Analysis & Implementation of Sentiment Analysis of User YouTube Comments

Shivani Wadhwani, Prashant Richhariya and Anita Soni

April 2, 2022

Analysis & Implementation of Sentiment Analysis of User YouTube Comments

Shivani Wadhwani[1], Prashant Richhariya[2], Dr.Anita Soni[3]

[1] Research Scholar, Department of Computer Science & Engineering, TIT Advance, Bhopal (M.P.)
[2] Professor, Department of Computer Science & Engineering, TIT Advance, Bhopal (M.P.)

ABSTRACT

With the turning wheel of time, the influence of the social networking websites on the people has significantly increased. People are now connecting with each other in cyber space and show their sentiments in the form of comments in different social networking websites such as Twitter, Facebook and Google Plus. YouTube is considered as a king in the field of video sharing. It is a largest video sharing repository, where people come and share their thoughts regarding video in the form of comments. "Sentiment Analysis" is the process of extracting other people's (speaker or writer) opinions from a given original source (text) utilizing natural language processing (NLP), linguistics computing, & data mining. For the interpretation of meaning of each and every comment, "text mining approach" is used. For understanding the meaningfulness of any content, it is important to classify them into positive and negative comments on the basis of user opinion. In the present study, researcher has performed sentiment analysis on YouTube comments on the most popular topics nowadays by using Classifier techniques.

Keywords: YouTube, Sentiment Analysis, text mining, social networking websites, Classifier technique

INTRODUCTION

Social media usage is increasing rapidly due to its ease in usage & simple to create & share pictures from everyone even those people can share the videos that are technically unaware of social media. Non-textual content is also shared by various different platforms such as animations, videos, images, audios which allows user to give their review through comments. Among various web platforms, YouTube is considered as the most popular one. The rich content is analyzed and exploited by sharing it on YouTube according to the research community's interest. The retrieval potential of the video is investigated according to some of the studies which states that comments as well as Meta data must be used. YouTube is an appropriate way which carry outs huge amount of positive comments in comparison with the inappropriate videos. So, it can be said that comments are one of the appropriate ways which perceives about the video quality, relevancy as well as its popularity. On the other hand, the comments by the users are generally unstructured and cannot be analyzed easily.

A human eye is always needed to watch it as it is prone to human errors. The extraction of the attitude or mood represented in a block of unstructured data in connection to the subject of the document being studied is called "sentiment extraction (SE)". There are various steps taken for the extraction as follows: Adjectives and adverbs are used which determines the polarity of the sentence at the very first level. Some of the positive words which can be used as rating are: good‖,‖awesome‖,‖excellent‖ and soon. On the other hand, some of the negative words are: —bad‖, —poor‖, —abomination‖ and so on.

Sentiment Analysis

Essential information & data regarding business profits and other important factors of scientific and commercial industries is contained in sentiment analysis. "Sentiment analysis" is needed because it is not possible to extract the information or keep manual tracking on them. Sentiments or opinion of the users on a particular product, subject or area are expressed which is known as Sentiment Analysis. Computational linguistics, natural language processings and also the text analytics are using sentiment analysis for identification of subjective data from the base or primary data. The sentiments are divided into categories such as: positive or negative. So, the general attitude is determined of the speaker or writer according to the topic.

Objectives of the Study

- Acquisitions of raw youtube data and converts into standard datasets.
- To analyze the content from youtube.
- Prediction of content from youtube.
- Content Classification and build the opinions of particular events.
- Develop Sentiment Analysis Tool

LITERATURE REVIEW

For different aspects of "youtube video features", several researches have been performed [1]. Decision is made on the basis of one of the important comment among several comments on any specific video [2]. The video objects are annotated by these comments [3,4]. The behaviour of the user is also analyzed through the comments and it can be used to find the troll users [5]. The sentiments can be analyzed for the comments and the positivity & negativity of the sentiments can also be analyzed [6]. The videos are categorized into different categories on the basis of comments on particular videos [7].

[8]. Sentiment Analysis is a methodology for analysing information in the form of text and determining the substance of thoughts from the content. Emotion mining or feeling mining are the other name given to Sentiment Analysis. YouTube, Facebook and Twitter are some of the social media platforms which are also known as "On-line communication channels" and nowadays these channels are creating so much interest in human life. On these social media channels, people share their thoughts, feeling etc. The after effects of studying different "Machine Learning" and "Lexicon investigation method" are represented in the present study. An evaluation study is performed is performed to study the outcomes of the analysis. The estimations of the present study are also analyzed. By understanding the present beginnings in this examination, it will be helpful in future investigations.

[9]. It is important for airline operators to get feedback from their customers particularly for operational planning and strategic planning. To get such type of feedbacks, social media sites are becoming trending. On the other hand, social media sites are not trivial task for the analysis, categorization and generation of useful data. The capabilities of "Naïve Bayes classifier" are investigated in the present study for the sentiment analysis of "airline image branding". The impact of data size is also examined in the present study on the accuracy of the classifier. For some online conversations, the data is collected according to the incident for airline's security operations which are managed by the passengers in a case study. A loss of $1 billion was observed about the incident for the company's corporate value. From Twitter, the data was collected and it was processed & analyzed through "Naïve Bayes Classifier". For determining the optimum training size, there is a demand for the mechanisms for observing the finest precision of the classifier. A damaging effect was observed on the negative perception of the customers on particular brand which will main to a catastrophic loss in the organization.

[10]. "Multi-label text classification" method is used for sentiment analysis of Hinglish comments on YouTube using Deep learning. Different parameters were used with "Multi-layer perceptron (MLP)" to investigate several sentiments in the comments. MLP has been modelled and estimated by making variations in optimizers, neurons, activation functions and several other engineering techniques. From the results, it was observed that 98.53percent of accuracy was observed in Kabita's Kitchen and 98.48percent accuracy was observed for

Nisha Madhulika in MLP. During the present study, careful tests were performed for the evaluation of the experiment.

[11]. Research on the detection of social media content, especially Twitter, has been done. Twitter content detection is based on classifying content or words made by users (tweets) into two groups, namely positive and negative. Research to detect negative content or harsh words in Indonesian tweets is still rare. There are several studies that have been conducted, the detection of negative words is only limited to certain categories, such as pornography, hate speech, and others, so that if the negative word only includes one category, then if there are other negative words that do not belong to that category, this word will not be detected. This is a challenge for researchers to classify texts in Indonesian. First, this paper will briefly explain the NB and SVM, then proceed with an explanation of the general research framework that has been carried out. In the results and discussion section, a comparison of the results achieved by existing researchers is explained. And based on these results, another approach will be proposed to detect negative content on Indonesian twitter.

[12]. Procedure of exploring the emotions, sentiments, ideas and thoughts in particular sentence are analyzed through sentiment classification which can be expressed by people. “Sentiment classification” is used to judge the sentiments and feelings of the people on the basis of analyzing the reviews, comments on social media etc. In sentiment classification, “Lexicon based techniques” and “Machine learning techniques” are used widely for the analysis of sentiments from comments and reviews of customers. Various learning algorithms are used in Machine learning techniques to make judgements on the sentiment through support vector machines, Navie bayes, etc. On the other hand Word net, Senti Word net are the other techniques used in Lexicon Based techniques. Complete image of “sentiment classification techniques” is provided by this survey. A comprehensive overview was provided by the survey paper in recent and historical researches on the sentiment classification.

METHODOLOGY

Sentiment Analysis on YouTube

Sentiment analysis is not an easy way to detect the sentiment polarity in social media particularly YOUTUBE. The main reason behind its challenging behaviour is challenges in current sentiment dictionaries. No proper sentiment is created by community in the present dictionaries. In overall YouTube traffic, 10percent was the Internet traffic and 20 percent was the web traffic present, according to the study conducted by [24, 25]. For the judgement of views and opinions on particular video by the users, various mechanisms are present of YouTube. Rating, voting and sharing are the mechanisms of YouTube. For several applications, useful data can be achieved through analysis of user comments. Various applications are: personal recommendations, filtration of comments as well as user profiling. For sentiment analysis, different techniques are used for user comments and for the same purpose “Senti word Net” is used.

Machine learning techniques

Machine learning (ML) is a data analytics technique which acquires knowledge from the previous practice and provides information directly to the systems without being programmed explicitly. Definition says "a computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E". ML has become a key technique for solving problem and for making better decisions and predictions. Mainly we used machine learning techniques every day to make significant decisions in stock trading, medical diagnosis, weather forecasting, and many more.

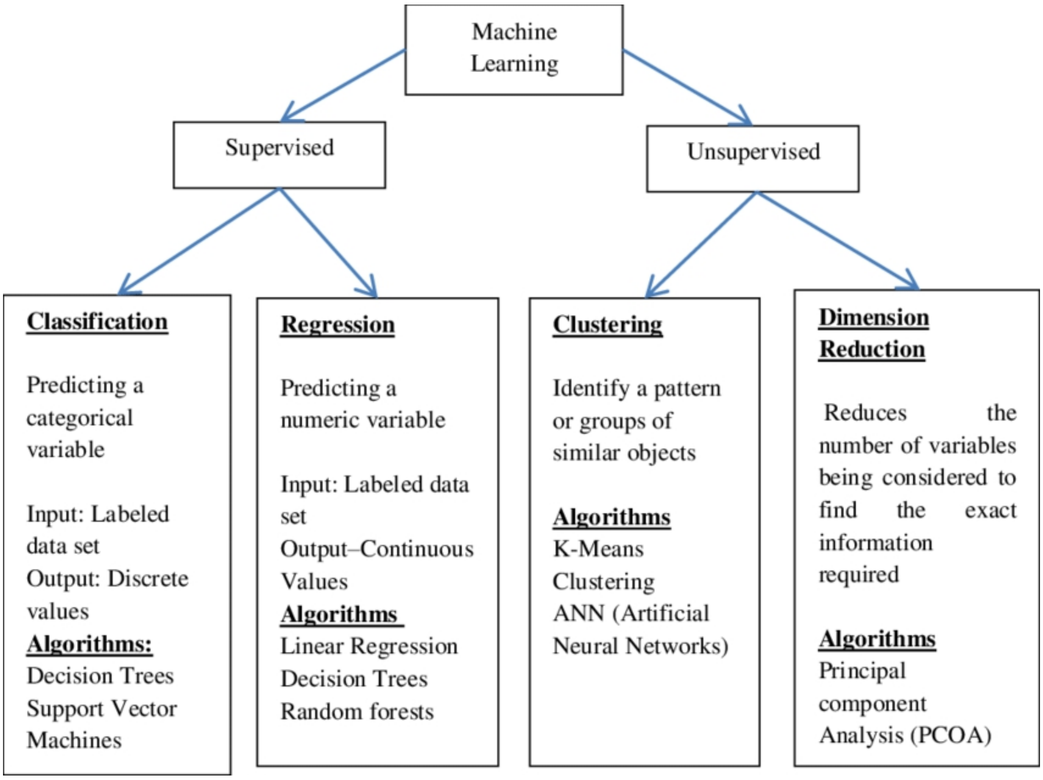

Figure 1: Flow chart of ML Technique

RESULTS AND DISCUSSION

Selection of the Sentiment Analysis Approaches

For the selected particular YouTube videos, this section shares the acquisition of comments from Kurzgesagt – In a Nutshell - "The Coronavirus Explained & What You Should Do".

A focused crawler is implemented in order to address this task. Comments are extracted from the video by web API which uses HTTP GET methodology according to the video URL. However, on the basis of languages & various notions, the extracted comments are heterogeneous. Some pre-processing methods are used on unstructured comments to generate the data sets.

After extracting the comments, following changes are performed:

- According to the proposed method, all the expressions must be removed which are irrelevant such as ("May 20- 2021"or "25-6-2021"), link (“www.imdb.com, www.tmdb.com etc.”), numeric (12, 20 etc.) as well as specific characters (“"*","/","!","@","?","#","&","$"”), emojis (“ ,<3 etc.”) and various languages (“Chinese, Arabic, Bangla , Hindi etc.”).
- All the punctuation marks must be removed for example, period (“.”), space ("-"), commas (","), semicolon (“;”), hash (“-“) etc.

Network Analysis

Network analysis is becoming a widely used tool for scholars specially to deal with the complexities of interrelation between actors as well as all sorts. Instead of seeing actors as isolated entities, Network analysis provides a placement of significant on the relationship among the actors. For network analysis, there are numerous applications and the network graph is also created for example: cytoscape and gephi. In network analysis, R has been developed as the powerful tool; however it is not particularly designed for the same. In comparison with the “strength of R”, the stand-alone network analysis software is a threefold. Reproducible research is enabled by R at first which is not possible with GUI. Robust tools are provided by the data analysis power of R for the manipulation of data especially for network analysis. R is transformed into a complete network analysis tool with the ever growing range of packages designed. Statement suite of igraph and packages are included in the essential network analysis packages for R.

Nodes and Edges

“Multitude of separate entities” as well as their connection is the two essential aspects of the network. In this, the vocabulary used might be inconsistent and even technical among the packages, softwares and disciplines. Either nodes or vertices, the entities are considered for the graph and the connections are performed as links or edges. Nomenclature of edges and nodes are mainly used here except the discussed packages which were using different vocabulary. The data must be in particular way in the network analysis packages so that special type of objects can be created which are utilized in each package. Adjacency matrices, sometimes termed as socio matrices, are used to create the object classes network, igraph, and tidygraph.

Edge and node lists

In 1585, Daniel van der Meulen received letters to form a network object from the database. In the present study, the researcher also making both edge list & node list. The data frame of letters is manipulated through the necessitated use of the dplyr package which was sent to Daniel and was divided into two different fragments consisting of structure of edge and node lists.

Network Objects

Tidygraphy and igraph are closely related with the network object classes. It is not difficult to translate in between the igraph object and network object. It is important to keep both the packages and objects separately. It will be best if only one package loaded at one time because capabilities of network can overlap the igraph.

Performance Measures

Precision, accuracy and recall performance evaluation are used for the analysis of accuracy value of the SVM method's results compared with the confusion matrix method. Confusion Matrix is used for the evaluation which includes False Negative Rate (FN rate), True Positive Rate (TP rate), False Positive Rate (FP Rate) and True Negative Rate (TN Rate). These are the indicators used here. For positive class, TP rate is used and for negative class TN rate is used. FP rate is classified as positive class however it is class negative. FN rate is classified as the negative class however it is a class positive. [28]

In performance measures discuss the performance evaluative measures to youtube datasets. Following are the evaluative measures of classifiers. These evaluative measures find the goodness of classifier and suitability of data.

True Positives (TP): TP are the predicted values here which means that value of actual class is YES as well as the value of predicted class is also YES. For example: If the same thing is indicated by actual class and the predicted class that this passenger survived.

True Negatives (TN): TN are the predicted values here which means that value of actual class is NO as well as the value of predicted class is also NO. For example: Actual class and predicted class will say the same thing such as passenger did not survive. When actual class contradicts with the predicted class then false positive and false negatives are obtained.

False Positives (FP): FP are the predicted values here which means that value of actual class is NO and the value of predicted class is YES. For example: Passenger did not survive according to actual class, but passenger will survive according to predicted class.

False Negatives (FN): FN are the predicted values here which means that value of actual class is YES and the value of predicted class is NO. If actual class value indicates that the passenger survived and the predicted class states that passenger will die.

The following section describes the parameters for the calculation of Accuracy, Precision and Recall.

Accuracy: Accuracy is the ratio of correctly predicted observation and the total observations. It is considered as the most intuitive performance measure. People might think that if they got high accuracy then their model is best. And yes accuracy is the great measure but only one condition when values of false positive and values of false negatives are almost same. So, it is important to evaluate the performance of the model by taking other parameters also in consideration.

Accuracy = TP+TN/TP+FP+FN+TN

Precision - The ratio of correctly observed positive observations of predicted class and the total number of observations of predicted class is defined as Precision. The challenge that this measure answers is how many of the passengers who were identified as having survived obviously did. The low false positive rate is related to high precision.

Precision = TP/TP+FP

Recall (Sensitivity) - The ratio between the predicted positive observations and the observations in actual class are defined as Recall. The answers provided by recall are: All the passengers who really survived and how many of them were labelled.

Recall = TP/TP+FN

Table 1: Accuracy Performance Measure

Methodology	**Accuracy (%)**
Existing work [29]	81%
Proposed work	89.3%

CONCLUSION

YouTube is defined as the source of comprehensive video information on the web. In all the social media sites, YouTube is one of the most popular sites. As on this site, users can directly interact with rating, sharing and commenting on videos.

Sentiment analysis model is proposed here for YouTube video comments. Naïve Baise Algorithm is also used here. The “neural network's output” is a categorization of negative, positive, or neutral sentiments.

Following are the problems emphasised in the present work to identify the polarity of the comments given by the YOUTUBE users.

1. Challenges in present sentiment dictionaries.
2. Users are using informal language.
3. On the basis of community-created terms, sentiments are estimated.
4. Proper labels must be assigned to the events.
5. To attain satisfactory classification performances
6. Challenges involved in “social media sentiment analysis”.

In comparison with the statistical model, the results showed that this model achieved better accuracy. The range of classification accuracy is in between 70percent to 89percent.

REFERENCES

1. S. Siersdorfer, S. Chelaru, J. S. Pedro, I. S. Altingovde and W. Nejdl, “Analyzing and mining comments and comment ratings on the social web,” ACM Transactions on the Web (TWEB), 8(3), 2014, pp. 1-39.
2. S. Siersdorfer, S. Chelaru, W. Nejdl and J. San Pedro, “How useful are your comments? analyzing and predicting youtube comments and comment ratings,” In Proceedings of the 19th international conference on World wide web (ACM), 2010, pp. 891-900.
3. E. Momeni, C. Cardie and M. Ott, “Properties, Prediction, and Prevalence of Useful User-Generated Comments for Descriptive Annotation of Social Media Objects,” In Proceedings of ICWSM, 2013.
4. O. Uryupina, B. Plank, A. Severyn, A. Rotondi and A. Moschitti, “SenTube: A Corpus for Sentiment Analysis on YouTube Social Media,” In LREC, 2014, pp. 4244-4249.
5. E. Momeni, B. Haslhofer, K. Tao and G. J. Houben, “Sifting useful comments from Flickr Commons and YouTube,” International Journal on Digital Libraries, 16(2), 2015, pp.161-179.
6. H. Lee, Y. Han, Y. Kim and K. Kim, “Sentiment analysis on online social network using probability Model,” In Proceedings of the Sixth International Conference on Advances in Future Internet, 2014, pp. 14-19.
7. K. Filippova and K. B. Hall, “Improved video categorization from text metadata and user comments,” In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, 2011, pp. 835-842.
8. Mehta, P., & Pandya, S. (2020). A Review On Sentiment Analysis Methodologies , Practices And Applications. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, *9*(02), 601–609.
9. Odim, M. O., Ogunde, A. O., Oguntunde, B. O., & Phillips, S. A. (2020). Exploring the Performance Characteristics of the Naïve Bayes Classifier in the Sentiment Analysis of an Airline ’ s Social Media Data. *Advances in Science, Technology and Engineering Systems Journal*, *5*(4), 266–272.
10. Donthula, S. K. (2019). Man is What He Eats : A Research on Hinglish Sentiments of YouTube Cookery Channels Using Deep learning. *International Journal of Recent Technology and Engineering (IJRTE)*, *2*, 930–937. https://doi.org/10.35940/ijrte.B1153.0982S1119
11. Andana, K., Andana, E. K., Othman, M., & Ibrahim, R. (2019). of Advanced Trends Computer Engineering Comparative Analysis of Text Classification Using Naive Bayes and Support Vector Machine in Detecting Negative Content in Indonesian Twitter. *International Journal of Advanced Trends in Computer Sci Nce and Engineering*, *8*(1).
12. Rizwan, M., & Rana, R. (2018). *A survey on sentiment classification algorithms , challenges and applications*. *1*, 58–72. https://doi.org/10.2478/ausi-2018-0004