Smart Monitoring of Comedian Transcript

Major project report submitted in partial fulfilment of the requirement for the degree of

Bachelor of Technology

in

Computer Science and Engineering

Submitted by:

Rishabh Sharma (181280)

UNDER THE SUPERVISION OF

Mr. Surjeet Singh Assistant Prof. (Grade II) Of Dept. Of CSE & IT

Department of Computer Science Engineering and Information Technology Jaypee University of Information Technology, Waknaghat, 173234, Himachal Pradesh, INDIA

TABLE OF CONTENTS

CERTIFICATE

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report entitled " Smart Monitoring Of Comedian Transcripts" in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering/Information Technology submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from Jan 2022 to May 2022 under the supervision Of Mr. Surjeet Singh Assistant Prof. (Grade II) Department Of CSE & IT.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Rishabh Sharma (181280)

This is to certify that the above statement made by the candidates is true to the best of our knowledge.

Mr. Surjeet Singh
Assistant Prof. (Grade
II)CSE & IT

ACKNOWLEDGEMENT

I take the opportunity to express my heartiest thanks and gratefulness to almighty God for their divine blessing makes it possible to complete the project work successfully.

I am grateful and wish my profound indebtedness to our project supervisor Of Mr. Surjeet SinghAssistant Prof. (Grade II) Dept. Of CSE & IT, JUIT, Waknaghat for his deep Knowledge & keeninterest in the field of Data Science to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project. I would also like to generously thank each one of those individuals who have helped me directly or indirectly in making this project a success. Finally, Imust acknowledge with due respect the constant support and patients of my parents.

Rishabh Sharma (181280)

Computer Science & Engineering
DepartmentJaypee University of Information
Technology

List of Abbreviations

Py	Python
Np	Numpy
Pd	pandas
NLP	Natural Language Processing
Pck	pickle
DTM	Document Term Matrix
TS	Transcripts
BS	Beautiful Scope
STL	ScrapsFromTheLoft
SW	Stop words
TW	Top words
SA	Sentimental Analysis
TM	Topic Modelling
TG	Text Generation

List Of Figures

ABSTRACT

Data science is simply asking the right question, visualize and looking for a right approach It gives you the answer to the question, “What is likely to happen?” through Predictive analysis. This is study of computer algorithms that improve automatically through experience.
There is a great amount of value added to the company through various ways. Forecasting, estimating the future based on past and present data. Predictive Modelling, performing predictions more granular, example, “Who are the customers who are likely to buy in the next month?”. Data Science is a method of teaching machines to learn things and improve predictions/behavior based on data on their own.

Data Science is being used in various platforms like, social media, Banking, E-commerce, Search Engines, etc. And its evolving and growing for decades. This project comprises a collection of datasets using ScrapsfromtheLoft and thereafter cleaning of this dataset. Further with the help of various analysis algorithms we cover many aspects of public opinion. Firstly, weimplement sentiment analysis. Wherein we extract the sentiment from the dataset and formulate meaningful observations from it. Secondly, we perform topic modelling which helps in identifying the main topic from the tweets captured. For this we use the Latent Allocation Algo; it takes into consideration each document as some topics in a fixed proportion.

CHAPTER 1: INTRODUCTION

1. Introduction

Natural Language Processing what is it well I think of NLP in two different parts there's a naturallanguage piece and the processing piece so for natural languages one of these three here is not like the other and we can probably tell pretty quickly that Python is not like the other two so English and Chinese are natural languages because these were languages that were created naturally over time that people use to communicate with each other whereas Python is a coding language and it was specifically created for coding so English and Chinese are natural languages because they're used for communication and then for the processing piece if you think about a processes are in your computer it's how a computer carries out instructions so if you put it all together NLP is basically how a computer is able to deal with language or text data and natural language processing falls under the greater umbrella of artificial intelligence which is just a computer performing tasks that a human can do so if you think about it like you might have heard of computer vision that's a computer trying to be a human's eyes and see objects for naturallanguage processing it's a computer that's trying to interpret text just like a human would.

Some of the import terms we used in the entire projects are as follows:

Some essential metrics which we used a lot in project-
The New Comedian Ratio is a useful measure of whether we're having trouble attracting newtranscripts of each comedian.
The retention rate indicates how many comedians we analysis at a given period. We'll demonstrate no. of top words and stop words .

DS Model (Data Science Model):

Programming: That's computer and computer science and knowing how to code. This is the mostbasic skill required by the data scientist.

Math’s and Stats: This includes some linear algebra and calculus some statistics used to solve themaths algo and machine learning problems

Communication: After we done all that number crunching and coding can we wrap it all togetherin a story and communicate our insights and there's this one part here that we just want to mention that is the danger zone of data science so if we are really good at programming and we also have communication skills, but we don't have the math and stats background this is what wecall the danger zone

And all three depend on each other directly and indirectly. And we studied on them in brief with the help of calculations and the visualizations in this project

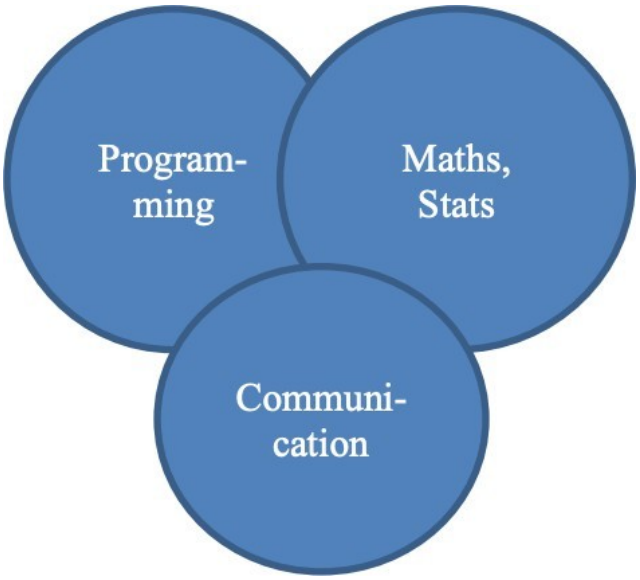


Fig.1 DS Model

2. Problem Statement

We all go through Internet everyday and there is a lot of content some is restricted some is public and some of that is private. As internet is accessible to all and because of which the restricted contend which is not for children that is of age below 18 and in comes countries below 21 so we mate a model of using natural language processing in which we use different techniques like sentiment analysis, topic modeling text generation and LDA. Using these all techniques we can categories the speech or words in speech, which will be suitable for which age group. By this we can recommend them the videos based on their filter, which they want to see based on their preference. In pour model we mainly used the data set of Ali Wong and some other comedians and filtered them on bases of the negative words or restricted words in their comedy. This model can be linked and run on different sites like YouTube also so that so that the individual who don't want to listen to negative words or abusive words will not get to see and interact with restricted content.

3. Objective

The main objective of our project is to help analyze and research to get insights of how to analyze the speech input and process it to filter it on the bases of viewers preference that they want comedy with restricted words or content or not and in our model we mainly focused on inputs of different comedian. So we considered many aspects in our mind to analyze them and tried to bring up the results and in the end with the best possible model and conclusions.

So, here the main motive is to analyze all the data deeply to get best result and see what is beneficial to the audience and what's not and then act basis on that only to enhance the growth ofthe comedy content and to find the proper path also on which direction audience should concentrate their focus and on which content audience should avoid wasting their energy. Many times, audience give all their efforts and subscription in the wrong direction for a long period of time without realizing that it's not working well for the viewers and they need to change their

strategy about it to get the best result and audience can have the best content of their preference. So keeping all that in our mind we should take actions.

This project where we have done all the analysis from selecting best content for the viewers frominternet and data from scraps from the loft to predict the next comedy or video for building a bestpossible model to get the best result out of it, will really help a lot of audience to select the right content and come up with the most beneficial for viewers. What we are doing is data driven growth which means we took a dataset and then did a deep analysis on all the possible aspects and came up with the best possible result to enhance the growth of a content on basis of user preference which is the main objective of this major project.
And we have done all this using data analysis, machine learning, artificial intelligence, Natural Language Processing algorithms (Sentiment analysis, Topic Modeling, Text Generation and LDA) along with algorithms (like linear regression, logistic regression, k-means clustering etc.), elbow method, Jupyter Notebook, Python Programming language, and other libraries imported which will be discussed in the coming pages.

4. Methodology

1-) Data Cleaning

We first took data from scraps from Scraps from the loft which has transcripts for a bunch of standup comedy routines and then We also used IMDB data to filter down the data so to specifywhich comedians We should look at and this step is a very difficult step in the data science project it's where we have to define the scope based on your domain expertise and so there's no right or wrong answer to this step it's all about what you think makes the most sense and as longas you can prove that makes the most sense then you've got it right and so at the end of the day this is how I decided to limit my scope. We looked at comedy specials from the past five years ones that had at least a 7.5 rating with over 2,000 votes on IMDB so that meant it's a good comedy special and it's a popular one and then if we alchemy Dean had multiple specials, I just kept the one with the highest rating to make my life easier so at the end of the day. We

ended up with these twelve comedy routines and if you take a look at them, you can see that they make sense these are some popular stand-up comedians and in the way that we actually got this data is.We scraped this data from scratch from the loft and we used a couple of different Python librariesto do this first we used the request library and the formal definition of what it does is it makes HTTP requests but the simple way to think about it is a request does is it finds a website online and it's able to pull information from that website that's what requested for me. Then we use beautiful soup and what beautiful soup does is it takes all the content of a web page and it's able to extract certain parts from that web page so let's say I just want the title or just want a certain image I can get that with beautiful soup and then finally I once I got all this data. We pickled it sopickling is basically taking an object in Python and then saving it for later and we are going walkthrough all of this in a do Paterno book at the end of the section as well okay so that was my day they're gathering piece. We were able to scrape all that data the next thing we did was we had to take all that scrape data and clean it in some way so the goal of this step is to get the data in a clean standard format for further analysis and different types of analysis actually require the data to be in different formats so, We put all the data in two different formats one is just a corpus which we 'll talk about. We were in a bit and then another is the document term matrix okay so first let's talk about this corpus so a corpus is a fancy name for a collection of texts and this is what it looks like so We have a comedian Ali Wong and then transcript all in its RAW format there and so this here is my corpus and the way we got our data into this corpus format is we used pandas which is a Python library for data analysis and we put all of this transcript data into a data frame which is essentially a table within pandas alright so that was the first format which is a corpus it's pretty straightforward all you have to do is you take all of your transcript data and you put it into a table or we put it into a data frame this next format is a little bit more complex it's called a document term matrix and to create this document term matrix we need to follow these three steps the first thing you do is we need to clean the text so we’re going go over some popular data cleaning techniques the next thing you need to do is tokenize the text so break downthat huge transcript text into smaller pieces and then finally we can put it all into a matrix so that a computer can process it okay so here is a line from john Mulvaney special alright petunia okay so it looks pretty straightforward to a human but if we were to give this to a computer the computer would have a hard time reading it so what we need to do is take this raw

text and put it into that document term matrix format okay so the first thing we're going do is want to clean this data and there are so many ways that you can clean data but whenever we working with text data ,We always start with a couple the first thing we do is to remove punctuation. We always lowercase all our letters and then I remove numbers as well so there are many other things you can do at this point but it's good to just start with something simple and see if you're going to be able to get interesting results from it okay so if we do that then my text looks like this so it's already starting to look a little bit simpler for the computer to read and the way we do this is within Python We're going use a library for regular expressions so a regular expression is think of it as a really-really powerful control find so you know when we're workingwith some document in Excel or Microsoft Word and we're trying to find a word and you do ctrl F and then you can find a certain word well with regular expressions what you can do is instead of finding a word you can actually find a pattern so any word that starts with a capital letter or any word that's this many characters things like that so it's just a really powerful find technique and we're going to use regular expressions to do some data cleaning okay so now that you have the data cleaned the next thing to do is you have to tokenize the data and what tokenization means is splitting your text into smaller pieces and each piece is then called a token so the most popular way to tokenize something is to tokenize in two words we can also tokenize it into thingslike sentences that are a little bit lengthier but again the most popular thing is words so that's what we're going to do here we've taken all the words from the data cleaning staff and we've tokenized them into these individual tokens here and now that every word is on its own we can do some filtering so what we can do here is filter out words that have little meaning so these are called stop words so think of things like a or the or on those won't have that much meaning and we can get rid of those for now so if we do that then we end up with just a few words and so at this point we've tokenized our text and remove stock words so right here this is called a bag of words model and what that means is it's a really simplified way of looking at your text data and it's just throwing all these words into a bag where order doesn't matter even though it's such a simple way to a simple way to represent your text it is really powerful as a first round of analysiswhen you're doing NLP okay so now that we've talked in eyes our text what we're going to do is we're going to put it all into a matrix and the reason we need to put it all into a matrix is because now this petunia example I'm only showing you one line of text but we have text for a bunch of different comedians so we need to put it into a matrix so that the computer can process the information for each comedian so that final matrix would look something like this so you can

seeon the left we have in every row we have a different comedian and then every column is a different word and you can see John Mulvaney he says right petunia wish me luck and then Ali Wong she says hello Ali and thank you and things like that and so this right here is called a document term matrix each row is a document each column is a term and all the values inside areword counts all right so this is our second format the - min term matrix and the way to create thisis using two libraries first you can use scikit-learn where actually it's just one library you can usescikit-learn which is a Python library from machine learning to create this there's a specific function called count vectorizer and what that allows us to do is create a - mentor matrix all rightso just to reiterate the goal of all this was to get all of our data into a clean standard format for further analysis and we put it into two types of standard formats the first was a corpus as you cansee here we're just putting the transcript data into a table and the second is this document term matrix where every row is a document and then every column is a different term and then all the values are word counts so the input into this data step was the question how is Ali Wong differentI was able to gather all the data by scraping data from a website and I was able to clean all the data by putting it into a standard format and now the output of this step is going to be a corpus and a document term matrix so what We're going to do now is we are going to jump into a Jupiter notebook to show you how, We actually went through these steps all right here's the data cleaning Jupiter notebook again you can find it on GitHub under a dash of data okay so for this data clean the step what I recommend is that you go through it and detail yourself We are going go through the high-level here and then hopefully they'll give you a good enough background to then go into everything in detail so again for data cleaning.

We shared it with my problem statement and then the next thing I wanted to do is gather the data and if we remember from the presentation the way we did this was we scraped some data from a website and used a couple libraries. We used requests, We use beautiful soup We're also using pickle and request some beautiful soup allow me to scrape all that data and pickle allows me to take all our scrape data and save it for another notebook so this was a function that we wrote thatrequests the data from a URL and then what it does is it supervised that text and then in this casewe specifically looked for anything on that website that was of the class post content and this is specific to the scraps from the loft website that we were scraping from .We found that all the comedian transcript text was in sections that had a class of post content and so this is a function

we wrote that specifically works for the scratch my loft website and then what we did was we putin all the URLs of the transcripts that we wanted to scrape and you can see if we uncomment thisthis line right here will actually scrape all of the websites for me and pull out those transcripts and then at this point in this step I've pickled those files for later use We've commented this out for now because We've already pickled the files and so after you pickled the files you can also load them back in to have your Jupiter notebook work with and then what I always do now that I pulled in this data into data so again data has all of my transcript data in it what we are going to do now is just make sure that my data has been loaded properly so if we look at data my data is adictionary where my keys are all of the comedians and my values are it's all the transcript data soif we look at data and we look at the keys you can see we have the 12 comedians right here and then if we specifically look at Louie and we look at the value for him you can see here is all the transcript data for one of his comedy routines okay so that was the data gathering step which wasall about scraping the data from a website next is the data cleaning step and for the data

Cleaning step this is an iterative process it takes a really long time to clean text data and what wetypically do is, We start by using some common data cleaning techniques such as making everything lowercase removing punctuation and so on and then I take a look at how my data looks and then I continue some other okay I could do using some other data cleaning techniques from there so we're going do here is first we always like to just look at my data again my key is Lois here and then if you take a look at my value it's all the transcript data okay so the first thing I'm going to do is if you looked at this transcript the way was scraped was that instead of having all of the paragraphs or different sections of the comedy routine in one value it actually gave me a list of texts and so the first thing we're going to do is we're just going to combine all that text into one large chunk of text set this step here combined text and what I'm doing is I'm taking thatentire list and we're putting it all together alright so that is my function to combine all my text and then this next section we actually combine it okay so at this point you can either keep your data in a dictionary format if that's what you're used to but ,We really like using pandas and so we put everything into a panda's data frame and you can see here that I've taken my dictionary and We've converted it into a data frame so now we have this really nice data set that has all of the comedians and all of their transcripts okay so let's take a look at the transcript for Ali Wong so we specifically looking where index is equal to Ali and you can see here is all of her transcript

data and one value all right so now that we have that we're ready to apply a first round of text cleaning techniques again this is an iterative process so I'm going to show you two rounds of textKing that I did but you can absolutely do more so I'm gonna import two different libraries here one is regular expressions again regular expressions are used to find patterns and text and then string in this case We're going to use it to look at a bunch of punctuation marks so We've createdthis function and what it does is it makes all of my text lowercase it looks at all these punctuationmarks and it replaces them with nothing it looks at oh sorry in this case it looks at all the punctuation marks and it replaces them with nothing in this case it's looking for anything that's inthose square brackets and replaces them with nothing and in this case it's looking at anything thathas some digits and is surrounded by either text or digits basically any token that has a numeric value in it or a digit in it and it's going to replace that with nothing as well okay so this is my firstround of cleaning texts and if we apply this to my data frame then you can see here my text already looks a lot cleaner everything was lowercase I have removed things that have numbers inthem we removed punctuation and so on but if you look here there's still more that I can do in fact if you look hearing specifically it looks like my string dot punctuation lists did not include these specific type of quotes and so what I'm going to do now is apply a second round of cleaning and in this case I'm going to specifically put in those other type of quotes and then I'm also going to get rid of some of these slash ends that were up here see if I can see any in this sample of text so you can't see in this sample of text but there are these backslash ends that actually came through my text so I'm going to remove those and so now I've applied a second round of cleaning to my text and you can see those special quotes have been removed and those backslash ends have also been removed so at this point you can continue to apply text pre-processing techniques or text cleaning techniques but it could go on pretty much forever so at one point you need to stop and for me these two rounds were a good first pass for me so.

We're going keep moving on with my analysis okay so I've done some data gathering some data cleaning and now We're going organize the data so We will put it into the two formats that I mentioned earlier first is a corpus so just a collection of text and then second is that document turn matrix so the corpus is actually something we've already created so if you look at that data frame from earlier this is the corpus .We have every single comedian along with their transcript and then what we're going do here is just add another column that contains their full names so

we can use that for some of my visual issue visualizations later on okay and at the end of this step I'm going to pickle it or basically save this object for later use okay the second format We're going to create is called this document term matrix and what's going happen here is We are goingto use this count vectorizer and what count vectorizer does is we can call count vectorizer and we can input in my transcript data tell it what stock words we want remember stock words are words that don't have much meaning that we want to get rid of and then after It transform my transcript data into this count vectorizer form we can put it into a data frame and it looks something like this again every row here is a different comedian or in the document term matrix context every row is a different document and then every column here is a different word or a different term and then certain terms have already been excluded which are the stop words like the or a and so on and so now my text is in this document term matrix format.

2-) Exploratory Data Analysis

We've done the first two steps of the data science workflow we've come up with a question which is how is Ali Wong's comedy routine different than everyone else's number two is we've gathered data and we've cleaned it and we put it into some standard formats for NLP and we are going to be going through exploratory data analysis so the input into this step is data in that standard format so either as a corpus or as a document turn matrix and what we're going to do during this step is we're going to summarize the main characteristics of the data that means is wewant to take a look at the data and see if the data makes sense or if there are trends in the data at all and if you remember from a previous example this is actually typically done using visual techniques so in my example earlier we had a table of data but we put it into a scatter plot like this so you can figure out if there are any trends immediately in the data and with example beforewe saw that yes there was this upward trend so now the output of the EDA step:
First of all find these trends and then second of all get comfortable with your data and see if it makes sense so before we move on to any fancy data science techniques we always make sure we do EDA first okay so let's dive into EDA so again my question is how is Allie long different from other stand-up comedians and some ways that you can think of to explore this data right

away is maybe looking at the top words that she uses versus other comedians you can also take alook at her vocabulary we just have a small vocabulary a big one what about other comedians and finally We're going to look at amount of profanity so originally we were doing this analysis. We didn't think that profanity would be such a big part of a stand-up routine stand-up comediansand their routines but it did end up being a big part which why I've included it here okay so let's go through the steps of how to do EDA for each of these questions the first thing I'm going to dois look into top words so what are the top words of every comedian and first thing we need to figure out is which data format should we be looking at to figure out top words should we be looking at a corpus or should we be looking at a document term matrix so there's my corpus there's my document term matrix now that you've gotten comfortable with these you probably know that this document term matrix would be perfect for this situation because we already havefor each comedian all of the words that they use and how often they use those words so now thatwe've gotten the data the next thing we need to do is we need to aggregate that data in some wayso how can we a get this data to find the top words well for Age comedian .We can select the columns that have the largest values and now that I've aggregated them the next step is to visualize this so let's say we found the top 30 words for each comedian how could we visually communicate this well we could maybe create a bar chart or in this case I've created some word clouds and then finally what are some insights we can get from this so now that we have a bunchof worked clouds we can visually answer a few questions does the day to make sense and does further cleaning need to be done and then also what are some initial findings and how the comedians are similar or different okay so those were the four steps of EDA data aggregate visualize and insights.

Now what I'm going to do is We're going follow these EDA steps for all three of the questions I've had here so what are the top words for every comedian what is the vocabulary look like for each comedian and what is the profile level for each comedian and the way I'm going to do this is I'm going to use visualization techniques so We're going to use the word cloud library to create word clouds in Python and then I'm also going to be using matplotlib which is a standard data visualization library in Python so to summarize the section before I go into the code the input into EDA the EDA step is data in a standard format so for NLP that's going to be a corpus or document term matrix and then for EDA we're going to summarize the main characteristics ofthe data using the four steps here and then finally the output to EDA is just getting comfortable

my data and seeing if it makes sense so let's walk through this Jupiter notebook again you can find all my code on github a dash of data all right so here's the EDA notebook and what we're going to do is I'm going to answer these three questions that I've listed out and first I'm going to start with most common words so to do this the first thing I'm going to do is I'm going to pull in my data from my previous notebook so if we Remember from the data cleaning step I Pickled my data which should which means that I've taken that data and We've saved it so that another jupyter notebook can then read that data and you can see that here i read in my document term matrix and now we have every single document - sorry we have every single document every single word and we transposed it in this case to make my next steps a bit easier okay so again mygoal is to find top words here so what We're going to do is We are going find the top 30 words said by each comedian and the way I'm going to do that is .

We're going to create a dictionary and every key in my dictionary is going to be a comedian and then every value in my dictionary are going to be their word a word and how many times they say it and then I'm going to sort all of that so now .We have for every comedian this word how many times they say it and this is the most frequent word their second most frequent word and allthe way down and here We're going to print the top 15 words said by each comedian just to get an idea what it looks like so as I'm looking through this I see that there are still words in here thatdon't have much meaning so even though I've gotten rid of stop words in my previous step we think we can include a couple more stoppers in there especially words like like so on so what We're going to do down here is

I am going to add words to my stop word list so the way I'm going do that is we are going to see if looking at these top 30 words for each comedian if a lot of comedians say those top 30 words then We're going to remove them so these are all the top words said by the comedians and out ofmy 12 comedians all of them have like at the top word all of them have I'm as a top word and so what we are going do is We're going to get rid of them and so I've created this list comprehension here that says if more than half of the comedians have these words as their top word add it to my stop word list so you can see this I'm doing here so what we'll doing is pullingin my data we'll be adding in some stop words we're going recreate my document term matrix toremove those stop words as well and we're going pickle it for later so now at this point I've I have my data set and I've clean it just a little bit further before going into EDA and again for thisstep my goal was to figure out what are the top words said by each comedian and can I show that in a visual way and so I'm going to do that using word clouds and so in this case right

here. We've import a word cloud and these are a bunch of parameters that you can tune with in word cloud based on how you want it to look and then in this point here I'm going to actually plot it and create a sub plot for each comedian so if you see my plots here this is what every comediantalks about and at this point this is when I found out that there's a lot of profanity in stand-up comedy and so that's why we dig into that later but the findings here where I find that we found that Ali Wong says the S board a lot and talks about her husband and then a lot of people use thef-word and more into that later so that was the

EDA step I did was I looked at top words and I create award clouds the next thing I want to look at is vocabulary so what we want to do here is we want to know do some comedians have a bigger vocabulary than others so we're going to find the number of unique words that each comedian uses and the way we're going to do that is We going to We're going to look at unique words or first we going to look at all the nonzero items in the document term matrix so that means those are words that are actually in a community's vocabulary and then we are going to find the count of those and the way I'm going to and then after that I'm going to put it all into a data frame and so then you get this so We have it for every comedian what's the number of unique words alright so here I have the number of unique words for every comedian and you cansee this person is pretty low and this person is pretty high and then I decided to take it a step further and also calculate the words per minute of each comedian so do they talk fast or do they talk slow and the way I'm going to do that is we found the runtimes for every single um comedy routine and I added that as a column to my data frame so I have every comedian that unique words we also have their total words that they used here are the runtimes and then the words per minute is just the total words divided by the runtime and then you can see this is how quickly they speak okay so now that all this is a table what I'm going to do is I am going to visualize it insome way so that's easier to interpret and the way I'm going to do that is creating two different subplots and you can see I have number of unique words and number of words per minute or the speed at which someone speaks and what we found is that in terms of vocabulary this one the leftRicky Gervais and Bill burr use a lot of words Louie CK and Anthony Lesnick have smaller vocabularies and then for chocolate these two talk quickly these two talk slowly and then Ally Wow who is the person that I'm most interested in is somewhere in the middle so this was some EDA that I've done but unfortunately there's nothing too interesting here for me to report so now

we're going to move on to amount of profanity okay so earlier I said that we come back to this because we just found a lot of profanity the word clouds and if you look at the most common words you definitely see some profanity in here too so I'm kind of isolate just these bad words and you can see here for every comedian this is how many times they say the f-word and the s word in their routines and when I see two columns like this is perfect for a scatterplot so here I'mcreating a scatterplot and you can see this is the number of bad words using a routine here's a number of s words and here's the number of F words you can see Joe Rogan and Jim Jefferies there's a lot of efforts have a routine okay so looking at this I found that there are some people who swear a lot but then there's also some people who have cleaner humor so you can see here s words Ali Wong uses the S word a lot and what was really interesting as Mike Birbiglia actually has no profanity in his routines and he's another comedian that I kind of like and so you can see that maybe I like comedians that don't swear too much so that was an interesting part of this EDA. So now I've done three types of EDA and again what was the goal for all this it was be able to be able to take Anna Lanisha look at our data and see if the results made basic sense and my conclusion is it does for a first pass it's not perfect but especially for NLP it's going to take a long time for you to be perfect and clean your data perfectly and so We always try to do as little as I can at the beginning and then We can always come back and make things better so there's my life motto there which is let go perfectionism especially when working with NLP.

5. Organization

Here we have divided our work into some parts. Detailing them out below:

1. Getting to the data we get from Scrap from the Loft
2. Segmenting the base on our filter.
3. Figuring out the positive and negative words.
4. Prediction using churn to increase retention.
5. Best comedy content of user choice prediction.
6. Building Models based on Dataset.
7. Uplifting the model
8. Testing Analytics strategies (A/B Testing)

CHAPTER 2: LITERATURE SURVEY

S.No.	Literature	Topic Discussed
1.	Purohit, H., Hampton, A., Shalin, V. L., Sheth, A. P., Flach, J., & Bhatt, S. (2013). What kindof# conversation is Twitter? Mining# psycholinguistic cues for emergency coordination. Computers in Human Behavior,29(6), 2438–2447.	Data science, data science solution, bigdata.
2.	Y. Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning:detecting, diagnosing, and editing data abnormalities. PLoS Medicine, 2, no. 10, e267	Data visualization, visualization, data analysis,databases.
3.	Yi, S., Li, C., & Li, Q. (2015). A survey of fog computing: Concepts, applications and issues. InProceedings of the 2015 workshop on mobile big data (pp. 37-42). ACM.	Optimisation,Big data, data visualization, data mining.

4.	Stieglitz, S., & Dang-Xuan, L. (2013b). Socialmedia and political communication: A social media analytics framework. Social Network Analysis and Mining, 3(4), 1277–1291.	Data mining, clustering algorithms.
5.	Severo, M., Feredj, A., & Romele, A. (2016). Soft data and public policy: Can social media offer alternatives to official statistics in urban policymaking? Policy & Internet, 8(3),	Comedian segmentation , clustering, datamining, pattern clustering.
6.	Jha, R. (2018). Regional inequality and indirect tax reform in India. In Facets of India's economyand her society volume II (pp. 119–148). London: Palgrave Macmillan.	Unsupervised machine learning, customer segmentation,spending behavior, data models
7.	S. IBM (2019) [Online], Available: h from thescrpas from the loft IEEE (pp:8975) , Ref -09873	Comedian Transcripts ,customer segmentation,predict algorithms,logistics

8.	Yi Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2018). Polarization and acculturationin US election 2016 outcomes–can twitter analytics predict changes in voting preferences. Technological Forecasting and Social Change. https://doi.org/10.1016/j.techfore. 2018.09.009.	Demand response, datamining, load management,feature extraction

1. Existing system

Build relationships with comedian for a fewer challenging means of growing deals, and don't let comedian relationships sour after the original transcripts cleaning. Acquire how to expand transcripts from:

A. Getting transcripts from comedian
B. Utilizing promoting computerization to assemble long haul corpus with comedian
C. Address progressing comedian corpus
D. Foster new corpus focused on your present comedian
E. Upsell and strategically pitch as a feature of your typical advertising technique ratherthan zeroing in just on new comedian securing
F. Make a reliability program
G. Train a comedian driven group
H. Comedian overall existence corpus

1. Research your market

Statistical surveying and investigation open ways to new freedoms inside your current comedian base. By investigating the market, you serve, you can focus on who your comedian transcripts

are and what they need. A few strategies for expanding your insight into your comedian come from an assortment of sources including:

A. Gathering information by directing a scrapsfromtheloft.
B. Getting input through studies, corpus and DTM.
C. Paying attention to what comedians say across friendly stages.
D. Perusing exchange articles and industry-related distributions.

Likewise, take a gander at information from the Bureau of Labor Statistics and the Economics and Statistics Administration. Mine information from Google Analytics to comprehend basic components. You can utilize it to impart your image's story in web-based media, email, live talk, and any remaining channels where your clients may lock in. You can find out about client venture planning here. Watchwords that drive the greatest traffic, just as you are looking for these catchphrases in your scenario.
Set goals that are similar to the different comedian of different activities on your site to decidethe corpus from each channel and each publicizing message, the presentation of your greeting pages, and the commitment of different components on your site Concentrating on your comedian transcripts assists you with taking apart your current comedian manners of thinking,their normal pathways (ventures) prompting buy, and in any event, recognizing neglected requirements that reflect openings for you to stand apart from the opposition.

2. Proposed system

The main proposed system that we have used here is more focused on the new tools and technologies introduced in the past few years. Data science, Machine Learning, Deep Learning, Artificial Intelligence, Natural Language Processing, and other technologies are among them.

We have followed the data science process of initially finding the dataset, analyzing it and employing the model we developed after training and testing to fuel the growth of any business

In the proposed system, instead of just letting the humans do all the work, most of the analysis isautomated which is being done by Jupyter Notebook, Python libraries that we imported like Keras, tensorflow, numpy, panda, scikit, sklearn, and using these libraries to do automation and build a model to visualize how the change is happening in the given dataset.

3. Feasibility Study

Any vital stage in the transcript improvement process has been procured. Permits engineers to get a functioning item that has been tried. Alludes to item exploration that may be done as far as transcripts results, application execution, and specialized help expected to utilize it. A potential examination ought to be completed dependent on an assortment of conditions and conditions.

1. Economic Feasibility

Monetary recuperation is the contrast between the benefits or results we get from an document term matrix and the general expense we spend to upgrade it. The formation of another comedian improves framework accuracy and paces up application and announcing handling in the current framework.

2. Probability Feasibility

The exhibition of a transcripts to make it work is alluded to as accessibility. A few things might perform honorably during analysis and visualization, yet they might break down, in actuality. it involves exploring the required transcripts just as their specialized information. The contained information, refreshed data, and reports for corpus are precise and speedy in the current framework.

3. Technical Feasibility

The expression "specialized execution" identifies whether or not the product right now accessibleis able to totally support the current framework. lt examines the advantages and downsides of using explicit advancement programming, just as its practicality. lt additionally figures out the amount additional time comedian should make the application work. The current framework's UIis easy to use and doesn't require much information or preparation. lt simply takes a couple of mouse snaps to finish exercises and produce reports. Since comedian need fast admittance to sites with an undeniable degree of safety, the product used to update is most appropriate for current applications. This is cultivated by joining a web server and an information server in a similar actual area.

CHAPTER 3: SYSTEM DEVELOPMENT

1. Design and development

In any case, most importantly, why do we do division?
Since you can't treat each user the same way with a similar choice and views. They will find another choice which comprehends them better.

Some methods we used in this project are the Sentiment Analysis, Topic Modeling and Text Generation etc.

- Sentiment Analysis: Let’s say you are a manager of a company that sells hats and also shirts and you want to know what your customers are thinking about your hats and shirts do they have positive feelings about them or negative feelings about them so then you go to your call center and you see that a bunch of people have called in without your hats and your shirts andyou could go through all of these and listen to every single message but that would take you areally long time so instead you can use an NLP technique to automatically tag these as positive or negative calls and then at the end of the day you can figure out that people tend to think that your hats are pretty good and that your shirts are not very good so this concept is called sentiment analysis. In this we use text blob sentiment analysis in which we give polarity to the word in range of -1 to +1 in order to decide which one of them are negative and which are positive.

- Topic Modeling: The task of discovering themes that best characterise a set of documents isknown as topic modelling. Only throughout the topic modelling process will these themes arise (therefore called latent). Latent Dirichlet Allocation is the topic modelling technique (LDA) we used in our mopdel.

There are some other methods also that we have used.

Classification: Classification is the process of identifying a function that aids in the classificationof a dataset based on several factors. A computer programmed is trained on the training dataset and then categorizes the data into distinct classes based on that training. The classification algorithm's goal is to identify the mapping function that will convert the discrete input(x) to the discrete output(y) (y).

Algorithms for classification can be further classified into the following categories:

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machines
- Kernel Support Vector Machine
- Naive Bayes
- Decision Tree Classification

Naive Bayes- It is a classification algorithm that may be used to classify binary and multiclass data. It is a supervised classification technique that uses conditional probability to assign class labels to instances/records in order to categories future objects. It plays an important role in thisproject.

Regression: The technique of discovering correlations between dependent and independent variables is known as regression. It aids in the prediction of continuous variables such as markettrends, house values, and so forth. The Regression algorithm's goal is to identify the mapping function that will translate the continuous input variable (x) to the discrete output variable (y) (y).

Regression Algorithm Types:

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression
- Support Vector Regression

- Decision Tree Regression
- Random Forest Regression

2. Algorithms

Machine learning algorithms which we've used are:

Linear regression: Linear Regression is a simple machine learning algorithm that is used to solve regression problems and falls under the Supervised Learning technique.
It is being used to anticipate a measured process variable using control variables. Linear regression is used to find the best-fit line for predicting the outcome of a continuous variable. When only one regression analysis is used, simple linear regression is being used. If there are more than two types of variables to predict, the Multiple Regression Model will be used. By picking the optimal fit line, the algorithm sets up the correlation between the dependent as well as relationship between the independent variable.

Among many only well Machine Learning techniques used during Directly controlled Learning approaches is logistic regression.. It can be used to solve both classification and regression problems, however classification is the most typical application. Logistic regression is used to predict the categorical dependent variable using independent factors. A Logistic Regression problem can only have two possible outcomes: 0 and 1. When calculating the probability between two groups, logistic regression can be used. For example, whether it will rain today or not, whether it will rain today or not, whether it will rain today or not, true or untrue, and so on. Probabilistic prediction is often used in logistic regression. In this case, the observed data should be considered the most plausible. In logistic regression, we pass the weighted sum of inputs through an activation function that can transfer values between 0 and 1. A type of activation function is the sigmoid function.K-means clustering-based unsupervised machine learning method Using K-Means for Cluster analysis

3. Model Development

Data wrangling: The process of cleansing and integrating chaotic and complicated data sets for easy access and analysis is known as data wrangling. With the amount of data and data sources continuously increasing and expanding, it is becoming increasingly important to organize vast amounts of data for analysis. In most cases, this procedure entails individually converting as wellas mapping data through one numerical form to the other in order to facilitate data consumption and association.

The goals of Data Wrangling are to accumulate information from diverse sources in revealing"profound intellectual capacity." Reduce the time it takes to collect and organize unorganizeddata before it can be used. Allow data scientists and the data analysts to concentrate on data analysis rather than the data wrangling. Senior executives in an organization should be encouraged to improve their decision-making skills.

Crucial Steps in Data Wrangling

- Data Acquisition: Locate and gain access to the information included in your sources.
- Data integration is the process of combining altered data for the future assessmentincluding using.
- Data cleansing entails reorganizing the data into a more useful and functional manner, aswell as correcting or removing any incorrect information.

In monitoring the comedian transcript module we did data wrangling.

Feature engineering: When developing a predictive model using machine learning or statistical modeling, feature engineering refers to the process of leveraging domain expertise to choose andconvert the most important variables from raw data. The purpose of feature engineering and

selection is to make machine-learning (ML) algorithms perform better. The construction, transformation, extraction, and selection of features, also known as variables, that are most conducive to constructing an accurate ML algorithm are all part of feature stuff
Hyper parameter tuning: A mathematical model containing a number of parameters that must be learned from data is referred to as a Machine.
Hyper parameters, on the other hand, are a type of parameter that cannot be learned directly fromthe standard training procedure. They are normally fixed prior to the start of the training procedure. These parameters describe crucial aspects of the model, such as its complexity and learning rate.

EDA (Data Exploration Analysis): EDA is a data assessment strategy that employs a variety of (mostly diagrammatical) methods to optimize comprehension of a data set. This apart from fitting the infrastructure to available information, we can fit the same parameters of the model.by building the classifier with existing data Recognize underlying structure, extract significantfactors, detect outliers and anomalies, exam fundamental assumptions, construct parsimoniousmodels, and identify the optimal factor settings.

4. Mathematical

Here are some important formulas we used:

- Transcripts = Active Comedian Count * Order Count * Average transcripts per comedian
  We used this formula to calculate the transcripts as we consider transcripts as our dataset is long enough.

- We calculated the three types of D.T.M as:
  Conversion D.T.M: Conversion rate of test group - conversion rate of control groupOrder D.T.M: Conversion DTM * number of converted customer in test group Revenue DTM: Order DTM * Average order $ value

We used all these in the uplifting the sales module where we simply calculated the three types raises

- DTM score: $P_{ali} + P_{louis} - P_{james} - P_{jonas}$

  Here in this formula buyers who would only purchase only when they receive approval have been listed here.

  Buyers who would not buy when they're not consulted with just a proposition are madereference

5.Implementation

Dataset used

We majorly used comedian transcript dataset for this project which we obtained from web scrapping, We also used a few other platforms to select suitable datasets like IMDB. All of these datasets were required for the analysis.

The online set of transcript data largely contains all the data of famous comedians; some of the most important sections within This set of data that we are working on is listed ; many of the most essential columns in that data source which we are working on are mentioned some of the most important columns in that dataset that we used

- Name of Comedian
- Unique words
- Total Words
- Run times
- Words per minute

As we can make use of them to predict the polarity

	comedian	unique_words	total_words	run_times	words_per_minute
1	Anthony Jeselnik	984	2905	59	49.237288
3	Bo Burnham	1272	3165	60	52.750000
0	Ali Wong	1341	3283	60	54.716667
9	Louis C.K.	1098	3332	58	57.448276
4	Dave Chappelle	1404	4094	67	61.104478
6	Jim Jefferies	1313	4764	77	61.870130
10	Mike Birbiglia	1494	4741	76	62.381579
11	Ricky Gervais	1633	4972	79	62.936709
8	John Mulaney	1391	4001	62	64.532258
5	Hasan Minhaj	1559	4777	73	65.438356
2	Bill Burr	1633	5535	80	69.187500
7	Joe Rogan	1435	4579	63	72.682540

Table 1. Dataset

5. Tools and Technologies used

- Anaconda : Jupyter Notebook
- Python Programming language
  - Data science process
  - Libraries like numpy, pandas, scikit, matlab, pyplot etc.
  - Machine learning models/algorithms

Libraries and the packages used in bit

brief-

Numpy: NumPy is a Python module which allows users to interact with the arrays. Numpy evenhas capabilities for trying to deal with algebraic expressions but also linear advanced mathematics. NumPy is referred vas the Numerical python.

Why NumPy:

We have lists in Python those acts like the arrays, however they are pretty slow to process.

NumPy intends to deliver a 50-fold quicker array object than ordinary Python lists. NumPy's array object is called ndarray, and it has a lot of features.of helper functions to make working with it a breeze. In data research, when speed and resources are critical, arrays are widely employed.

Pandas: Pandas is one of the most popular and well-liked data science tools for wrangling and analyzing data in the Python computer language. In the real world, nowadays data is inherently messy. When it comes to cleaning, transforming, manipulating, and analyzing data, Pandas is a game changer. Pandas, basically, assist in the cleanup of the mess completely.

Matplotlib: Matplotlib is a Python graphical interface as well as diagrammatical plotting packagewhich is a statistical enhanced version NumPy which keeps running. As a result, it provides an open source alternative to MATLAB.

Seaborn is a Scripting language visualisation kit that's also premised on matlab. It has a high-level interface for creating visually appealing and instructive statistics visuals.

Plotly: Plotly allows the users to study and visualize the data by importing, copying and pasting,or streaming it. . Plotly allows you to save, share, and collaborate on Python scripts.

Datetime: The datetime module is used for the manipulation of the dates and times.

Sklearn (Computational tool) is by far the most functional and reliable pattern recognition open source Library. It makes advantage of a Python consistency interface to provide a collection of machine learning capabilities. And statistical modelling, such as classification, regression, clustering, and dimensionality reduction.

Keras: Basically in the predicting sales module we concentrated on the Long Short-term Memory (LSTM) approach, which is a prominent Deep Learning method. In order to implement LSTM in our project, we used Keras. Kera’s is a Google-developed high-level deep learning API for implementing neural networks. It is built in Python and is used to make neural network implementation simple. It also allows for the computation of numerous neural networks in the

backend. Tensorflow is one of the frameworks that Keras supports.

Tensorflow: It's a free artificial intelligence programmed that creates models using data flow graphs. It enables programmers to build large-scale neural networks with multiple layers. Some of the best uses of tensorflow are Classification, understanding, discovery, predicting and creating.

Chapter 4: Performance Analysis

1. Data Analysis

A) Data Exploration

Data exploration is very important topic in data science because by exploring we can understand the nature of data and its usefulness to the problem objective. It has many steps

l**) Reading the data file** - In python data can be read using the pandas in csv, excel and many more formats.

```
In [12]: for comedian, top_words in top_dict.items():
             print(comedian)
             print(', '.join([word for word, count in top_words[0:14]]))
             print('---')
```

```
ali
like, im, know, just, dont, shit, thats, youre, gonna, ok, lot, gotta, oh, wanna
---
anthony
im, like, know, dont, got, joke, thats, said, anthony, day, say, just, guys, people
---
bill
like, just, right, im, know, dont, gonna, got, fucking, yeah, shit, youre, thats, dude
---
bo
know, like, think, love, im, bo, just, stuff, repeat, dont, yeah, want, right, cos
---
dave
like, know, said, just, im, shit, people, didnt, ahah, dont, time, fuck, thats, fucking
---
hasan
like, im, know, dont, dad, youre, just, going, thats, want, got, love, shes, hasan
---
jim
like, im, dont, right, fucking, just, went, know, youre, people, thats, day, oh, think
---
joe
```

Fig. 2 Reading Dataset

2) Variable identification - It is very important to identify which variables are independent andwhich are dependent and by how much and also the types (categorical or continuous). This waywe can only those variables which are more likely to change the outcomes.

Univariate analysis is used to explore the variables one at a time, summarize them and make sense out of the summary to gain insights or discover anomalies etc. Univariate

analysis of continuous variable can be done graphically by plotting histogram or by using describe() function in python. Analysis of categorical variables can be done graphically by plotting bargraphs or using counts() function to create a frequency table.

```
In [13]: from collections import Counter

         words = []
         for comedian in data.columns:
             top = [word for (word, count) in top_dict[comedian]]
             for t in top:
                 words.append(t)

         words
```

```
Out[13]: ['like',
          'im',
          'know',
          'just',
          'dont',
          'shit',
          'thats',
          'youre',
          'gonna',
          'ok',
          'lot',
          'gotta',
          'oh',
          'wanna',
          'husband',
          'got',
          'time',
          'right',
```

Fig.3 Top words from Dataset

3) Bi-variate analysis is used to determine independence and also to detect anomalies. Bi-variate analysis of continuous-continuous variables can be done graphically by scatter plot or by using corr() function in python to find correlation between those variables. Analysis of continuous categorical variables can be done graphically by bar plot or by using two sample t-

test between those variables. Analysis of categorical-categorical variables can be done two waytable or by using chi-squared test for those variables.

3) Treating missing values - There can be missing values in the dataset therefore we need to treat them accordingly. There can be many reasons such as no response from 12 filling party, error in data collection (faulty equipment) and error in reading. We use isnull() function in python to find both continuous and categorical missing values. WE can deal with them either by deleting or by replacement. We can delete the entire row or entire column. This leads to loss of data therefore it is not recommended. While replacing the missing values we have option to replace them by zero, mean of the column, mode of the column, median of the column or by creating a regression model and find value in accordance with the model. We can also use classification model to replace missing categorical values.

5) Outlier Treatment - There can be some outliers to our data which do not make any sense. We have to remove them in order to get good results. Reasons for outliers to appear in our dataset can be data entry errors, measurement errors, process error etc. Univariate outliers canbe detected with box plot and bi-variate outliers can be detected with scatter plot. There are several ways to treat outliers such as deleting, transforming and binning values, replacing outliers like missing values etc.

6) Data transformation - Data can be transformed by replacing values by some mathematical function. Data is transformed to get desirable scale of the variable, to convert non-linear relation to linear relation or to create symmetric distance from skewed distance. Many methodsare available to transform the data such as log arithmetic reduction, polynomial reduction, binning

```
In [18]: import matplotlib.pyplot as plt

plt.rcParams['figure.figsize'] = [16, 6]

full_names = ['Ali Wong', 'Anthony Jeselnik', 'Bill Burr', 'Bo Burnham', 'Dave Chappelle', 'Has
              'Jim Jefferies', 'Joe Rogan', 'John Mulaney', 'Louis C.K.', 'Mike Birbiglia', 'Ri

for index, comedian in enumerate(data.columns):
    wc.generate(data_clean.transcript[comedian])

    plt.subplot(3, 4, index+1)
    plt.imshow(wc, interpolation="bilinear")
    plt.axis("off")
    plt.title(full_names[index])

plt.show()
```

Fig. 4 WordCloud Library Use

2. Output at various stages

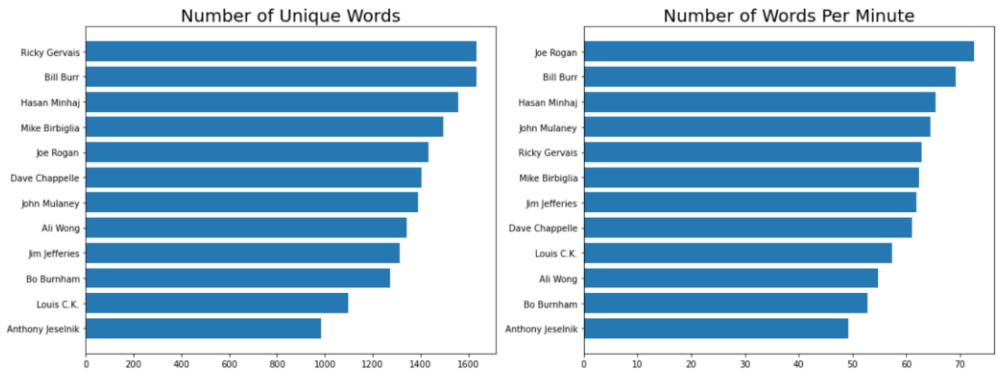


Fig. 5 New vs Existing stages

We have creating two different subplots and we have number of unique words and number of words per minute or the speed at which someone speaks and what we found is that in terms ofvocabulary this one the left Ricky Gervais and Bill burr use a lot of words Louie & Anthony have smaller vocabularies and then for chocolate these two talk quickly these two talk slowly and then Ali wong who is the person that I'm most interested in is somewhere in the middle sothis was some EDA that we have done.

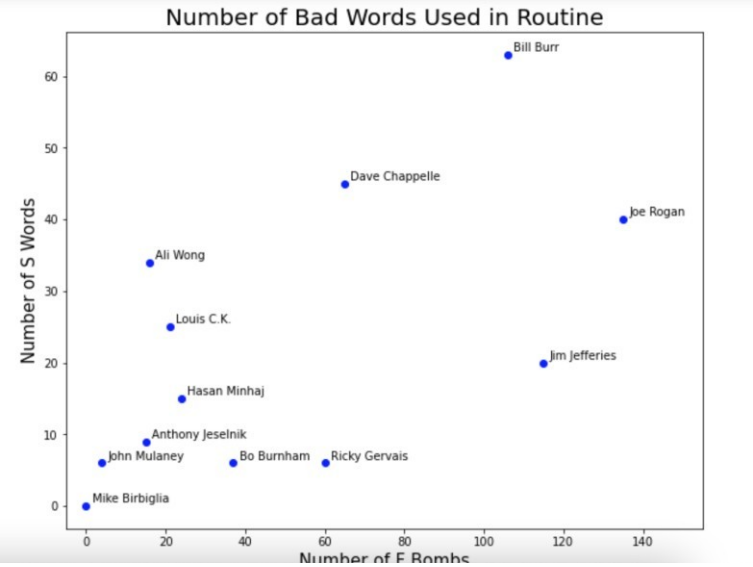


Fig. 6 No. of Abusive words

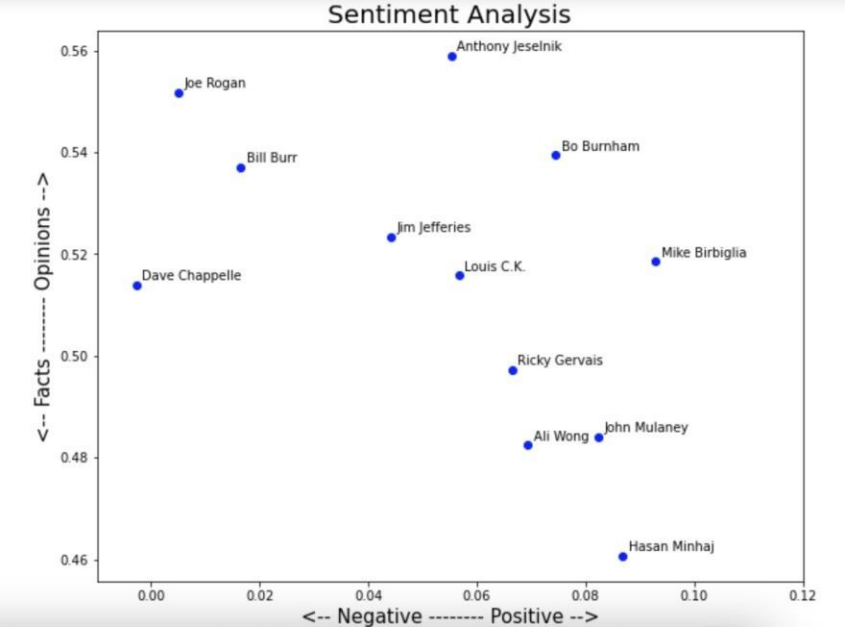


Fig. 7 Sentimental Analysis

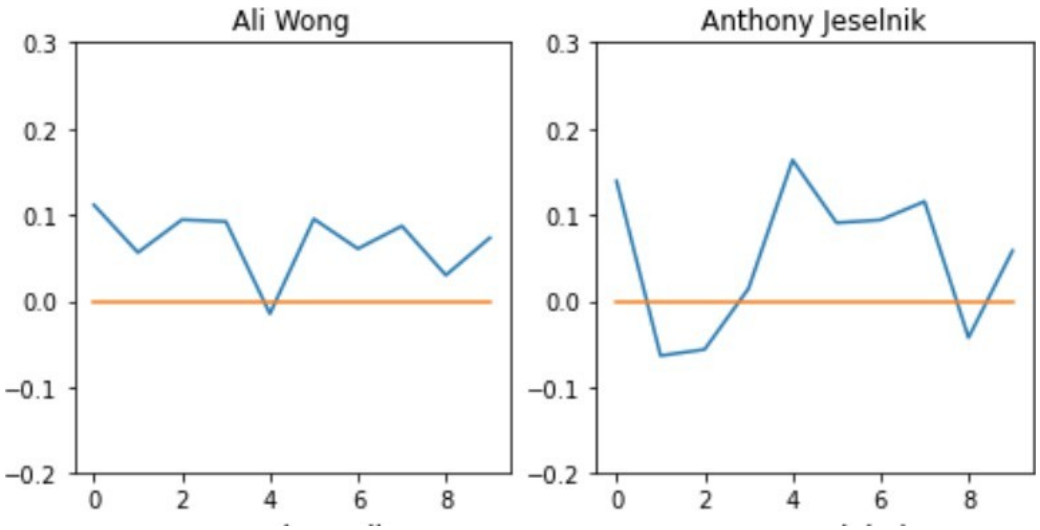


Fig. 8 Polarity Graph and Subjectivity

It shows the polarity behavior for different comedian

```
In [39]: lda = models.LdaModel(corpus=corpus, id2word=id2word, num_topics=2, passes=10)
         lda.print_topics()

Out[39]: [(0,
            '0.006*"cause" + 0.006*"went" + 0.005*"really" + 0.005*"thing" + 0.005*"fucking" + 0.005*"d
         ay" + 0.005*"goes" + 0.005*"good" + 0.005*"going" + 0.004*"say"'),
          (1,
            '0.008*"fucking" + 0.007*"shit" + 0.006*"fuck" + 0.006*"say" + 0.005*"want" + 0.005*"going"
         + 0.005*"didnt" + 0.005*"theyre" + 0.004*"hes" + 0.004*"did"')]
```

Fig. 9 Topic Modelling

Topic modelling input is a document term matrix. A document term matrix is a matrix where the rows are different documents and then the columns are different terms and the values in the matrix are the valve the word counts so what's going to happen is each topicwill consist of a set of words and in this case, order doesn't matter so we're going to workwith the bag of words

CHAPTER 5 – CONCLUSION

1. Conclusions

The dataset from Scrap from the loaf proved very helpful it gave us freedom to test and train ourmodel on real word data and transcripts. All the steps of cleaning data worked very well to filterdata and gather meaningful information from it. All the Machine learning and Natural Languageprocessing algorithms like clustering, sentiment analysis, text generation, topic modeling and EDA worked very well to obtain expected results. We can use this model to differentiate between positive words and negative words present in transcript, which will give users only the type of content, the want and require. It also provides very helpful to its user's underage who don't want to get exposed to negative content.

2. Future Work

We in our model we mainly focus on comedian Ali Wong for future work we can perform the same analysis on other comedians.
We can also collaborate with online audio streaming platform to automatically recommend next track.

3. Applications

We can use this model to differentiate between positive words and negative words present in transcript, which will give users only the type of content, the want and require. It also provides very helpful to its user's underage who don't want to get exposed to negative content.

REFERENCES

1. Stieglitz, S., & Dang-Xuan, L. (2013b). Social media and political communication: A social media analytics framework. Social Network Analysis and Mining, 3(4), 1277–1291.
2. Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2018). Advances in social media research: Past, present and future. Information Systems Frontiers, 20(3), 531–558.
3. Chung, W., & Zeng, D. (2016). Social-media-based public policy informatics: Sentimentand network analyses of US immigration and border security. Journal of the Association for Information Science and Technology, 67(7), 1588–1606.
4. Business Standard News (2017a) [Online], Available: http://www. business-standard.com/article/economy-policy/mrp-retailers-cansell-gst-inventory-with-new-price-stickers-till-30-sep117070500333_1.html.
5. Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2018). Polarization and acculturation in US election 2016 outcomes–can twitter analytics predict changes invoting preferences. Technological Forecasting and Social Change. https://doi.org/10.1016/j.techfore. 2018.09.009.
6. Pfeffermann, D., Eltinge, J. L., Brown, L. D., & Pfeffermann, D. (2015). Methodologicalissues and challenges in the production of official statistics: 24th annual Morris Hansen lecture. Journal of Survey Statistics and Methodology, 3(4), 425–483.
7. Briscoe, G. 2014. “Digital innovation: The hackathon phenomenon”.
8. M. Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," in IEEE Access, vol. 7, pp. 106111-106123, 2019, doi: 10.1109/ACCESS.2019.2930410.
9. Cardoso, J, Garcia Castro, LJ, et al. 2020. “Towards semantic representation of machineactionable Data Management Plans”. In: DaMaLOS – First Workshop on Data and Research Objects Management for Linked Open Science: Co-located at the InternationalSemantic Web Conference ISWC 2020. PUBLISSO. DOI: https://doi.org/10.4126/ FRL01-006423289

10. Cardoso, J, Jones, S, et al. 2020. *Mapping of maDMPs to Funder Templates*. Version 1.0.0. DOI: https://doi.org/10.5281/zenodo.3944458
11. Yi Wang, Qixin Chen, Chongqing Kang, Mingming Zhang, Ke Wang and Yun Zhao, "Load profiling and its application to demand response: A review," in Tsinghua Science and Technology, vol. 20, no. 2, pp. 117-129, April 2015, doi: 10.1109/TST.2015.7085625.
12. Garcia, L, et al. 2020. “Ten simple rules to run a successful BioHackathon”. In: *PLOSComputational Biology*, 16(5). Publisher: Public Library of Science, e1007808. ISSN:1553-7358. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007808 (visited on 05/08/2020). DOI: https://doi.org/10.1371/journal.pcbi.1007808
13. Hasan, A, Fouilloux, A and Jacquemot, C. 2020. *Integrating maDMP with the data management process*. Version 1.0.0. DOI: https://doi.org/10.5281/zenodo.3944434
14. Karimova, Y, et al. 2020. *Research Data Management Workflows and maDMPs*. Version 1.0.0. DOI: https://doi.org/10.5281/zenodo.3944468
15. Klar, J, et al. 2020. *maDMP export from RDMO*. Version 1.0.0. DOI: https://doi.org/10.5281/zenodo.3944448
16. Manghi, P, et al. 2019. *The OpenAIRE Research Graph Data Model*. Version 1.3. DOI: https://doi.org/10.5281/zenodo.2643199

APPENDIX

In [44]:
```
data = {}
for i, c in enumerate(comedians):
    with open("transcripts/" + c + ".txt", "rb") as file:
        data[c] = pickle.load(file)
```

In [45]:
```
data.keys()
```

Out[45]:
```
dict_keys(['louis', 'dave', 'ricky', 'bo', 'bill', 'jim', 'john', 'hasan', 'ali', 'anthony',
'mike', 'joe'])
```

In [50]:
```
data_combined = {key: [combine_text(value)] for (key, value) in data.items()}
```

In [51]:
```
import pandas as pd
pd.set_option('max_colwidth',150)

data_df = pd.DataFrame.from_dict(data_combined).transpose()
data_df.columns = ['transcript']
data_df = data_df.sort_index()
data_df
```

Out[51]:

	transcript
ali	Ladies and gentlemen, please welcome to the stage: Ali Wong! Hi. Hello! Welcome! Thank you! Thank you for coming. Hello! Hello. We are gonna have ...
anthony	Thank you. Thank you. Thank you, San Francisco. Thank you so much. So good to be here. People were surprised when I told 'em I was gonna tape my s...
bill	[cheers and applause] All right, thank you! Thank you very much! Thank you. Thank you. Thank you. How are you? What's going on? Thank you. It's a ...

In [60]:
```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(stop_words='english')
data_cv = cv.fit_transform(data_clean.transcript)
data_dtm = pd.DataFrame(data_cv.toarray(), columns=cv.get_feature_names())
data_dtm.index = data_clean.index
data_dtm
```

Out[60]:

	aaaaah	aaaaahhhhhhh	aaaaauuugghhhhhh	aaaahhhhh	aaah	aah	abc	abcs	ability	abject	...	zee	zen	zeppelin
ali	0	0	0	0	0	0	1	0	0	0	...	0	0	0
anthony	0	0	0	0	0	0	0	0	0	0	...	0	0	0
bill	1	0	0	0	0	0	0	1	0	0	...	0	0	0
bo	0	1	1	1	0	0	0	0	1	0	...	0	0	0
dave	0	0	0	0	1	0	0	0	0	0	...	0	0	0
hasan	0	0	0	0	0	0	0	0	0	0	...	2	1	0
jim	0	0	0	0	0	0	0	0	0	0	...	0	0	0
joe	0	0	0	0	0	0	0	0	0	0	...	0	0	0
john	0	0	0	0	0	0	0	0	0	0	...	0	0	0
louis	0	0	0	0	0	3	0	0	0	0	...	0	0	0

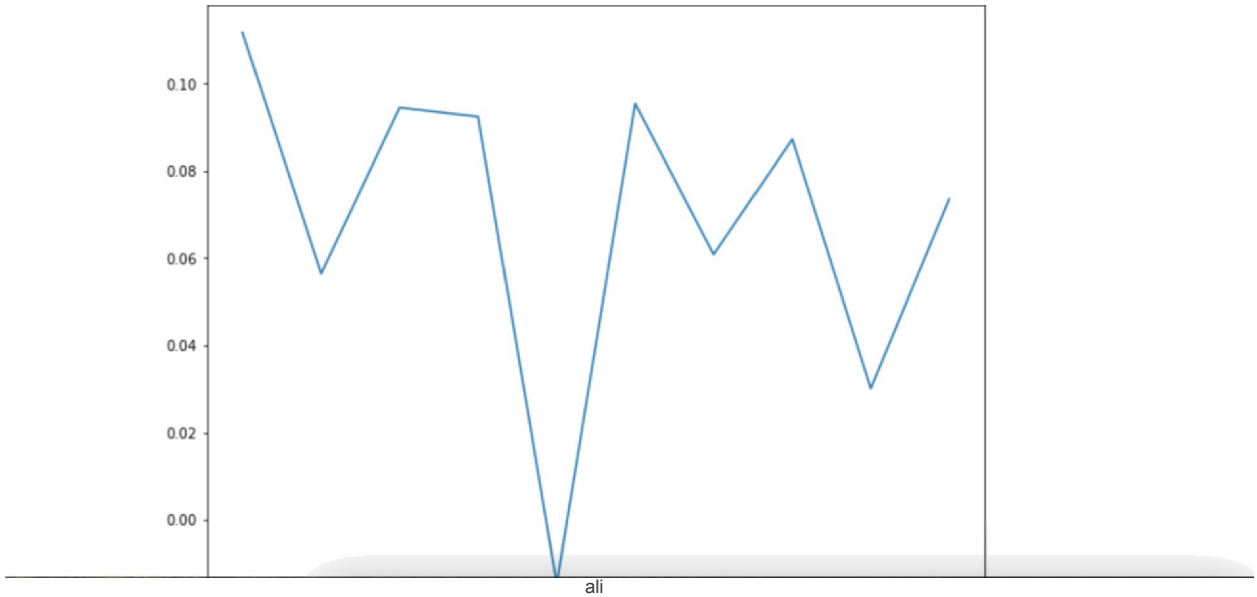
```
data_words['total_words'J = total_list
data_words['run times'] = run_times
data words['words per minute'] = data words['total words'] / datawords['run times']

data_wpm_sort =
data_words.sort_values(by='words_per_minute')data_wpm_sort
```

Out[20]
:

	comedian	unique words	total words	run times	words	er minute
1	Anthony Jeselnik	984	2905	59		49.237288
3	Bo Burnham	1272	3165	B0		52.750000
0	Ali Wong	1341	3283	60		54.716667
9	LouisG.K.	1098	3332	58		57.448276
4	Dave Chappelle	1404	4094	67		61.104478
6	Jim JeWeries	1313	4764	77		61.870130
10	Mike Birbiglia	1494	4741	F6		62.3B1579
11	Ricky Gervais	1633	4972	79		62.936709
B	John Mulaney	1391	4001	62		64.532258
5	Hasan MinhaJ	1559	4777	73		65.43B356
2	Bill Burr	1633	5535	80		69.187500
7	Joe Roaan	1435	4579	63		72.682540

In 10]:
```
pit.plot(polarity transcript[D])
pit.title(data['full_name'].index[0])
pit.show()
```



ali

```
generate_sentence(a1i_dict)
```

In tl2]:

In [11]:
```
import random

def generate_sentence(ohain, oount=15):
      'Input a dictionary in the format of key = current word, value = list of next words
       along with the number of words you would like to see in your generated sentence.'''

    wordl = random.choice(list(chain.keys()))
    sentence = wordl.capitalize()

    for i in range(count-l):
        word2 = random.choice(chain[wordl])
        wordl = word2
        sentence +'        + word2

    sentence =
    returo(sentence)
```