms and applications which can have three main classification levels in SA:

document-level, sentence-level, and aspect-level SA. They are presented an overview on the recent updates in SA algorithms and applications. s. After analyzing these articles, it is clear that the enhancements of SC and FS algorithms are still an open field for research.[1]

[2]Maite Taboada of Simon Fraser University has published (Simon Fraser University) A paper based on Semantic orientation (SO) is a measure of subjectivity and opinion in text. It usually captures an evaluative factor (positive or negative).From these they are presented with a word-based method for extracting sentiment from texts. This attribute to this criteria for selecting and ranking words, which include excluding ambiguous words and including fewer rather than more words.[2]

[3]Karthik Konar 1MCA Student, Dept. of Computer Engineering, NMIMS Mukesh Patel School of Technology Management & Engineering, Vile Parle(West) Mumbai has published (NMIMS Mukesh Patel School of Technology Management & Engineering, Vile Parle(West) Mumbai) in their paper which can be based on Sentimental analysis is that particular domain, where you try to understand human emotions with the help of a software. Human emotions are in written form, and we classify those sentiments as positive, negative, neutral. This paper provides a detailed comparison of various applications of sentiment analysis which was implemented using different approaches such as lexicon, machine learning, and VADER sentiment analysis.[3]

III) Proposed work :

A.Dataset

The dataset has been taken from Kaggle
Structured or unstructured data : Structured Data in CSV format

B.Dataset Description :

The data set is related to movie reviews given on IMDb platform. It doesn't specify about the movies like for which movie does that review belong.
It have two features as 1)Review, 2)Sentiment
Each review has been labeled as positive,negative or neutral which help to verify after predicting the sentiment. Here we are going to calculate the sentiment and sentiment value and then verify with given sentiment

C. BLOCK DIAGRAM :

For our project we have proposed the block diagram that explains how the data undergoes the execution in the whole program along with detailed explanation .

Dataset
It contains a record of reviews of different movies. It which contains over 5000 reviews with their respective sentiment for future analysis
Preprocessing Module
In this section preprocessing steps are carried out on a dataset. It includes data cleaning, tokenization, lemmatization which is explained later on in detail.
Text Identification
This module performs tasks where words are identified and classified with the help of dictionaries.
Dictionary
It contains sentiment resources which helps for Text Identification. These sentiment resources(dictionaries) contain positive and negative sentiment words. Some misspelled words are also included in the lexicon as they appear frequently in the social media text.

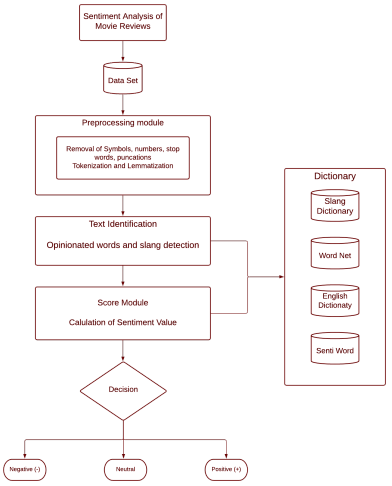

Fig . 3.C Proposed block diagram of the project

Score Module
It will help us to calculate the sentiment value of the review with the help of a dictionary which has sentiment values for words.It calculates positive, negative and neutral sentiment value a review separately and then it would calculate the final sentiment score

1. Here the dataset is first imported from the database
2. Preprocessing steps are carried out on a dataset. It includes data cleaning, tokenization, lemmatization
3. It identifies the words based on adjectives and classified with the help of dictionaries.
4. Then with the help of score module, calculate sentiment value
5. The review will be classified into positive, negative, neutral based on sentiment value

D. Proposed methodology

The selected approach for our project is "**Lexicon Based** ". Where we will be using a Dictionary based approach. The dataset has reviews of movies. The problem statement is to predict whether the review is positive, negative or neutral. It is referred to as a classification problem as we are identifying the sentiment of the review. Here we are using dictionaries which have set words classified sentiment wise and having scores which would help us to identify the sentiment of review by comparing.

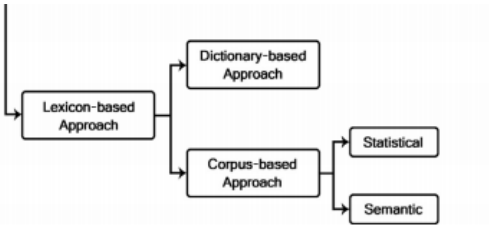

Fig 3.D The diversity of the types of problems in Lexicon

E. Algorithm analysis :

Step 1: Importing the data set
Step 2: Removing single letters and numerics from the reviews
Step 3: Snippet to tag the parts of speech for each review
Step 4: Narrow down each review with only adjectives and verb forms
Step 5: Making a list of tuples which has review and its respective class
Step 6: Separating the data set by each class i.e. "Positive", "Negative", "Neutral"
Step 7: Splitting the words in each list to get the word frequencies
Step 8: Finding the frequencies of the unique words in each class
Step 9: Identifying the similar words for each class in each review
Step 10: Calculating the score of positive, negative and neutral words in each review
Step 11: Predicting the class based on the score.

IV. IMPLEMENTATION :

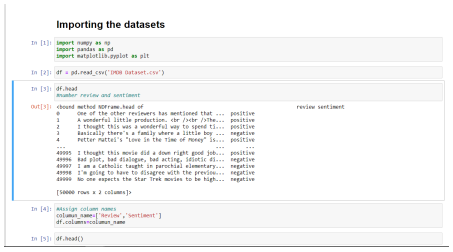

Fig 4.1 The loading of the dataset

Fig 4.2 Carrying out data preprocessing steps

Out[41]:

	POS	NEG	NEU	truth	PRE
0	1.00	0.93	0.89	pos	films adapted from comic books have had plent...
1	1.00	0.95	0.95	pos	every now and then a movie comes along from a...
2	1.00	0.96	0.98	pos	you ve got mail works alot better than it des...
3	0.91	0.94	1.00	neutral	jaws is a rare film that grabs your atte...
4	0.95	1.00	0.94	neg	moviemaking is a lot like being the general m...

Fig 4.3 Final analysis of the review

Out[43]:

	POS	NEG	NEU	truth	PRE
0	1.00	0.93	0.89	pos	pos
1	1.00	0.95	0.95	pos	pos
2	1.00	0.96	0.98	pos	pos
3	0.91	0.94	1.00	neutral	neutral
4	0.95	1.00	0.94	neg	neg

Fig 4.4 Predicting the Sentiment

V. RESULTS :

Out[43]:

	POS	NEG	NEU	truth	PRE
0	1.00	0.93	0.89	pos	pos
1	1.00	0.95	0.95	pos	pos
2	1.00	0.96	0.98	pos	pos
3	0.91	0.94	1.00	neutral	neutral
4	0.95	1.00	0.94	neg	neg

In [44]: `pd.crosstab(UL.truth,UL.PRE)`

Out[44]:

PRE	neg	neutral	pos
truth			
neg	783	85	0
neutral	62	332	52
pos	0	75	611

Fig 5.1 Sentiment analysis

Out[44]:

PRE	neg	neutral	pos
truth			
neg	783	85	0
neutral	62	332	52
pos	0	75	611

In [45]: `print accuracy_score (UL.truth,UL.PRE)`

0.863

In [46]: `print recall_score(UL.truth,UL.PRE, pos_label= 'neg', average='micro')`

0.863

Fig 5.2 Accuray of the analysis

CONCLUSION :

This paper about Sentiment Analysis of IMDb Movie Reviews using Lexicon approach. We have presented a Dictionary-based method for extracting sentiment from texts The algorithm provides a clear idea about the importance and working of the Sentiment predictions. It provides accuracy up to 86%

REFERENCES :

[1]Walaa Medhat , Ahmed Hassan , Hoda Korashy “Sentiment analysis algorithms and applications: A survey ” School of Electronic Engineering, Canadian International College, Cairo Campus of CBU, Egypt b Ain Shams University, Faculty of Engineering, Computers & Systems Department, Egypt Issued: 19 | April 2014

[2]Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Vol, Manfred Stede “ Lexicon-Based Methods for Sentiment Analysis ” Department of Computer Science, University of British Columbia, 201-2366 Main Mall, Vancouver, B.C. V6T 1Z4 Canada. Issue: 28 | September 2010.

[3]Karthik Konar , Riddhi Patel “ A Systematic Study on Sentimental Analysis and its Applications ” MCA Student, Dept. of Computer Engineering, NMIMS Mukesh Patel School of Technology Management & Engineering, Vile Parle(West) Mumbai Issue: 06 | June 2020

[4]Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau Department of Computer Science Columbia University New York, NY 10027 USA Issued: 23 | June 2011

[5]Fazal Masud Kundi, Aurangzeb Khan, Shakeel Ahmad, Muhammad Zubair Asghar “Lexicon-Based Sentiment Analysis in the Social Web ” Institute of Computing and Information Technology, Gomal University, D.I.Khan, Pakistan, Institute of Engineering and Computer Sciences, University of Science and Technology Bannu, Pakistan Issued: 26 | May 2014