**TREND DETECTION IN SOCIAL MEDIA DATA**

Minor project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

# Computer Science and Engineering

By

## SUDEEP (191323)

## RITIK (191324)

**UNDER THE SUPERVISION OF**

Mr. PRAVEEN MODI (**Assistant professor)**



Department of Computer Science & Engineering and Information Technology

# Jaypee University of Information Technology, Waknaghat, 173234, Himachal Pradesh, INDIA

**TABLE OF CONTENT**

|  |  |
| --- | --- |
| **Title** | **Page No.** |
| **Declaration** | **I** |
| **Certificate** | **II** |
| **Acknowledgement** | **III** |
| **Abstract** | **IV** |
| **Chapter-1 (Introduction)** | **1-3** |
| **Chapter-2 (Feasibility Study, Requirements Analysis and Design** | **4-6** |
| **Chapter-3 (Implementation)** | **7-18** |
| **Chapter-4 (Results)** | **19-20** |
| **References** | **21** |

I

**DECLARATION**

I hereby declare that, this project has been done by me under the supervision of Mr. Praveen Modi**, Assistant Professor (Grade-I),** Jaypee University of Information Technology. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

**Supervised by:**

**Mr. Praveen Modi**

Assistant Professor(Grade-I)

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology

**Submitted by:**

**Sudeep (191323)**

**Ritik (191324)**

Computer Science & Engineering Department

Jaypee University of Information Technology

II

**CERTIFICATE**

This is to certify that the work which is being presented in the project report titled “Trend Detection in Social Media Data” in partial fulfillment of the requirements for the award of the degree of B.Tech in Computer Science And Engineering and submitted to the Department of Computer Science And Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by “Sudeep (191323) and Ritik (191324)” during the period from January 2022 to May 2022 under the supervision of Mr. Praveen Modi, Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

Sudeep (191323)

Ritik (191324)

The above statement made is correct to the best of my knowledge.

Mr. Praveen Modi

Assistant Professor (Grade-I)

Computer Science & Engineering and Information Technology

Jaypee University of Information Technology, Waknaghat,

III

**ACKNOWLEDGEMENT**

Firstly, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the project work successfully.

I really grateful and wish my profound my indebtedness to Supervisor **Mr. Praveen Modi, Assistant Professor(Grade-I)**, Department of CSE Jaypee University of Information Technology, Waknaghat. Deep Knowledge & keen interest of my supervisor in the field of “**Machine Learning**” to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Mr. Praveen Modi,** Department of CSE, for his kind help to finish my project.

I would also generously welcome each one of those individuals who have helped me straightforwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

Sudeep (191323)

Ritik (191324)

IV

**ABSTRACT**

Sentiment analysis or assessment mining is the field of study connected with dissecting conclusions, sentiments, assessments, perspectives, and feelings of clients which they express via web-based entertainment and other web-based assets. The insurgency of web-based entertainment destinations has additionally drawn in the clients towards video sharing locales, like YouTube. The internet based clients offer their viewpoints or sentiments on the recordings that they watch on such destinations. This undertaking presents a short overview of strategies to examine sentiments posted by clients about a particular video. Opinion mining or remarks toward disposition assessment, individual substance, are normally called sentiment. Everybody is allowed to offer perspectives related to the current conclusions on youtube. Consequently individuals have a through and through freedom to offer their viewpoint in regards to the exhibition. Because of the rise of numerous critics that show up in a short measure of time, there is a need to lead an analysis on opinion mining. The method involved with looking or following the normal language to observe examples or states of mind of society against specific items, individuals or points is called Sentiment Analysis.

[1] Sentiment analysis procedures can uphold numerous choices in numerous situations. This review utilizes three class ascribes, which are positive, neutral and negative, in light of the fact that in the web the remarks that seem can be positive, neutral and negative comments.

[2] TextBlob is a Python (2 and 3) library for handling literary information. It gives a basic API to jumping into normal regular language handling (NLP) undertakings, for example, part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and that's only the tip of the iceberg.

**1**

**Chapter 01**

**INTRODUCTION**

* 1. **Introduction**

For Trend Detection/Analysis we have different substance over the web, for example, message, picture, video and we picked message and we performed NLP tasks to it to get the right sentiment analysis

* Essentially we need to break down the sentiment of the client in view of the dataset. Also examined the Positive and Negative remarks
* Sentiment is essentially an inclination or an idea of an individual or basically we can say that an assessment of one individual is communicated. For example:- Let's say that you have told a person that you look revolting means you have negative sentiment for that individual. What's more, also a positive sentiment is some great idea, great inclination or great assessment
* What's more, to do that sentiment analysis we need to introduce a package called "textblob".
* Textblob is an open-source python library for handling literary information. It performs various procedures on literary information, for example, noun phrase extraction, sentiment analysis, classification, translation, etc.
  1. **Objective**

The target of the review is to decide whether a straightforward forecast in light of the sentiment of the comments can be valuable for anticipating like extents on Youtube recordings. We are checking whether the youtube video is having a positive impact or negative impact. In this project we have scrapped the comments of youtube videos and applied functions available in the TextBlob library to check the polarity of comments. We will likewise consider how neutral comments can be utilized in the forecast and think about performance while training on Youtube comments.

* 1. **Motivation**

It is normal to openly post opinions via online entertainment, like Twitter, Facebook or YouTube, and these opinions can give fundamental data to realize individuals' opinion on items and administrations. Besides, these opinions could

**2**

impact others' opinions and influence their choices in whether they would be keen

on a specific item, association or point. This client feedback may, thusly, illuminate about broad

patterns like ubiquity and quality. YouTube is essentially a stage where its clients can transfer their own recordings, yet in addition a stage for organizations to promote their items, media sources to share data on recent developments from there, the sky is the limit. YouTube offers an accomplice program for its clients, where content makers can adapt their recordings and procure promotion income. How much cash a maker can procure is straightforwardly connected with how much premium, in the types of perspectives, evaluations and remarks, that they can produce on their substance. Besides, well known recordings could assemble significant client feedback, for example, remarks, where programmed ways of knowing the feeling extremity is attractive. The video creators' reliance on their watchers' endorsement raises the significance of their feedback, which makes it alluring additionally to concentrate on the connection between the different types of client feedback.

* 1. **Language Used**

Python is a simple to learn, strong programming language. It has efficient high-level data structures and a basic however powerful way to deal with object-oriented programming. Python's rich linguistic structure and dynamic composing, along with its deciphered nature, make it an optimal language for prearranging and quick application advancement in numerous areas on most platforms.

* 1. **Technical Requirements (Hardware)**

PC hardware particulars are unit specialized portrayals of the PC's components and abilities. Processor speed, model and manufacturer. Processor speed is normally demonstrated in gigacycle each second (GHz). the lower the amount, the speedier the pc. Random Access Memory (RAM). This is normally demonstrated in gigabytes (GB).The extra RAM in an extremely huge PC the extra it will do simultaneously. Exhausting disk (sometimes referred to as ROM) house. This can be normally demonstrated in gigabytes (GB). Processor Intel Pentium/Core - one. 7GHz and higher than Memory 1GB and higher than Storage 80GB least free house Graphics 1GB and higher with a stable internet connection.

**3**

* 1. **Deliverables/Outcomes**

A model has been trained for the sentiment analysis of the Youtube comments followed by the pre-processing of the dataset. Pre-processing incorporates data naming, lowercasing of the text, stop words expulsion, feature extraction.

**4**

**Chapter 02**

**Feasibility Study, Requirements Analysis and Design**

**2.1 Feasibility Study**

While doing this project, different advances were taken to guarantee that the data and the outcomes gathered can be repeated to as near the depicted numbers as could really be expected. In any case, there are a few contemplations, specifically, that ought to be accentuated. Endeavoring to aggregate YouTube comments and YouTube video insights will undoubtedly not bring about similar accurate data as it was gathered during this project. A few recordings are as yet seen and connected incidentally, in this way the comments, view count and rating count may be unique. Further, the creators of the recordings can make recordings utilized in this project unlisted or private, making them out of reach. To produce the outcomes accomplished during this project, the models can be stacked and assessed with the given YouTube comment dataset, which ought to bring about precision as portrayed.

**2.1.1 Problem Definition**

In the present situation there are large no of videos available on youtube. It is difficult for humans to classify each and every video present on youtube on the basis of good or bad impact, so we need a software which helps us to find the videos which have good impact and useful to subscribers.

**2.1.2 Problem Analysis**

There are large number of videos present over youtube that’s why it is difficult for any person to find best video for his/her requirement .Let’s take an example ,consider that any student want to learn topics of computer networks but there are many videos created by youtubers are present on youtube but he doesn’t have time to open and watch every video but on the basis of comments or feedbacks he can find which channels videos are providing best explanation on each and every topic of computer networks. This is also a time consuming task to check each and every comment of the video. So if any software is available which compares the comments or feedback given to videos to find the perfect channel to study computer networks it helps students to increase knowledge with correct data.

**2.1.3 Solution**

Hyberg and Isaacs have composed a bachelor thesis on the subject of predicting Youtube video like extents from remark opinion. They additionally utilized a straightforward formula in light of the quantities of remarks named positive or negative by supervised learning classifiers for expectation however their preparation was just done on tweets and their classifiers just ordered

**5**

remarks as sure or then again negative. They implemented SVM, logistic regression, and

multinomial Naive Bayes classifiers. Their outcomes showed Pearson correlation coefficients

between the real like extent and the anticipated extent in view of their formula for their various classifiers and Youtube channels that went from - 0.1 to 0.62. Their anticipated like extents were practically all in the 0.3 - 0.8 range, while the real like extents had a bigger spread from around 0.05 to 0.95. This review means to develop their work via preparing Youtube remarks, incorporate more classifiers, remember unbiased remarks for the grouping, and test a few formulas for expectation. Kumar Verma and Sahu have composed a course paper incompletely on the subject of predicting Youtube video-like extents from the feeling of the remarks. They utilized a neural network furthermore, added the remark count and view included adjacent to the remarks' opinion in their forecast. Kumar Verma and Sahu observed that in light of their review the connection between remark feeling and video like ratio was feeble.

We used the Textblob package to analyze the comments of every youtube video. Textblob is an open-source python library for processing text based data. It performs various procedures on text based data, for example, thing phrase extraction, sentiment analysis, classification, translation, etc.

**2.2 Requirements**

Framework necessities Specification might be a record or set of documentation that portrays the choices and conduct of a framework or code application. It incorporates a determination of parts that makes an endeavor to frame the alleged practicality required by the clients to fulfill their totally various clients.

**2.2.1 Functional Requirements**

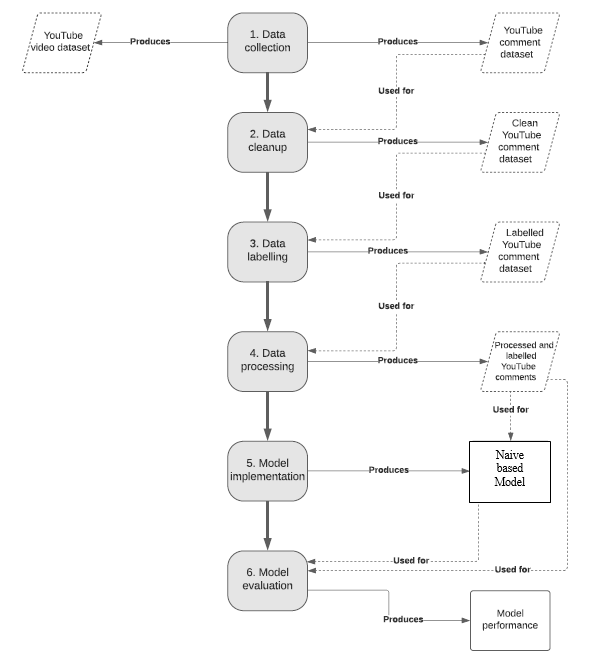
Practical requirements of the portrayal reports and tracks the normal fundamental data to effectively depict business and requests. Practically the depiction report is fabricated inside the perceivability part of attempt. Its fundamental object is to work with, to activate, meander, reinforce, client/client, and any ally you own insight/regard in the required definition framework is required.

**2.2.2 Non-Functional Requirements**

Non-functional requirement (NFR) is a requirement that explains the state of you will generally select framework execution, rather than indicating conduct. Invalid requirements characterize code quality ascribes. The code will be decided for help support, security, capacity, functionality, versatility, accessibility and solace.

**6**

**2.3 Data-Flow Diagram (DFD)**

****

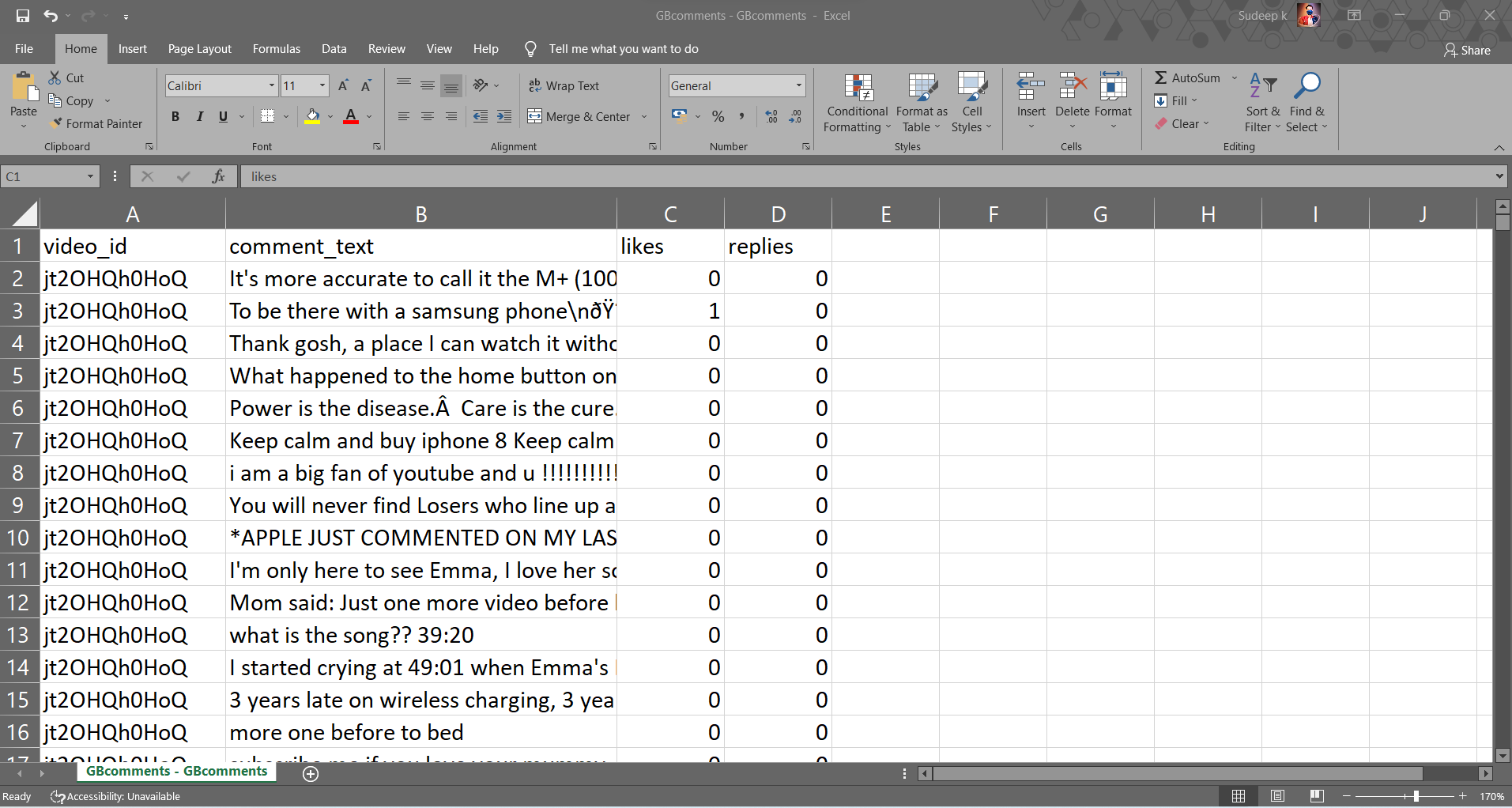
**7**

**Chapter 03**

**IMPLEMENTATION**

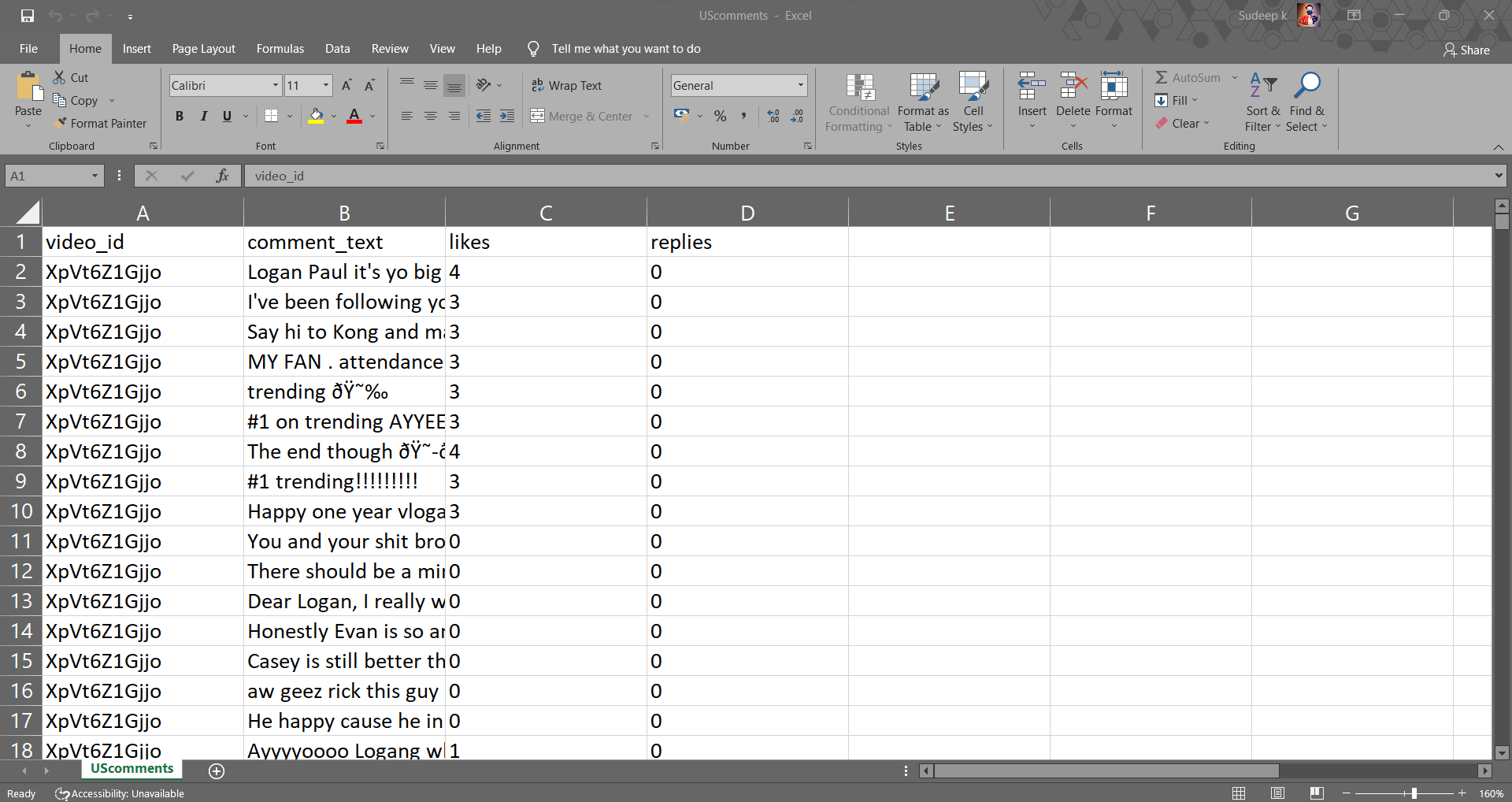
**3.1 Date Set Used in the Minor Project**

GBcomments.csv



**Training Dataset**

UScomments.csv



**Testing Dataset**

**8**

**3.2 Data Set Features**

**3.2.1 Type of Data Set**

GBcomments.csv (Training dataset)

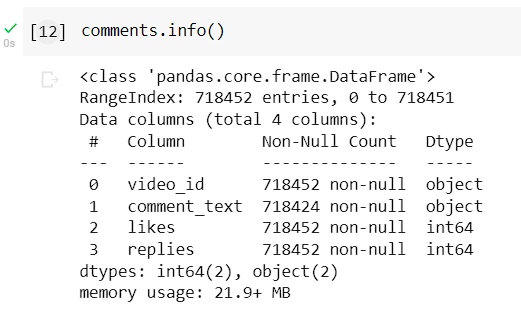
UScomments.csv (Testing dataset)

**3.2.2 Number of Attributes, fields, description of the data set**

UScomments.csv

No. of Columns=4

No. of rows =691408

****

GBcomments.csv

No. of Columns = 4

No. of Rows =718458

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Attributes** | **Description** | **Data type** |
| **1** | video\_id | Id of the youtube video | String |
| **2** | comment\_text | comments of youtube video | String |
| **3** | likes | Count of no of likes | Integer |
| **4** | replies | Count of no. of replies | Integer |

**Attributes of Training Dataset and Testing Dataset**

**3.4 Screenshots of the various stages of the Project**

**3.4.1 Tokenization**

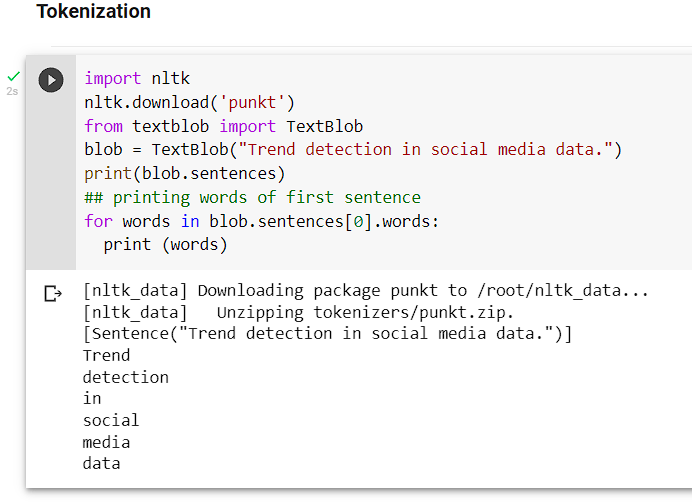
Tokenization alludes to partitioning a message or a sentence into a succession of tokens, which generally compare to "words". This is one of the fundamental errands of NLP. To do this utilizing TextBlob, follow the two stages:

1. Make a textblob object and pass a string with it.

2. Call functions of textblob to do a particular errand.

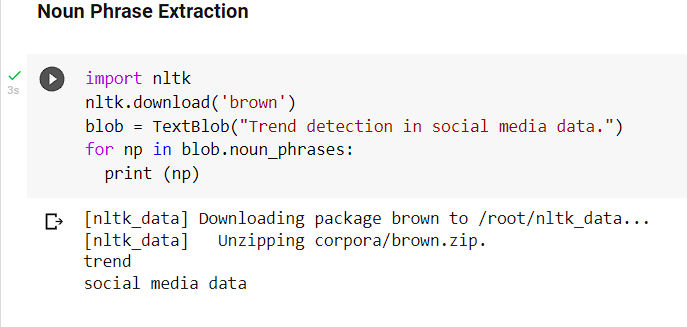
**9**

Along these lines, how about we see this in code



**3.4.2 Noun Phrase Extraction**

Since we extracted the words in the past area, rather than that we can simply remove the noun phrases from the textblob. Noun Phrase extraction is especially significant when you need to examine the "who" in a sentence. Lets see an example below.



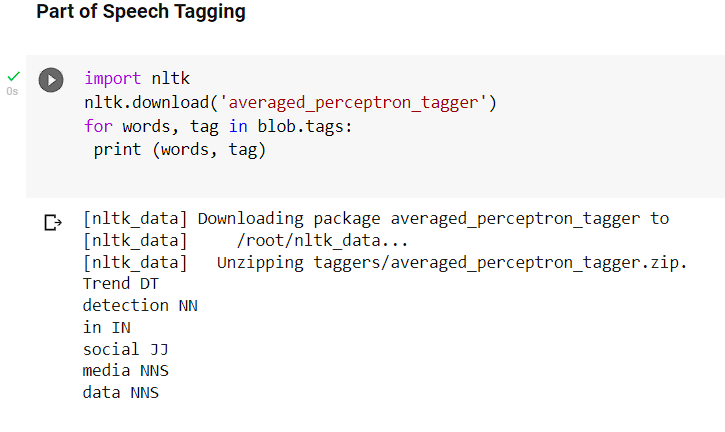
As we can see that the outcomes aren't totally right, however we ought to know that we are working with machines.

**3.4.3 Part-of-speech Tagging**

Grammatical form labeling or linguistic labeling is a strategy to check words present in a message based on its definition and setting. In straightforward words, it tells whether a word is a thing, or a descriptor, or an action word, and so on. This is only a total rendition of noun phrase extraction, where we need to track down every one of the grammatical features in a sentence.

**10**

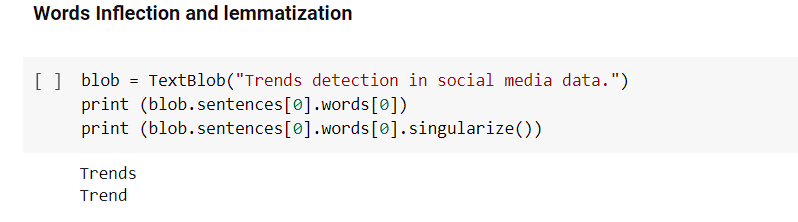
We should really take a look at the labels of our textblob.



Here, NN represents a noun, DT represents a determiner, etc.

**3.4.4 Words Inflection and Lemmatization**

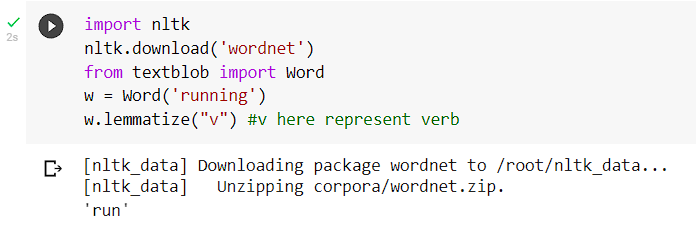
Inflection is a course of word development wherein characters are added to the base type of a word to communicate linguistic implications. Word inflection in TextBlob is exceptionally straightforward, i.e., the words we tokenized from a textblob can be effortlessly different into singular or plural.



TextBlob library also offers an in-build object known as Word. We just need to create a word object and then apply a function directly to it .

We can also use the tags to inflect a particular type of words .

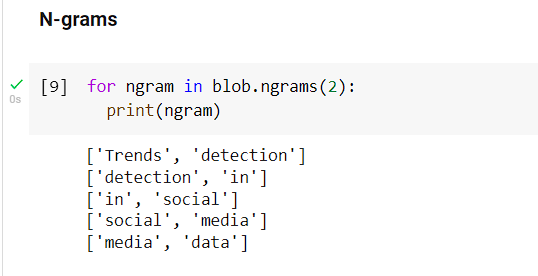
Words can be lemmatized using the *lemmatize* function.



**11**

**3.4.5 N-grams**

A mix of numerous words together are called N-Grams. N grams (N > 1) are for the most part more enlightening when contrasted with words, and can be utilized as highlights for language displaying. N-grams can be handily gotten to in TextBlob utilizing the n-grams function, which returns a tuple of n progressive words.



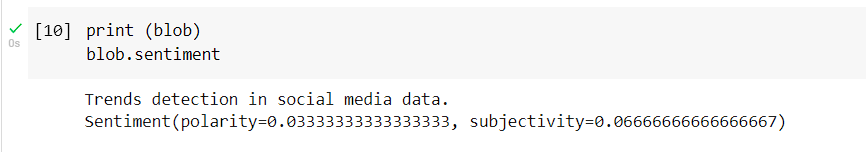
**3.4.6 Sentiment Analysis**

Feeling examination is basically the method involved with deciding the demeanor or the feeling of the author, i.e., whether it is good or pessimistic or unbiased.

The sentiment function of a textblob returns two properties, polarity and subjectivity.

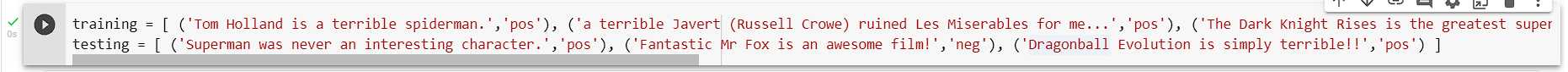
Polarity is a float which lies in the range of [-1,1] where 1 method positive proclamation and - 1 method a negative assertion. Abstract sentences for the most part allude to closely-held conviction, feeling or judgment while objective alludes to verifiable data. Subjectivity is also a float which lies in the range of [0,1].

We should actually look at the opinion of our blob.



**3.4.7. Text classification using TextBlob**

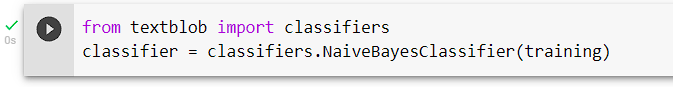
We should assemble a straightforward text order model utilizing TextBlob. For this, first, we want to plan training and testing information.



**12**

Textblob provides an in-build classifiers module to create a custom classifier. So, let’s quickly

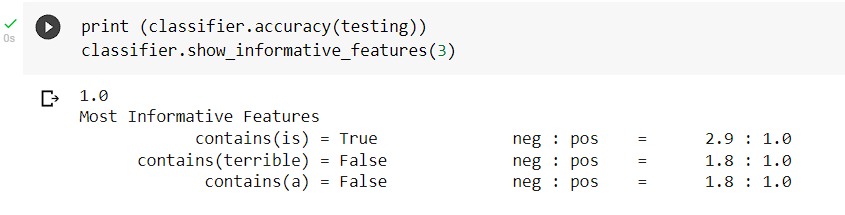
import it and create a basic classifier.



As you can see above, we have passed the training data into the classifier.

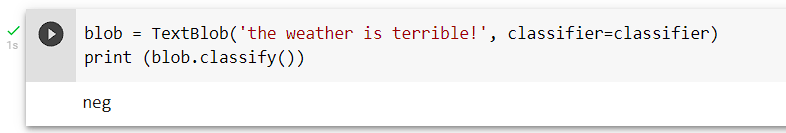
Note that here we have used the Naive Bayes classifier.

Now, let’s check the accuracy of this classifier on the testing dataset and also TextBlob provides us to check the most informative features.



As, we can see that if the text contains “is”, then there is a high probability that the statement will be negative.

In order to give a little more idea, let’s check our classifier on a random text.



So, based on the training on the above dataset, our classifier has provided us the right result.

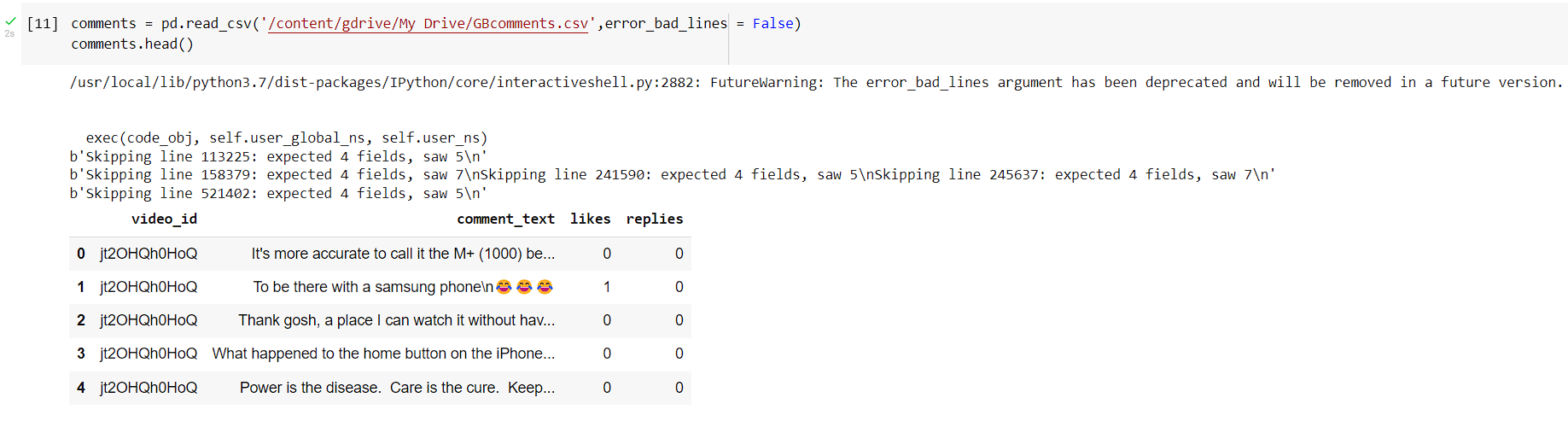
**13**

**How could we achieve it ?**

● Importing every one of the vital libraries



● Investigating the dataset

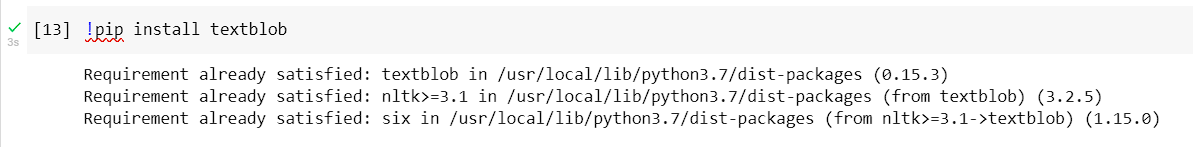


View the dataset utilizing .head() technique.

Presently we played out the sentiment analysis based on the comment\_text feature given in the dataset.

● And to do that sentiment analysis we need to introduce a package called "textblob".

Installing the TextBlob library

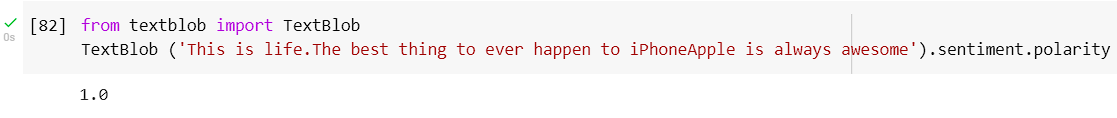


● Importing the TextBlob to check whether the assertion is a positive assertion and to do this we will actually look at the polarity of any sentence. Polarity basically characterizes the direction of the communicated assertion,

○ i.e.; assuming that the sentence decides positive sentiment the worth of the polarity will be over 0 and not exactly or equivalent to 1 and on the off chance that the sentence decides negative sentiment

**14**

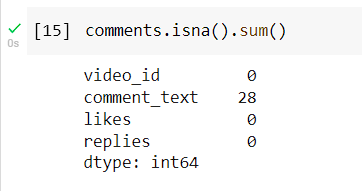
○ then, at that point, the worth of the polarity will be under 0 and more noteworthy than or equivalent to - 1. And in the event that the assertion decides impartial sentiment, the worth of polarity will be between - 1 and 1.



● Cleaning the Dataset

● Checking for missing values

Till this point,I stacked the dataset and every one of the libraries. Presently, here comes the part to observe regardless of whether there are any missing values in the dataset. And I did likewise utilizing this code :-



● Eliminating the missing values from the dataset

○ This shows that we have 28 missing values in the segment comment\_text. So,simply drop those 28 values utilizing the accompanying code :-



● Presently as the missing values are dropped how about we again check whether it reflects in the first dataset.



● Putting away all sentiment of sentence into a variable

○ Here I have no missing values. So we will basically begin actually looking at the polarity of every assertion in the segment "comment\_text" and store that polarity values in a rundown utilizing the accompanying code :-

**15**



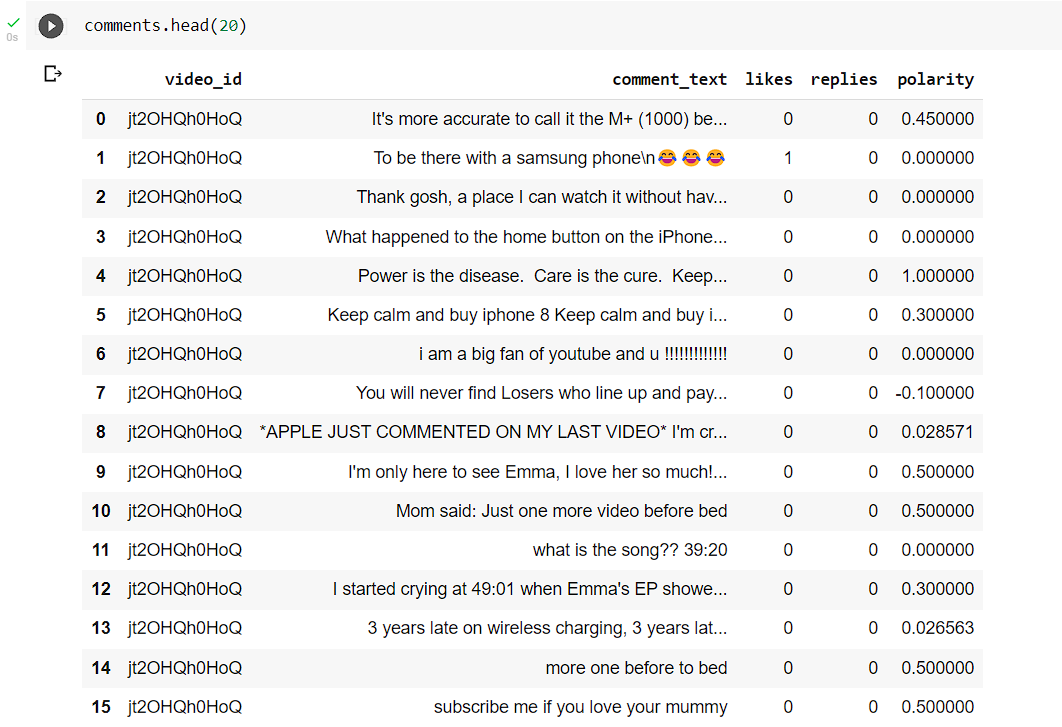
We have put away the polarity values in the rundown named polarity.

Presently added another segment in the dataset and name it as polarity and add every one of the values put away in the rundown named polarity utilizing the accompanying code :-



● Refreshing the remarks dataset

This will make another section named "polarity" in the dataset remarks. Presently we can cross check our dataset utilizing .head() command and see regardless of whether all the polarity values have been transferred.



Here we can see that polarity values have been refreshed in the dataset and the segment name is polarity.

Presently till here I have examined the sentiment of the sentence. So our most memorable errand of performing sentiment analysis is done here. Presently, we need to focus on our next task i.e.; to perform exploratory information analysis on negative and positive sentences.

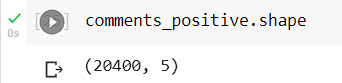
**16**

● Investigating the positive sentences

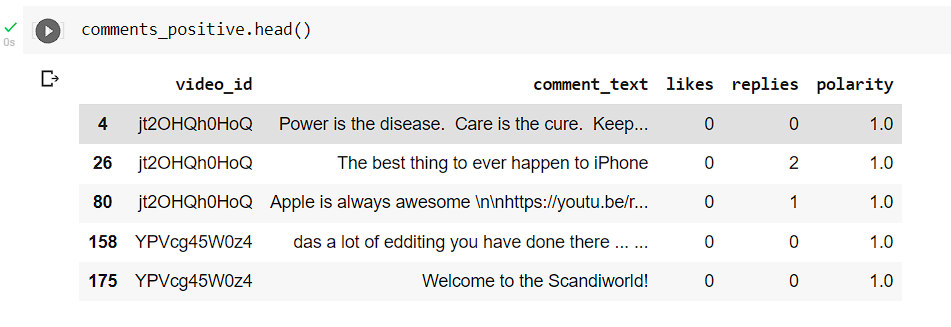


Stacking every one of the positive sentences information in the comments\_positive dataframe

Presently taking a look at the lines and sections in the comments\_positive dataset by utilizing this code :-

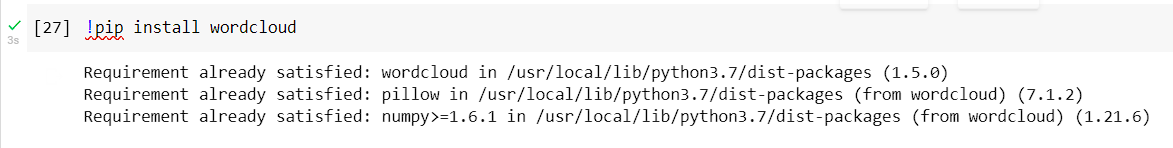


And get to realize that there are 20400 lines and 5 sections and presently we should see the comments\_positive dataset utilizing the .head() command :-



Here you can see that all sentences having the polarity = 1

● Installation of word cloud for perception



● Import the wordcloud and stopwords :-



Subsequent to importing fundamental libraries we will allot stopwords the worth of STOPWORDS, utilizing the accompanying code :-

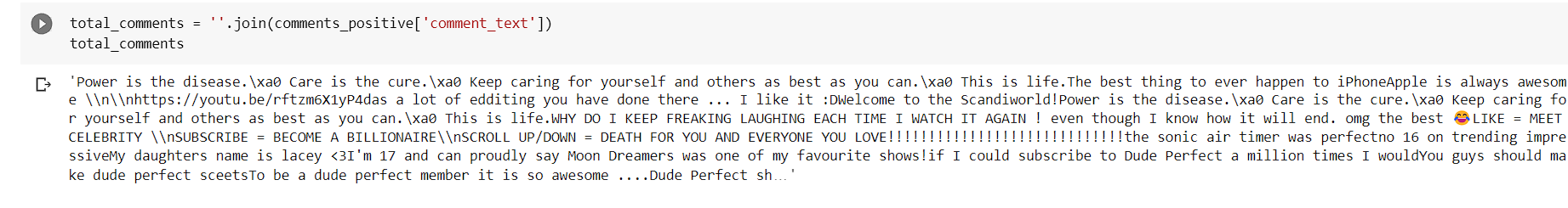


This will take novel words that are not important to us. That is the reason we've taken here to take just the novel words and dole out it to variable named stopwords.

After this, I have taken every one of the sentences that are in the comment\_text into a variable and

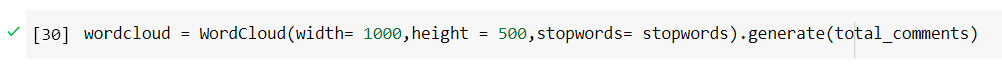
**17**

named it as total\_comments utilizing the accompanying code :-



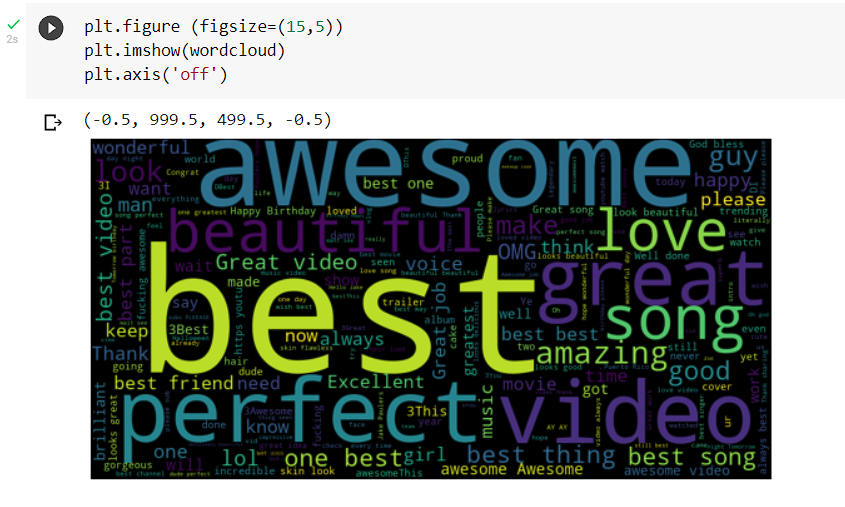
This code will store every one of the sentences that are independently put away in the section comment\_text.

After this, we will shape the WordCloud utilizing the accompanying lines of code :-



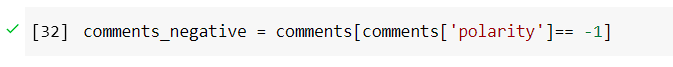
● Picturing involving WordCloud for positive sentences

word cloud utilizing the accompanying lines of code :-



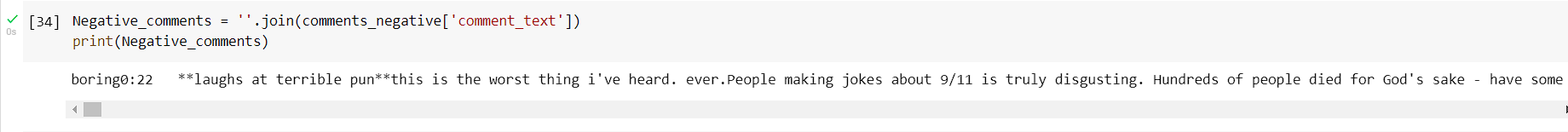
● Investigating negative sentences

Now,I took every one of the sentences having polarity as - 1 and store them in a variable called comments\_negative utilizing the accompanying lines of code :-



Presently, I combined every one of the remarks into one and name it as total\_comments, utilizing the accompanying line of code :-

**18**

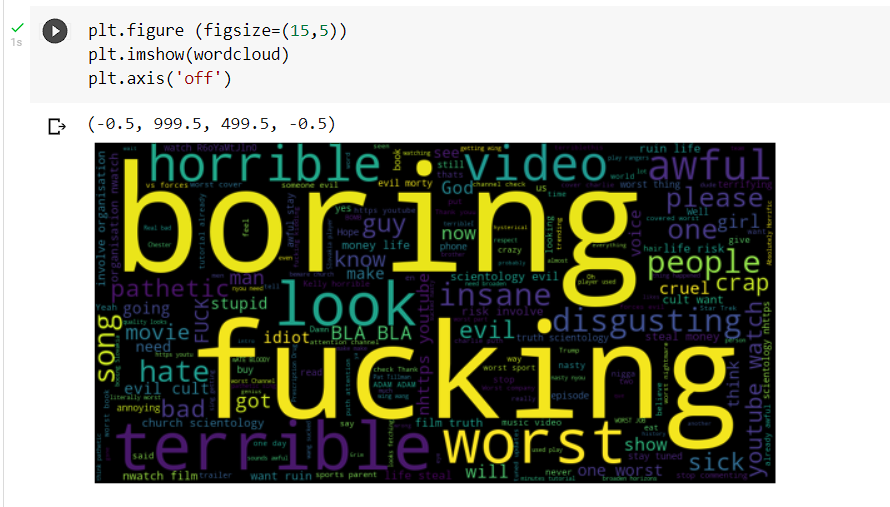


Word cloud code :-



Envisioning involving WordCloud for negative sentences

After this, I just showed the word cloud utilizing the accompanying lines of code :-



**19**

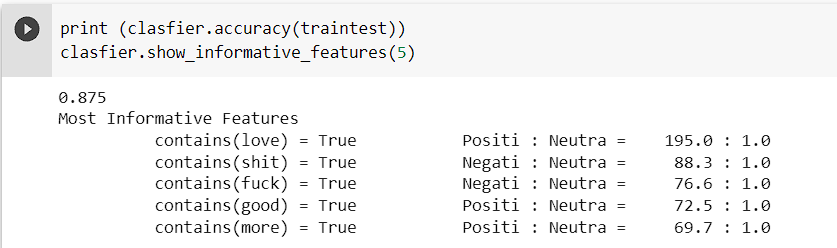
**Chapter 04**

**RESULTS**

**4.1 Discussion on the Results Achieved**

A model has been prepared for the sentiment analysis of the youtube remarks followed by the pre-processing of the dataset. Preprocessing incorporates data labeling, lowercasing of the text, stopwords evacuation, data parting, highlight extraction.

For the sentiment arrangement into two classes positive and negative calculation has been utilized and accomplished a precision score of 87.5% on test data.



Grouping of general occasions and discovery of Sentiment Polarity of client remarks in YouTube is a difficult undertaking for analysis up until this point. A ton of work is done in such manner yet at the same time has far to go to conquer this issue. In this task, We have underscored on the accompanying issues to observe the polarity of remarks given by the clients of YOUTUBE.

1)Current sentiment word references having constraints

2) Achieve palatable grouping execution

3) Challenges including web-based entertainment sentiment analysis.

4) The dataset isn't excessively perfect.

Youtube remarks can’t be greatly trusted as they can have a place with any area, and this textblob strategy just gives polarity in light of the text and not in connection with the setting of the video. For instance on the off chance that the video is about a cell phone audit and a remark is "today is a great day‟ then this remark will get polarity 1 and included in the positive remark part. Though, these remark are not the slightest bit connected with the audit of cell phone

**4.2 Application of the Minor Project**

* This project can reduce Negativity on youtube. For example if a video is viral on youtube then by using our ML model we can know whether this video is good for the community or not.
* Creators can improve their videos by knowing which video is having a good impact on students and on what video is having a bad impact on students.

**20**

**4.3 Limitation of the Minor Project**

* In this project we haven’t  handled Emojis . Nowadays Emoji’s are very popular out of  10 comments 3-4 comments have emoji in it . So if we want to increase the accuracy of models we should also consider emoji’s in it.
* If some users are not using English language instead they are using their Regional language then our model will not be able to identify the language .

**4.4 Future Work**

* In this project we can also include Emoji . Nowadays Emoji’s are very popular out of  10 comments 3-4 comments have emoji in it . So if we want to increase the accuracy of models we should also consider emoji’s in it.
* If some users are not using English language instead they are using their Regional language then model will not be able identify the language so, we can do one thing we can use open source API of Google to convert any language to other languages.

**21**

**References**

* R. F. Alhujaili and W. M. S. Yafooz, "Sentiment Analysis for Youtube Videos with User Comments: Review," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 814-820, doi: 10.1109/ICAIS50930.2021.9396049.
* Hazarika D, Konwar G, Deb S, Bora DJ. Sentiment Analysis on Twitter by Using TextBlob for Natural Language Processing. InICRMAT 2020 (pp. 63-67).