IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11376

Sentiment Analysis on YouTube using Lexicon Based Approach

Arunima Mukhopadhyay[1], Sejal Patel[2], Viren Parmar[3]

Application Development Senior Analyst, Accenture, Mumbai, Maharashtra, India[1]

Senior Software Engineer, Larsen & Toubro Infotech Limited, Mumbai, Maharashtra, India[2]

MS in Computer Science, New York University, New York, USA[3]

Abstract: In today's generation, people write blogs, articles, record videos, audios, and upload it across the Internet for people across the globe to view. One of the benefits of this advancements is that now people can share their opinions directly and instantly with the artist. This large amount of feedback can be of great significance, when analyzed with necessary expertise and tools. This field is referred to as Sentiment Analysis which aims at identifying the sentiment of the text and also whether the writer has a positive opinion or a negative one. Categorization of these responses can help an artist to get insight of public review in order to take necessary steps in near future. In this paper we intend to use a lexicon based approach for sentiment analysis of comments on YouTube videos.

Keywords: Sentiment Analysis, Lexicon, YouTube and SentiWordNet

I. INTRODUCTION

Sentiment Analysis is a computational process in which the opinions expressed are analyzed and categorized to determine whether the writer's attitude towards a specific topic is positive or negative. It is a combination of text analysis, natural language processing (NLP) and computational linguistics. It is indeed difficult to correctly predict what the writer is trying to express since it may involve complex human expressions like sarcasm where we need to understand the tone to estimate that it was a sarcastic comment.Positive orientation means praise while negative orientation means criticism. Semantic orientation varies from positive to negative and degrees varies from mild to strong.

The domain in the project is restricted to YouTube comments. More than 400 hours of videos are being uploaded on YouTube every minute. Thousands of YouTubers are uploading their content over the World Wide Web. We are collecting live data from the YouTube website, as this dataset is real time.

Among different approaches towards sentiment analysis, we plan to use lexicon based dictionary approach. This dictionary consists of various words with both positive and negative sentiments. In this approach each word is mapped with its polarity and overall polarity of a text is calculated. After this some analysis of the results are shown which are explained in detail in the upcoming sections.

II. RELATED WORK

We have come across various methods that are used in sentiment analysis, as it is one of those fields that is expanding at a very fast rate. There are various algorithms that can be used for sentiment analysis which includes - Lexicon Based Approach, Naive Bayes, SVM, bags of word approach etc. The following literature survey provides methods regarding the approaches and the model that are used:

1) Lexicon Based Approach

Lexicon based approaches: Sanjida Akhter and Muhammad Tareq Aziz in their paper **'Sentiment Analysis On Facebook Group Using Lexicon Based Approach'[1]** explained about how Lexical based approaches are used for sentiment analysis and text analytics for classifying facebook comments. These approaches make use of lexicon of sentiment words i.e. a dictionary or dataset which contains all the words which lead to a sentimental orientation of text[1].

These words are associated with a particular sentiment value which leads to overall sentiment prediction of the text. Lexicon based approaches also comprise of linguistic rules which help in determining a more accurate sentimental value of the provided data.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11376

2) **Machine Learning Approach**

Machine Learning is one of the most modern and emerging fields which can greatly improve the precision of sentiment analysis. Machine learning methods are classified into two categories: Supervised Learning and Unsupervised learning. Supervised learning methods are based on training models on data whose actual outcomes are known and the prediction of outcomes for testing data set. On the other hand, unsupervised learning algorithms like 7 clustering focus generally on classification of data into different groups. Following are the different classification models used in machine learning:

- Naïve Bayes Approach

Naïve Bayes is one of the most widely used and efficient classification approaches. It follows the principle of probability to predict the category in which a particular data should be classified.

$$P(C|A_1, A_2...A_n) = \frac{(\prod_{i=1}^{n} P(A_i \mid C))\, P(C)}{P(A_1, A_2...A_n)}$$

Huma Parveen, Shikha Pandey explained in their paper **'Sentiment Analysis on Twitter Data-set using Naïve Bayes Algorithm'[2]** details about analysing the tweets of twitter using Bayes rule given by the formula:

- SVM (Support Vector Machine)

Support Vector Machine (SVM) has defined input and output format, which is used for polarity of a comment. Input is a vector space and output is 0 or 1 (positive/negative). They are transformed into format which matches into input of machine learning algorithm input.

3) **Bags of Word Approach**

The task of fully understanding text is hardly easy since it involves a variety of complex concepts that are difficult to implement in machines. The bag of words approach follows a simple methodology i.e. to count the number of times each word appears in the given text and associate sentiment weight to each word depending on the overall sentiment value of text. It emphasizes on the idea of having one feature for each word which proves effective for sentiment analysis. It is used as a baseline in text analytics projects and natural language processing. Pre-processing stages can dramatically improve the performance of bag of words approach.

III. LIMITATIONS OF AVAILABLE SYSTEM

In the present scenario, a YouTube user who uploads video can see the following :

1. The number of likes and dislikes that he has received for a particular video
2. The different comments by the viewers (which can be more than millions)
3. The origin of the comment, i.e. from which nation a viewer has commented on the video.

The available system does not take into consideration that a YouTuber can have millions of followers, and it is practically impossible for the YouTuber to go through each and every comment, and make changes accordingly.

The proposed system will ensure that the YouTuber is able to understand how well the video was received, and what changes to implement in his upcoming or future works. This system will help to improve the quality of the videos that are uploaded by the YouTubers.

IV. SCOPE OF THE SYSTEM

The proposed system aims at developing a portal that can classify textual reviews obtained from YouTube into positive, negative and neutral categories based on polarity of the text. Here we perform sentiment analysis on YouTube comments which are written in English. The following system is proposed where:

1) Numerical Rating

The video will be given a rating on the scale of 1 to 10 after analysis of the comments is done.

2) Classification of the comments

The comments are classified as positive, neutral or negative on the basis of the rating given so that the YouTuber can deduce how the video was received by the audiences.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11376

3) Comment with maximum likes
The comments which received most no. of likes by viewers will be shown to the YouTuber.

V. SENTIMENT ANALYSIS USING LEXICON BASED APPROACH

Lexical based approaches are traditional approaches for sentiment analysis and text analytics in general. These approaches make use of lexicon of sentiment words i.e. a dictionary or dataset which contains all the words which lead to a sentimental orientation of text. For this system, we are using the '**SentiWordNet Dictionary'**, which consists of various words with both positive and negative polarity.

These words are associated with a particular sentiment value which leads to overall sentiment prediction of the text. Lexicon based approaches also comprise of linguistic rules which help in determining a more accurate sentimental value of the provided data. Some issues that are usually faced while using lexicon based approach are:

1. Some words can be both considered as positive or negative. For example - 'dangerous'. This word can be both used to express negative as well as positive sentiments. In negative sense it can be stated as: 'This act performed is extremely dangerous'. In positive sense it can be stated as: 'This is dangerously tasty'.
2. Some statements can be made sarcastic, just to mock the video. Sometimes some viewers who dislike the video may just post comments like: 'Wow, this video is sooooo goodddddd'. This statement is totally contradictory to the sentiment that the viewer originally has for the video.

Some statements may have sentiment, but can be difficult to analyze using this approach. For example: A viewer may post- 'I visited this place, but they took a long time to serve us'. This statement may not contain the major keywords, but still imply that the viewer did not like the place.

VI. SENTIWORDNET DICTIONARY

Sentiwordnet dictionary is a dictionary with various words, and their polarities/scores. The polarities are both positive and negative ranging from 0 to 1.[5] The format can be shown by the example given below:

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00001740	0.125	0	Able#1	Able to Eat

The value (POS,ID) uniquely identifies a WordNet(3.0) Synset. The values Positive Score and Negative Score are the polarity assigned to the respective words.

Synset column reports the words along with their sense number. The sense number refers to the various meanings of the given word. For example, let's say a word - 'field'. The word 'field' can have various meanings. One can be regarding the knowledge domain, the other can be a piece of land. These meanings can be represented as field#1, field#2 respectively.

VII. PROPOSED SYSTEM

1. Fetching Data from YouTube:
After the user enters the link of the YouTube video in the portal, we are scraping the comments of that video and storing them in a database using SQLite server. Hereafter, the steps for analysing the comments include the steps given below.

2. Pre-processing of the comments:
The next step is **pre-processing** of the comments which are obtained from YouTube. This major step includes:
i**. Removal of hashtags** - People usually express their opinions via hashtags, and it is necessary to remove them and use the word in the hashtags, for calculation of the score for the process. For e.g. If the comment is - "Wow the video is so good, and I felt inspired watching the madam talk about Artificial intelligence #inspired #best."
This comment talks about the liking of a person regarding an online lecture on artificial Intelligence. The hash tags express his sentiments that are – 'inspired' and his opinion about the video being the 'best'. Thus if we remove hash tags, the word can be used for the calculation for sentiment analysis.
ii. **Removal of usernames -** Usernames are not required for sentiment Analysis, as they are not helpful during calculation of the sentiment values for the comment

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11376

iii. **Removal of repeating words** - Repeating words can be irritating, and of no use, because those words cannot be found in the SentiWordNet dictionary. e.g. hurryyy, soonnn, bestttt etc. These words are needed to be corrected, and assigned their true meanings. e.g. hurryyy can be changed to hurry, soonnn can be changed to soon, besttt can be changed to best.

3. Applying the Sentiment Analysis Algorithm:

SentiWordNet Dictionary with the words, and their polarities are stored in the same database where the comments are stored. The words in the comments are then mapped to the dictionary, and their polarities are evaluated. With this, we evaluate the score of the given comment.

The algorithm is displayed below:

Data: Preprocessed Youtube comment

Output: Sentiment value of the comment

```
Find the list of sentiment words SentiList, their position and sentiment value
Find the list of negation words NegList, and their position
Find the list of blind negation words BlindNegList, and their position
Find the list of intensifiers words IntensifierList, and their position
if SentiList and NegList then
   for each word in the SentiList do
      if word is atmost the distance of 2 from NegList then
                  Revert the polarity of the word
      end
    end
end
if IntensifierList then
    for each word in SentiList do
       multipy the polarity with the intensifying value
    end
end
if BlindNegList then
          if SentiList then
                  return only the  negativity of the words
          end
end
```

3.1 Use of Emoticons - Emoticons can be a great way to express sentiments, when it comes to YouTube or any other social media platform. Unicodes of the Emoticons are used for the system. We are creating a new database where we will store the unicodes of the most frequently used emoticons and the sentiment value associated with them. This database can be used later in the calculation of the sentiment, when we encounter that particular emoji in the comment.

3.2 Analysis of Conjunctions:

Conjunctions are words, that are used to connect sentences together. e.g. This video was awesome **and** inspiring. Here, 'and' is a conjunction. There are four types of conjunctions:-

i. **Additive Conjunctions**- In these type of conjunctions, we consider both polarities of awesome and inspiring, and add them.

ii. **Contrastive Conjunctions(but, however)**- Here, if both are noun phrases, then we consider the second polarity, else we add them.

iii. **Conditional Conjunctions**- We add both the polarities

iv. **Comparative Conjunctions**- We add both the polarities

4. Displaying results:

The final results will include the following, which is shown in the image below:-

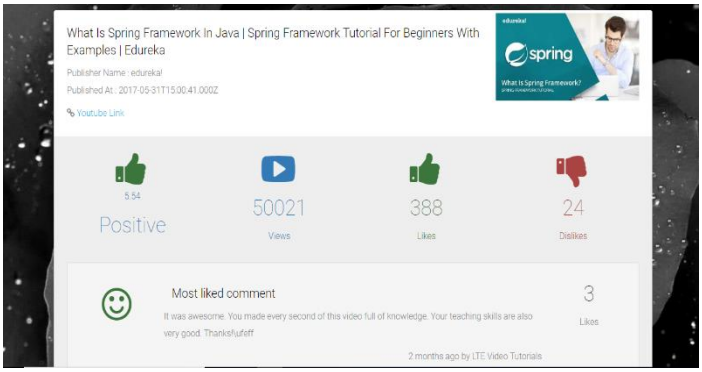
- Top 5 positive comments
- Top 5 negative comments
- Top 5 most liked comments
- Top 5 comments with most replies
- Overall score of the video.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11376



VIII. CONCLUSION AND FUTURE SCOPE

Sentiment analysis is based on lot of factors and achieving a high accuracy is difficult. We have gone through various algorithms such as Naive Bayes, Support Vector machine and previous projects and decided that we will be using a lexicon approach using dictionary. The lexicon approach makes use of lexicon of sentiment words i.e. a dictionary or dataset which contains all the words which lead to a sentimental orientation of text. In terms of future scope, this system can be implemented for various languages apart from English. Also, the algorithm used above can be refined for a specific domain of videos in future prospects.

REFERENCES

[1] Sanjida Akhter, Muhammad Tareq Aziz, *Sentiment Analysis On Facebook Group Using Lexicon Based Approach,* IEEE 2016
[2] Huma Parveen, Shikha Pandey, *Sentiment Analysis on Twitter Data-set using Naïve Bayes Algorithm,* IEEE 2016
[3] Gayathri Deepthi, K.Sashi Rekha, *Opinion Mining and Classification of User Reviews in Social Media,* IJARCSMS, Volume 2, Issue 4, April 2014
[4] http://sentiwordnet.isti.cnr.it/
[5] https://textblob.readthedocs.io/en/dev/index.html