**Nagarjuna College of Information Technology**

**Sankhamul, Kathmandu**

**Department of Computer Science and Information Technology**

****

**Project Proposal on**

**“Youtube Video Sentiment Analysis”**

**Submitted By:**

Anish Basnet - 20883/075

Ganesh Tiwari - 20885/075

Madan Sunwar - 20889/075

**Submitted To:**

Department of Computer Science and Information Technology Nagarjuna College of Information Technology

**Supervisor**

Ramesh Singh Saud

**Table of Content**

**1. Introduction 1**

**2. Problem Statement** **1**

**3. Objectives** **1**

**4. Methodology** 2

**a. Requirement Identification 2**

**i. Literature Review 2**

**ii. Requirement Analysis 3**

**b. Feasibility Study 5**

**i. Technical 5**

**ii. Operational 5**

**iii. Economic 5**

**iv. Schedule (Gantt chart) 5**

**c. High Level Design of System 6**

**i. Methodology of the proposed system 6**

**ii. Flow Charts 7**

**iii. Working Mechanism of Proposed System 7**

**iv. Description of Algorithms 8**

**5. Expected outcome** 10

**6.**  **References** 10

**1. Introduction**

Sentimental analysis is additionally referred as opinion mining that means to find out or identify the positive, negative, neutral opinions, views, attitudes, impressions, emotions, and feelings indicated in the text.

In this work, we will collect the data from the you-tube comments of the public and measures the attitude of the user towards the aspects of a video which they describe in a text and then give rating to the video on basis of user comments. Sentiment analysis is useful for quickly gaining the whole idea by using large number of text data and it will be helpful to understand the user’s opinion. YouTube provides many social mechanisms to gauge user opinion and views a few videos by means of voting, rating, favorites, sharing and negative comments, etc. Text analytic is that the analysis of “unstructured” data contained in natural language text using various methods machine learning tools, and techniques. Text analysis offers a very low-cost method to gauge public opinion.

We have developed a system based on Naive-Bayes classification algorithm which is based on principle of Bayes theorem which plays vital role to determine polarity of a sentence. To improve our system performance, we have used different features selection techniques like tokenization, stop words removal, lower casing the comments and punctuation removal.

**2. Problem Statement**

Generally, users watch YouTube videos on the basis of like and views, but the videos might not have a positive opinion from the users. The actual opinion/comment of the users may not be justified by the likes and views videos get. This has proven to be a challenge to identify the response of users about the video.

**3. Objectives**

The main objective of the system is:

* To scrap the YouTube comments and analyze the sentiment of the users about the video checking the polarity using Naive Bayes theorem.

**4. Methodology**

**a. Requirement Identification**

**i. Literature review**

In 2019, Saad and Yang [1] have aimed for giving a complete tweet sentiment analysis on the basis of ordinal regression with machine learning algorithms. The suggested model included pre-processing tweets as first step and with the feature extraction model, an effective feature was generated. The methods such as SVR, RF, Multinomial logistic regression (SoftMax), and DTs were employed for classifying the sentiment analysis. Moreover, twitter dataset was used for experimenting the suggested model. The test results have shown that the suggested model has attained the best accuracy, and also DTs were performed well when compared over other methods.

In 2019, Afzaal et al. [2] have recommended a novel approach of aspect-based sentiment classification, which recognized the features in a precise manner and attained the best classification accuracy. Moreover, the scheme was developed as a mobile application, which assisted the tourists in identifying the best hotel in the town, and the proposed model was analyzed using the real-world data sets. The results have shown that the presented model was effective in both recognition as well as classification.

In 2019, Feizollah et al. [3] have concentrated on tweets related to two halal products such as halal cosmetics and halal tourism. By utilizing Twitter search function, Twitter information was extracted, and a new model was employed for data filtering. Later, with the help of deep learning models, a test was performed for computing and evaluating the tweets. Moreover, for enhancing the accuracy and building prediction methods, RNN, CNN, and LSTM were employed. From the outcomes, it was seemed that the combination of LSTM and CNN attained the best accuracy.

In 2018, Mukhtar et al. [4] have performed the sentiment analysis to the Urdu blogs attained from several domain with Supervised Machine learning and Lexicon-based models. In Lexicon-based models, a well-performing Urdu sentiment analyzer and an Urdu Sentiment Lexicons were employed, whereas, in Supervised Machine learning algorithm, DT, KNN, and SVM were employed. The data were combined from the two soruces for performing the best sentiment analysis. Based on the tests conducted, the outcomes were shown that the Lexicon-based model was superior to the supervised machine learning algorithm.

In 2020, Kumar et al. [5] have presented a hybrid deep learning approach named ConVNet-SVMBoVW that dealt with the real-time data for predicting the fine-grained sentiment. In order to measure the hybrid polarity, an aggregation model was developed. Moreover, SVM was used for training the BoVW to forecast the sentiment of visual content. Finally, it was concluded that the suggested ConvNet-SVMBoVW was outperformed by the conventional models.

In 2020, Xu et al. [6] have introduced a NB method for multi-domain and large-scale E-commerce platform product review classification of sentiment. Consequently, the parameter evaluation method was extended in NB for continuous learning fashion. Later, for fine-tuning the learned distribution on the basis of three types of assumptions, many ways were introduced for acquiring the best performance. The results have shown that the suggested model has high accuracy in Amazon product and movie review sentiment datasets.

In 2018, Smadi et al. [7] have proposed existing models on the basis of supervised machine learning algorithms for specifying the defects of feature-based sentiment analysis of Arabic hotel's review. Moreover, SVM and Deep RNN were developed and trained with word, lexical, morphological, semantic, and syntactic features. The reference dataset of Arabic hotel's review dataset was used for evaluating the proposed model. The outcomes have shown that SVM was performing well when compared over RNN model.

In 2019, Abdi et al. [8] have suggested a deep-learning-based technique for categorizing the opinion of the user mentioned in reviews. Moreover, a deep learning model was a unified feature set that was representative of sentiment shifter rules, word embedding, sentiment knowledge, linguistic and statistical knowledge has not been continuously explored for a sentiment analysis. Moreover, the suggested model used RNN that consisted of LSTM for considering the benefit of sequential processing and conquered many issues in conventional algorithms.

In 2020, Park et al. [9] have designed a deep learning approach for improving performance. In order to improve the performance, two questions have come into picture. The content attention was required for being sophisticated for merging many attention results non-linearly and assumes the whole context for mentioning the complex sentences. The test results have shown that the proposed model was attained as the best performance.

In 2019, Bardhan et al. [10] have explored a quasi-qualitative model for understanding the underlying the affects of gender mainstreaming in SRH management. In order to explore the stakeholder concerns, verbal narratives froms semi-structured intervies and concentrated on group discussions. For decoding the emotions over the stakeholders, sentiment analysis with machine learning algorithm of NLP is employed.

In 2017, Araque et al. [11] have introduced a deep learning model for enhancing the performance by incorporating the existing surface models with deep learning models on the basis of manually extracted features. With the help of linear machine learning and word embeddings methods, a deep learning-based sentiment classifier was introduced. Here, 7 datasets were used for verifying the efficiency of the suggested model. The results have confirmed that the presented method was effective when compared with traditional methods

**ii. Requirement Analysis**

System requirement is the ability of the system to meet the condition desired by the users. System requirement analysis is performed by grouping the needs into functional requirements and non-functional requirements

**Functional Requirements**

The functional requirements of the Sentiment Analysis System of YouTube comments are as follows:

1. The system can manage, mine and store YouTube comments as pandas dataframe.

2. The system can store the dictionary of root words of YouTube comments

3. The system can perform preprocessing of comments removing stop words, punctuation, lower casing comments and remove unnecessary data.

4. The system can display the sentiment of YouTube video comments after sentiment analysis result derived from comments submitted by users.

**Non-Functional Requirements**

Meanwhile, the non-functional requirements of the Sentiment Analysis System of YouTube comments are as follows:

1. The system can run in various web browsers which support the system environment,

2. The system gives a fast response.

3. The system has a user-friendly interface design.

**b. Feasibility Study**

**i. Technical Feasibility**

|  |  |  |
| --- | --- | --- |
| Tools | Used | Explanation |
| IDE | visual studio | It supports python and is easy to use. |
| Management and diagram | Wondershare EdrawMax | It is used to draw Gantt chart, scheduling task and resources. It is used to construct diagram such as usecase diagram, flowchart etc. |
| Operating system | Window 10 | It is easy to use and compatible |
| Programming  Language | python | It is used to develop application. |
| Framework | Flask | This framework is used in developing web based application |

**ii. Operational Feasibility**

The users want to watch a video that have positive sentiments from the users. They will be able to see the video rating according to the sentiment of the comments. Users can ignore the videos with the most negative sentiment and watch videos that have positive review.

**iii. Economic Feasibility**

The project should be completed with minimum resources and cost with an objective of profiting and giving the efficient and desired outcome. Using processors and computers that are low cost to store the data and run the application smoothly.

**iv. Schedule (Gantt chart)**

A picture containing graphical user interface

Description automatically generated

**c. High Level Design of System**

**i. Methodology of the proposed system**

To scrap YouTube comments, we are using packages selenium, web driver-manager and beautifulsoup4. We then store it in pandas dataframe in csv format after preprocessing then we are using naïve bayes classifier to find polarity of comments. During preprocessing we will lower cased the comments, removed stopwords, remove punctuations and any unnecessary data.

The system is designed using Flask framework for frontend. Flask providea fast and easy manipulation to create user friendly UI.

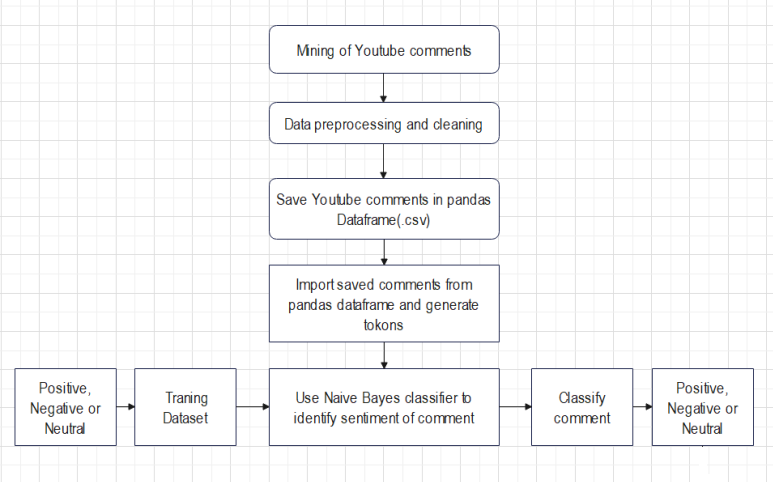


Fig. Structure Diagram

**ii. Flow Charts**

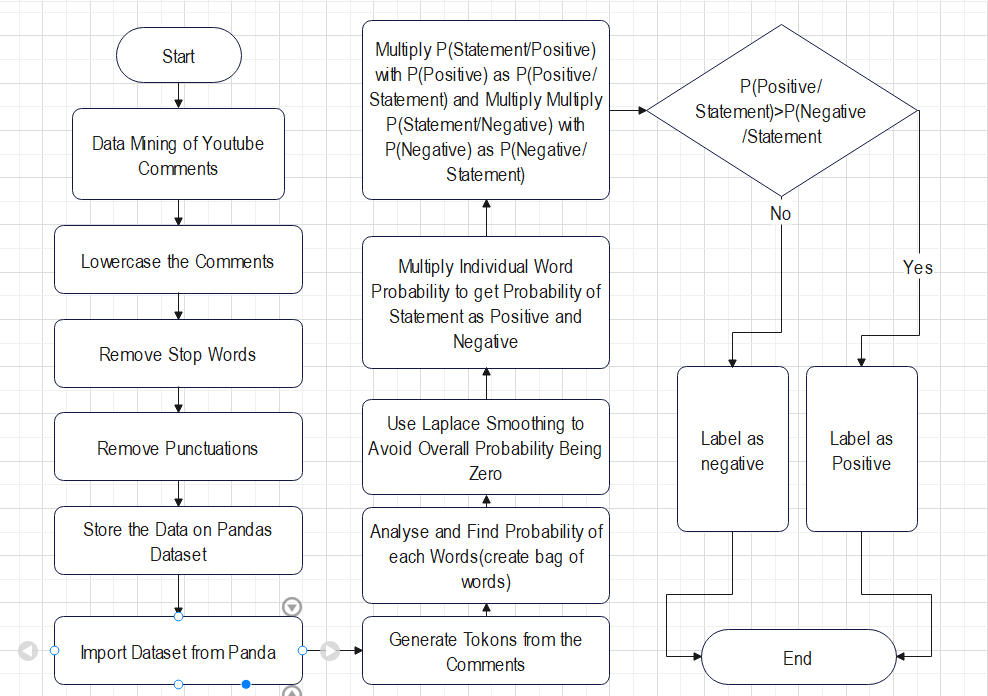
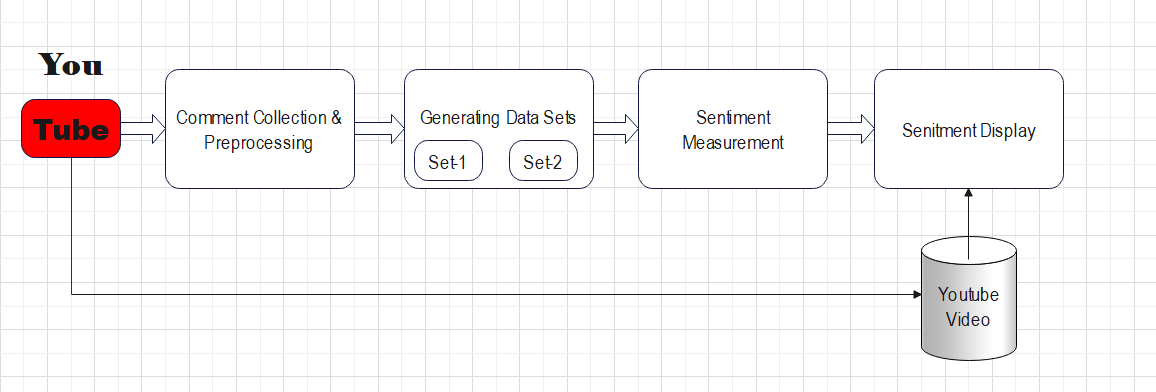


Fig. Flow chart of Naive Bayes classifier

**iii. Working Mechanism of Proposed System**



The system scrap the comment from the YouTube and do preprocessing to remove punctuation, making text lower case, remove stop words, etc. then save it in pandas dataframe. Then train the data sets for individual words then use Naive Bayes algorithm to check for the polarity of the comment.

**iv. Description of Algorithms**

For this project, Naive Bayes classifier have been used. This classifier works on the principles of Bayes Theorem. Naive Bayes is a supervised machine learning algorithm based on Bayes’ theorem. Bayes' theorem is defined mathematically as the following equation:

P(A|B) = (P(B|A) \* P(A))/ P(B)

* A is the polarity of statement depending on what probability we are calculating. (i.e., positive or negative)
* B is the statement for which we are calculating the probability of it being positive or negative.
* A and B are independent events (i.e., the probability of the outcome of event A does not depend on the probability of the outcome of event B).
* P is the probability.
* P(A|B) represents the probability of event A happening given that B is true.
* P(B|A) represents the probability of event B happening given that A is true.
* P(A) and P(B) are the probabilities of observing A and B without any prior conditions. These are referred to as prior probabilities.

For sentiment analysis we use formula derrrived from above Naive Bayes theorem ,After tokanization of sentence we get words represented as wi .

P(A|B) = P(A) \* Π P( wi | A) …..1

where i belongs to positions of tokanized words

**TRAINING THE MODEL**

**1. Calculate the Prior Probability:** We will first find the number of documents belonging to each class . Finding the percentage of the documents in each class will give us the required prior probability.

Let’s assume the number of documents in class Positive is Np.

Total number of documents is assumed to be total.

So , P(Positive) = Np / total which is similar for Negative class.

1. **Calculate the Likelihood Probability:** Our main goal is to find the fraction of times the word wi appears among all words in all documents of class Positive. We first concatenate all documents with category Positive and Negative into one big text ‘bag’ which is Bag of words. We will however face a very unique problem at this point . If the word doesn’t belong to the given class then overall probability will be zero. To combat this problem we will introduce an add-on , Laplace Smoothing Coefficient , to both the numerator and the denominator . Our equation will be modified as follows:

P( wi | Positive) = (count( wi | Positive) + a) / (bag(Positive) + length(bag) \* a)

P( wi | Negative) = (count( wi | Negative) + a) / (bag(Negative) + length(bag) \* a)

* P( wi | Positive) is word being positive
* P( wi | Negative) is word being negative
* count( wi | Positive) count of word wi in bag which are positive
* count( wi | Negative) count of word wi in bag which are negative
* bag(Positive) is length of positive word in bag
* bag(Negative) is length of negative word in bag
* length(bag) is length of bag
* a is the Laplace smoothing coefficient. We usually consider its value to be 1.

After calculating all the individual word probabilities and applying Laplace smoothing to each of them, we finally create a bag of words with each words negative and positive probability. Now for calculating overall probability of a statement we tokanize it first and use equation Naive Bayes theorem using our bag of words. After that, all we do is compare which class has the greatest probability; this class is the output of the classifier.

The input to this algorithm is the statement ,bag of words and prior probabilities of each class and it classifies the sentiment of sentence as output

**ALGORITHM STEPS:**

* STEP 1: START
* STEP 2: Calculate prior probability of each class label (positive or negative).
* STEP 3: Calculate conditional probability with each feature for each label.
* STEP 4: Multiply the conditional probability for each label.
* STEP 5: Multiply the label probability with STEP 4.
* STEP 6: The label with highest probability is taken as sentiment.
* STEP 7: STOP

**5. Expected outcome**

The expected system should be able to analyze the sentiments of the YouTube comments. The system should be able to identify the polarity of the comments as Positive or Negative.

**6. References**

[1] Patel, V., Prabhu, G. and Bhowmick, K., (2015). A survey of opinion mining and sentiment analysis. International Journal of Computer Applications, 131(1), pp.24-27.

[2] Sahayak,V.,Shete,V.,&Pathan, A. (2015). Sentiment analysis on twitter data. International Journal of Innovative Research in AdvancedEngineering(IJIRAE),voume 2(1), page178-183.

[3] Sharma, P., & Moh, T. S. (2016, December). Prediction of Indian election using sentiment analysis on Hindi Twitter. In 2016 IEEE international conference on big data (big data) (pp. 1966-1971). IEEE.

[4] Baid, p., Gupta, A., and Chaplot, N. (2017). Sentiment analysis od movies reviews

Using machine learning techniques. International Journal of Computer Application,

Volume 179, page 45-49.

[5] Wagh, R., & Punde, P. (2018, March). Survey on sentiment analysis using twitter dataset. In 2018 SecondInternational Conference on Electronics, Communication and Aerospace Technology(ICECA) (pp. 208-211). IEEE.

[6] Abbas, M., Memon, K.A., Jamali, A.A., Memon,S. and Ahmed, A.,(2019). Multinomial Naive Bayes classification model for sentiment analysis. IJCSNSInt*.*J *.*Comput. Sci.Netw*.* Secur, volume 19(3), p.62.

[7] Laksono, R. A.,Sungkono, K. R, Sarno, R and Wahyuni, C. S. (2019, July). Sentiment analysis of restaurant customer reviews on TripAdvisor using Naïve Bayes. In 2019 12th International Conference on Information & Communication Technology and System (ICTS) (pp. 49-54). IEEE.

[8]Puspsari,H.M., and Subarkah,P. (2022).Sentiment Analysis For Opinions on the Covid-19

Vaccination Program Using a Naïve Bayes Classifier. Journal Borneo Administrator, volume 18, page 213-230.

[9] Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers’ protest using twitter data. International Journal of Information Management Data Insights, volume1(2), 100019.

[10]. Shakhy P S and Vidya K S, "A Survey on KASR for Big Data Applications", International Journal of Computer Sciences and Engineering, Vol.3, Issue.4, pp.85-89, 2015.

[11]. Ashish Shukla, Rahul Mishra, “Sentiment Classification and Analysis using Modified K-means and naïve Bayes Algorithm” International journal of Advanced Research in computer science and Software Engineering. Volume 5, Issue 8, August 2015.