An Investigation on the Impact of TikTok on Fame

Nikodem Adamski
C18415776
Dr.Jenny Munnelly
COLLEGE OF BUSINESS, TU DUBLIN, CITY CAMPUS

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
B.Sc. in Business Analytics

2022

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of BSc Business Analytics, is entirely my own work and has not been submitted in whole or in part for assessment for any academic purpose other than in partial fulfilment for that stated above. The work in this project conforms to the principles and requirements of Technological University Dublin guidelines for ethics in research. Where required to draw from the work of others, published or unpublished, the Author has acknowledged such work in accordance with established scholarly and editorial principles.
Name: Nikodem Adamski
TUD Student Number: C18415776
Department: College of Business, School of Management
Signed:

Date: 24th May 2022

Acknowledgements

First of all, I would like to give my heartfelt thanks to my academic supervisor Dr. Jenny Munnelly, for her invaluable instruction, inspiration. Without her previous advice and guidance, this study could not have been completed. Big thanks go out to Dr.Wael Rashwan who helped me tremendously with any queries I had about the project. Also, I must express my sincere thanks to all the professors in Technological University Dublin, for their enlightening courses and lectures in these four years of study.

My acknowledgements also go to my classmates. Their support and generosity move me a lot. Last but not least, I have to thank my family members especially my parents.

Abstract

This project aims to analyze the significance of TikTok using web/scraping data mining a dataset. Use supervised machine learning models such as LSTM and ARIMA/SARIMA to determine that "TikTok" is the platform to Achieve Greatness, This resulted in TikTok users having an edge over none TikTok users. Furthermore, each application scraped was correlated with one another to identify any relationship—the results of this research present evidence of a significant relationship/ correlation between social media applications. The hypothesized defense was done using the Spearman and correlation function.

1. Introduction

TikTok is a Social Media Platform Founded in September 2016, Bought out and renamed "Musical.ly" by a Chinese company named DouYin.[1] TikTok as an application primarily attracts a younger audience to engage with content viewership, content creation and content commentation on the videos found on the application. A TikTok video ranges between 1 second to 3 minutes, with the average length being 30 seconds per video. This fast paste entertainment has led to the application's success since its launch, as Active users stand at 1 billion. TikTok also owns a sister application in China, Separated from Western TikTok, with a separate 600 million Asian users. TikTok is the west currently lands 7th as the quickest growing company to tackle the giants such as Instagram, Facebook, and Twitter. (Battersby, n.d.)

Web.2 is a common word used within the found pieces of literature in the Literature review section. What does it mean?

Web 2 is a term that was founded in 2004 to describe a new way in which software developers and end-users utilized/used the world wide web. This information means that a singular entity does not just upload online content; instead, it is made in collaboration/modification between multiple entities worldwide. While web 1 pages still exist on the world wide web, including personal pages such as web pages and encyclopedias, web 2 pages exist cohesively, such as blogs, vlogs, wikis and other collaborative projects. Although the worldwide web did not receive any specif update, it does require specific extensions for the full functionality of web 2. These Web 2 extensions include Adobe Flash (Common and popular for additional animation, activity and audio/video streams to a webpage).RSS(used to update Blog entries and news headlines in standard formatting.). Ajax(Asynchronous JavaScript, which allows web content to be updated without interfering with the display or behaviour of the entire page).[2] Web 3, on the other hand, is the third generation of the internet. While still not updating the entire world wide web. It focuses on using a machine-based understanding of data to provide a more data-driven and Semantic Web.[4]

Charlie D’Amelio Is a TikTok star. She began her TikTok journey back in 2018 And grew her account from 0 followers to over 141 Million followers. She was the first 16-year-old ever to gain over 100 million followers. Charlie's Net worth is approximated to be around 5-8 million euros as of 2022 through social media Sponsorships charging an estimate of 100 thousand per video, Which, as said previously, ranges from 15 - 30 seconds. Charlie does nott come from family of wealth but is currently raising her family as millionaires.This young star portrays the application's Power; moreover, she is not an anomaly. Bella Porch is a perfect example of a TikTok Star who also grew her account this time from 0 followers back in January of 2020 to almost 90 million followers today. Only eight months after her initial start on TikTok, Bella has gained the most viral video on the platform. This sudden burst of followers allowed the young artist to boost her way into the music industry and win her first-ever Song, "Build a b****" also Blowing up on applications such as Youtube and Spotify, gaining much traction, and bringing much newfound wealth to her portfolio. This would not be possible if she had not started a TikTok account back in January of 2020.

The evidence of TikTok bringing celebrities to a new height shows a certain amount of pushing Power, allowing an opportunity for anyone to become viral but “what about everyday people?”. This can be seen with even the author("Nick") of this research gaining over "28 million views" in the accumulated video views that contain multiple videos over “2 million views” and a single Viral video gaining over “18 Million views”. Even though celebrities every day gain unbelievable numbers, everyday people are dircouraged from the beginning. Reality shows that it is the completely the opposite. Examples worldwide show that anyone can achieve greatness, "Is TikTok the platform to achieve greatness?".

The research question in this work is "what is the impact of the use of social media platform TikTok on fame." and "Does TikTok hold a significant impact to the growth of a user's popularity outside of the application."

This research introduces 'TikTok' and "Fame" as it lacks literature online, and with the Recent surge of Development of web two and now web three, we are likely to switch online to a new era.

This project idea originated from What was observed in day-to-day life on 'TikTok' while scrolling. It was noticeable that the ("for you") page "home page" became packed with verified firm accounts. As Corporate giants entered TikTok and started to post videos, an idea popped up, 'is it a coincidence?' if so, 'why is everyone doing it ?'.

Moreover, four datasets from all four significant applications need to be recorded, including TikTok, Instagram, Twitter, and Youtube data. Although Facebook was excluded, it is not deemed irrelevant or more negligible of an application than the rest. Data preprocessing, Analysis and prediction will be made through Python.

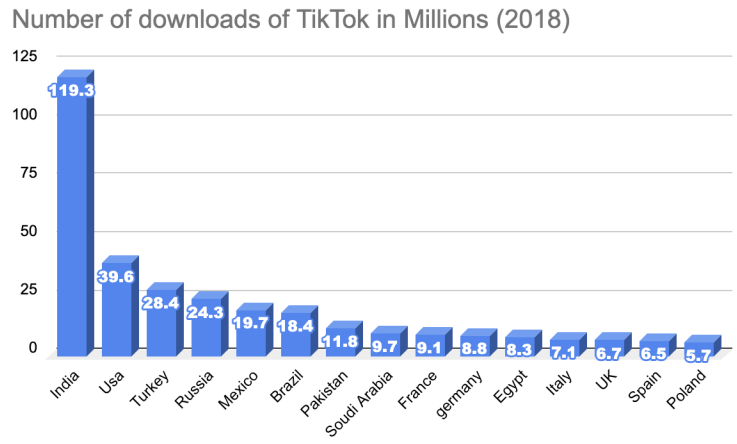
The Models used for predictions are LSTM, Arima and Sarima. However, many more methods, including unsupervised learning models and other supervised models, were carefully considered when deciding on the models used within the investigation.

To obtain the data sets, OctoParse was used. It is a piece of modern web information gathering application. Octoparse is a code-free software that allows to quickly scrape data from the internet in the format of the user choice. Furthermore, Google Trends, including Excel, were used to develop a TikTokers vs None TikTok users dataset.
This research dictates the significance of "TikTok" concerning fame, Provides objective evidence of significance in the Relationship between TikTok users and TikTok upload information, the Relationship between social media applications, and the Relationship between application and the original TikTok user growth to determine the significance of each.
Moreover, this Research analyses the potential mistakes and improvements made while working on such a broad topic and encourages future researchers to continue on the mantle of finding the truth behind the hypothesis.

2. Related Works

Social media is a commonly recurring topic in Literature, specifically its " trend analysis" and "prediction of time series data." Due to the subjective nature of "Fame," it is far more challenging to calculate the root cause, which is why it is less written about online until today, due to the nuance of TikTok as a whole. The earliest article on the topic was an article from 2018 written in English. Prior to this, TikTok mainly was spoken about by Indian researchers who are high due to the increased popularity of the application in that region[fig.1.],

Most likely due to the nuance of the application. While most kinds of literature focus on social media platforms such as Twitter, Facebook, and Instagram, the decision of TikTok was due to the lack of relevant work done on the application. Choosing a chronological order approach shows the creative power over time and how perception is constantly evolving. In this Literature review, pieces of Literature from 2010 until 2021 were discussed. These works of Literature were all reviewed to gain a good understanding of Trend analysis. Moreover, to dictate if there was any relevancy that could be duplicated, any gaps that could be identified were pursued through this research. In the end, it was found that the research online lacked data or information on "TikTok" concerning "fame," so that is the most significant gap that the research will accomplish. Another gap further found was the lack of Arima usage in trend analysis online, which is a missed opportunity covered by the current research and proved to be an accurate and reliable method of prediction. In the Literature, there was also evidence of the strength of sentiment analysis and evidence of usage of unsupervised learning as a whole. This information was taken into account when choosing the correct approach for the investigation.



[fig.1.] Recreation of a Diagram created on the number of Tiktok uploads in 2018

2.1 Literature/s in 2010

In 2010, very few studies explored social media as a valuable resource of media attention, But with the publication of 'Users of the world, unite! The challenges and opportunities of Social Media' (M. & M., 2010) by the following authors; A. M. Kaplan and M. Haenlein. The Literature writes, "***It is reasonable to say that Social Media represents a revolutionary new trend that should be of interest to companies operating in online space or any space, for that matter***" [1], indicating the potential importance to businesses. However, Through the Literature, scholars and businesses frowned upon the idea of social media and "free speech" due to misconceptions. "***There seems to be confusion among managers and academic researchers alike as to what exactly should be included under this term, and how Social Media differs from the seemingly-interchangeable related concepts of Web 2.0 and User Generated Content.***".[1]
These points acknowledge the potential of what was yet a new topic back in 2009. As the points made by the author mainly were composed of theoretical evaluation, the Study lacks statistical backing to portray the true significant importance of social media. It fails to provide multiple sources of social media presence. Luckily this is due to the lack of social media diversity in 2009 rather than the Literature. The Article helped this project to understand which data points are valuable and judge their evaluation based on if statements when finding the appropriate correlation to "fame".

Two researchers named Michael Mathioudakis and Nick Koudan presented a study by 'TwitterMonitor: Trend Detection over the Twitter Stream'. (Mathioudakis et al., 2010) This Study took a more statistical approach to trend analysis and focused on Twitter. User interaction on the social media platform(Twitter), also known as tweets, was researched through the

Literature. More specifically, they discussed "***establishing a system that identifies emerging topics (i.e. 'trends') on Twitter in real-time and provides meaningful analytic that synthesises an accurate description of each topic***".
Moreover, wrote "***popular interest in a particular topic and are often driven by emerging news or events. For example, a sudden rise in the frequency of the keyword 'NBA' may be linked to an important NBA match taking place***" [2]. This quote is a great insight that shapes the idea of popularity. The popularity of videos on youtube today can gain millions of views years after their original upload date due to a specific event happening in the world. The Literature highly reflects fame to be 'Time sensitive'. This study helped to understand further what drives popularity, which changed the research perspective on collected data mining features.
This Literature was one of the multiple primary papers with the highest correlation to own that were time series analysis oriented. Moreover, it provided great insight into trend analysis. However, it cannot compare to other applications and does not represent fame as the main topic as it was not the aim of the Literature.

Another study was named 'Predicting the future With social media' (S. & A., 2010) by I. P. Cvijikj and F. Michahelles. Multiple quotes in the Article were intriguing and helped form an appropriate topic analysis. A quote that helped to achieve this was, " ***Trend monitoring over Twitter stream has already been the subject of attention of scholars and professionals, resulting in numerous modified and new algorithms for information retrieval and commercial online tools. However, to the best of our knowledge, at the moment, there are no commercial or research efforts related to trend detection on Facebook.***" The Article indicates that old data/ ways of manipulating data improve over time. When conducting this research, the aim was to construct data capable of being reproduced with another application. Application interchangeability allows future generations to conduct the subsequent research on the next extensive application. This Literature exposed the possible predictive power social media has when predicting future values.
Moreover, providing evidence of using sentiment analysis of tweets to improve that prediction further. The gap in the Study was the applications used. Facebook and Twitter are Competitive giants, but new upcoming applications should not be ruled out of the equation as they might hold significance, which is the primary motivation of the project at hand.

2.2 Literature/s in 2012

Literature from 2012 by the name 'Mining social media: a brief introduction' (Gundecha et al., 2012) by P. Gundecha and H. Liu. This article Demonstrated the ability to mine social media information.

Furthermore, a description of social media data as a whole, stating, "***Social media data are vast, noisy, distributed, unstructured, and dynamic, which poses novel challenges for data mining***", was fully understood when scarping and understanding the dataset. This article was the guidance in understanding the project's longevity. It can be interrupted by other applications, which were one of the main underlying goals in this project, Just as data in the past on platforms such as Twitter / Facebook. Now in this Study, TikTok is presented. The best approach to overcome the lack of flexibility found in may Literature is to create the ability to obtain flexible data or rather a flexible solution for overcoming obstacles when the data source changes. In this Literature, several gaps can be included, such as the flexibility of the project, and the usage of twitterTracker disables the company to use their other application scarped data for analysis appropriately.

2.3 Literature/s in 2016

A study named "Stock Price Prediction based on Stock Big Data and Pattern Graph Analysis" (Jeon et al., 2016) by four researchers named 'Seungwoo Jeon, Bonghee Hong, Juhyeong Kim and Hyun-jik Lee'.
The Study primarily predicted stock prices based on the closing price and historical data. The Study showed an in-depth knowledge of patterns that can be found based on previous dates, which was the inspiration for careful data analysis in search of potential patterns when predicting the research of TikTok and fame. The only gap is the methods used for prediction. They did not use Arima and questioned its accuracy within the short term prediction, Sarima or Lstm, which are the models used in this research. In the Study, they said, "***Moreover, even though these studies have analyzed the significance of variables and increased the prediction accuracy by eliminating unimportant variables, the error rates of the predictions are high owing to the use of outliers***." This remarkable insight provided the idea of outliers to be the main focus of the project preprocessing, which included appropriate grouping and dealing with null values/deviations in the data. This research is an excellent example of constant improvement. There is always potential to improve a prediction based on a model or even think from an outside the box solution, such as applying unsupervised learning models in data to identify relationships further.

2.4 Literature/s in 2018

The first-ever Literature in English found on google scholar was in 2018. An article created by Hou, L. named 'Study on the perceived popularity of TikTok' (Hou, 2018) addressed the idea of interest around the application through quantitative data. The survey results in the article provide

insight that the application was only at its earliest stage of customer reach. The paper's conclusion found a significant correlation between the younger generation, 18-28. Secondly, it spoke about the benefit of product placement and marketing. What correlates with personal research the most was, '***The third part is the finding of the relationship between independent variables (product positioning, content variety, uniqueness) and dependent variable (perceived popularity of Tik Tok). Through the analysis of Pearson correlation, product positioning and content variety have a significant impact respectively on the perceived popularity of Tik Tok while uniqueness was not the contribution to prediction. In a word the hypothesis is not completely stand. Product positioning and content variety were proved to significantly contribute to the perceived popularity of Tik Tok, but the uniqueness is proved not the contribution to prediction***'. Which provides evidence that successful businesses or upcoming businesses who are looking for extensive reach could find that there might be a correlation between product placement and popularity within TikTok, and popularity on a single application adds up to the total overall growth of a company.
This article provided insight into the reality of the perception of the application at its earliest stage. This information proves to be a highly correlative topic to the one challenged, but this time with quantitative data rather than qualitative.

2.5 Literature/s in 2019

An article named 'Market Trend Prediction using Sentiment Analysis: Lessons Learned and Paths Forward' (Mudinas et al., 2019), made by the following authors,' Andrius Mudinas, Dell Zhang and Mark Levene'. This article approached the idea of using sentiment analysis instead of the previously used method.' ***Nowadays, many investment banks and hedge funds are trying to exploit investors' sentiments to help make better predictions about the financial market'***. The quote shows how businesses take every approach to have the edge over the competition, including jumping into unresearched deep water to hopefully turn over a more significant profit. Furthermore, This highly relates to personal research. It displays the lengths people/ businesses will go through to find the highest percentage of success, including using a new social media application to have an edge against their competitors.
A significant gap found in the study was in trend prediction, Which is only due to the nature of sentiment, which can only be calculated based on a boolean and frequency, which is the drawback of sentiment analysis. A mix of sentiment and regression would solve the issue.

Another piece of literature found in 2019 related to the project was named 'Machine-Learning Models for Sales Time Series Forecasting' (Pavlyshenko, 2019).' ***At present time, different time series models have been developed, for example, by Holt-Winters, ARIMA, SARIMA, SARIMAX, GARCH, etc.***' This article Provided in-depth knowledge on different models used with time series, Which helped develop knowledge of appropriate models used in personal

research and predictions. This article only has one flaw/gap: it does not specify how different data features might differ from different models used. Although a model might accurately predict future value, the gap lies in what can be used with predicted future value/s and how can the prediction be used to suit research.

An article named 'A comparison between Arima, LSTM, and GRU for time series forecasting.', (Yamak et al., 2019) created by 'Yamak, P. T., Yujian, L., & Gadosey, P. K.' Provided excellent insight into the differences between different time series models. This study was compelling to focus on Arima as the primary source of prediction as ***'it shows that the ARIMA Model gives the best accuracy and time***.' but also encouraged to attempt Lstm as it provided much in-depth knowledge on the topic.

2.6 Literature/s in 2020

In 2020, an article named ‘Popularity Prediction of Instagram Posts’ (Carta et al., 2020) was created by ‘Salvatore Carta, Alessandro Sebastian Podda, Diego Reforgiato Recupero, Roberto Saia and Giovanni Usai ’. This study focused on application popularity based on user interaction ‘Posts’. Future potential growth is predicted by using the x amount of likes and x amount of posts with a user data history. ***‘ In addition, finally, the development of new tools that, using as input the results of our approach, can potentially help users and social media managers to optimise their contents to achieve a good level of expected popularity for the new posts to be published. ’*** Through this article, the idea of your work being multifunction is an excellent factor. It allows for the idea of researched work having multiple meanings and different viewpoints. For example, although the research at hand is on predicting future tiktokers popularity value, it can also encourage none TikTok users, businesses who wish to choose it as an opportunity to grow and develop but also brands dictating the perceived power of a particular application to dictate the percentage of income that can be delegated to a specific application to appropriately and most efficiently grow online.

[Fig.2.]

Another valuable article was named ‘A Time Series Analysis of Trends With Twitter Hashtags Using LSTM ’ (Shams et al., 2020), created by ‘Monjur Bin Shams, Md. Junaed Hossain and Sheak Rashid Haider Noor’. This article provides valuable insight into the LSTM prediction model used with application features. This article incentivised the research to go down the rabbit

hole of finding new and valuable resources used in relevant work studies. The studies on the topic are new and have a relevantly low number of citations ranging from 0 to 10.

This literature review has applied excellent knowledge to this research allowing for the most accurate predictions.
Moreover, methodology and approaches were chosen due to the performance viewed in other literature. Multiple ways of completing the prediction in pieces of literature based on supervised or unsupervised learning and the different models that could be applied in personal research were tested in Trial research. Moreover, insight into the application features provided and how it was perceived in its earlier days allows us to understand better which features are at the highest value when feature ranking. This research also provided evidence of the importance of data preprocessing and how dealing with outliers earlier on could yield higher overall performance.

2.7 Literature/s in 2021

In 2021,' On the Psychology of TikTok Use: A First Glimpse From Empirical Findings' (C et al., 2021)was created by 'Christian Montag, Haibo Yang and Jon D. Elhai'. This literature provided knowledgeable insight into the history of TikTok as an application' ***TikTok (in Chinese: DouYin; formerly known as musical.ly) currently represents one of the most successful Chinese social media applications in the world. Since its founding in September 2016, TikTok has seen widespread distribution, in particular, attracting young users to engage in viewing, creating, and commenting on "LipSync-Videos" on the app.***' Nevertheless, Due to this literature focusing on the psychology of TikTok use, it lacks relevance to this current research.

2.8 Literature Review table

Year	Title	Citation	Keyword	Methods	Gaps
2010	TwitterMonitor: Trend Detection over the Twitter Stream	1164	Trend Detection	Data Mining, Twitter Monitor	Lack of Application Comparison
	Users of the world, unite! The challenges and opportunities of Social Media	26527	Social Media	Theoretical	Lack of Statistical/Coding backing
	Predicting the future With social media	3171		Linear Regression Visualisation	Applications used
2012	Mining social media: a brief introduction	278		Data mining, Tweet Tracker	Flexibility of the project
2016	Stock Price Prediction based on Stock Big Data and Pattern Graph Analysis	19	Prediction	(R, Hadoop) RSME Supervised / unsupervised learning	Limited Prediction models
2018	Study on the perceived popularity of TikTok	14	TikTok	Survey(qualitative data)	Statistical historical backing
2019	Market Trend Prediction using Sentiment Analysis: Lessons Learned and Paths Forward	29	Trend Prediction	LSTM SVM	Limited to sentiment
	Machine-Learning Models for Sales Time Series Forecasting	104	Machine learning	ARIMA RandomForest Lasso Neural Network Stacking	No information on what to do with data after prediction.
	A comparison between Arima, LSTM, and GRU for time series forecasting	71	time series forecasting	LSTM Arima GRU	Lack of information On Sarima
2020	Popularity Prediction of Instagram Posts	21	Instagram	XGboost Random Forrest	Flexibility Outdated Data mining
	A Time Series Analysis of Trends With Twitter Hashtags Using LSTM	3	LSTM	LSTM	Limited only to in-app activity.
2021	On the Psychology of TikTok Use: A First Glimpse From Empirical Findings	31	TikTok	Theoretical	No statistical/Coding Backing

2.9 Gap Explanation

These are the gaps found in all the pieces of literature. Some of them have already been proven In other works of literature, such as No statistical backing / Coding, lack of Sarima, Outdated Data mining, Flexibility of the project and lack of statistical backing. This research is here to provide relevance to the gaps that were not covered. These include correlating “tiktok” to fame, providing a Trend analysis on “TikTok” using Arima Model, Using “Bidirectional LSTM” for trend analysis and “Using GRU” concerning predicting “TikTok user” data.

3. Methodology

The Steps taken to partake in the methodology have been supplemented from "research Methadology in Business: A starte's Guide". (A & A, 2018) Where the Research onion was broken down in the following steps beginning with a Pragmatic approach, continued with Induction, followed by Multi-method Quantitative data sets that were composed of a Purposive and convenience sample of the top 50 celebrities of 2021, and the following data scraped was from those top 50 celebrities data. The Applications used to scrape the data were Google Trends and Octo Prase, and the Data preprocessing was performed on the data as Data Cleaning, Data Transformation, and data analysis. In the end, 5 data sets were then combined for further analysis resulting in a singular data set to compare the relationship between features.

3.1 Research Philosophy

Interpretivism argues that, unlike natural Phenomena, social phenomena are unique, making them too complex to be reduced to generalization rules and formulae. This approach is highly correlated with the objective view of "Fame." Fame by itself is none quantifiable, highly unpredictable, and spontaneous. Nevertheless, in the study, the research objectives attempt to find "some" correlation to understand the social phenomenon better while not necessarily the absolute truth.

Positivist adopts a scientific stance and aims to find general findings. Not all positivist queries can be justified to find absolute truth in the project objectives. However, some are, such as the "true" values found within the aim to "determine from the sample the correlation that applications have on user growth." this is a query that can be Positivistly justified yet still falls under the Interpretivism Tree, which brings me to Pragmatism.

Pragmatism is a research philosophy that focuses on the practical outcome of the research. Furthermore focuses on using "What works" methods while arguing that both methods can be used to achieve research objectives.

The research objectives found in this research are: "identify the growth of TT users and none TT user to determine the benefit of one over another,"

"Identifying the correlation between Application(TikTok) and User growth to see if it impacts fame in the concise term" and "Identifying short term impact of applications concerning fame." These objectives all contain the ability to satisfy the Pragmatism query. Moreover, the results found from the findings satisfy the Interpretivism contrary stance that the figures found do not fully represent the social phenomenon and are not the absolute truth but only the truth in this specific sample.
Pragmatism is the result of "Fame" being the main topic of the research, "Fame" cannot be justified in a single variable as it is a collaborative combination of plethra of variables that only in combination fully represent "fame."

3.2 research Approach

Induction, also known as a "Bottoms up " approach, is less concerned with generalization; instead, it better understands the research phenomenon. It is furthermore adopting a more flexible structure for investigation. This approach is the research approach undertaken. Data was collected, patterns were identified, a hypothesis was made on the results found, and a theory was formulated.
The deduction approach is known as the "top-down" approach, where, contrary to induction, it reverses how the research is complete. This approach can be seen in a glimpse during the research undertaken. Through analyzing the correlation of features ('Applications") have on the "Users," This is where a theory was made "Application holds a high significance" Then it was deducted through analysis and finally confirmed the analysis through observation. Although induction was the main focus, Both induction and Deduction were used to analyze the dataset collected firmly.

3.3 Research Methods

The main focus was Accuracy, Unbiased, Insightful, and Complete Report when deciding on the research methods. The first step to achieving this goal was to decide whether the Data type was collected Qualitative research or Quantitative research.
Qualitative research Depends on words rather than numbers, meaning that the research findings are not produced through quantification. Its primary focus is the exploration of data collected through means such as the unsupervised learning method to identify relations between features and theorize their impacts.
Quantitative methods investigate Phenomena by collecting Quantifiable data, usually numerical, and applying mathematical models and statistical techniques for data analysis. This method can be seen in supervised machine learning models such as linear regression, Arima, and any other model of that learning type. The result of the quantitative method is to gain generalized findings in the form of theories and formulae.

Before concluding which method to choose, it is essential to identify both methods' strengths and weaknesses to determine the correct approach in this investigation.
Qualitative data strengths lie in the description of complex phenomena, which would help understand "fame" through, for example, user insight such as comments on videos and questioners. Another strength lies in user interpretations and idiographic caution, providing evidence of multi-layered features associated with "fame" when analyzing "fame" more in-depth with quantitative data. The weaknesses in qualitative methods are that results may not be generalized, and it is mainly tricky to test hypotheses and test. At the same time, also qualitative studies may come to have a higher weight of bias to the researcher's influence.
Quantitative methods' strength lies in the testing and validation of the testing. As if numerical data were presented, it is easily programmable to identify the possible correlation and prediction based on previous values. Generalization could be made based on a random sample reducing bias, it is less time-consuming to collect this means of data, and it is commonly more used in larger data sets. At the same time, its weaknesses lie in the overall abstract, as it is too abstract and general for the direct application to specific.
The decision to which data type was complex as both methods have been found in literature online on 'Trend analysis,' which is the topic researched. The decision was made to create a quantitative data set. Which includes numerical data and provides the gap in this provide for future work. In potential future work, it is possible to look into the qualitative aspect of the project and find more concrete evidence of variables that hold weight to "FAME".
Furthermore, future work could implement the Mix method design through quantitative and qualitative datasets to analyze this project from a theoretical and statistical standpoint to determine a more validated result on TikTok trend analysis. This method is superior due to five reasons. Triangulation means increasing validity findings; Complementary meaning improves the interpretation of the data and gains more meaningfulness from the data in the results; development means utilization of the results to enhance validity. Initiation means analyzing from different perspectives, and finally, Expansion means possibilities of using different methods at different stages of inquiry. All previously mentioned benefit studies would, in the end, validate today's investigation.

3.4 Research Tool and Techniques

The sampling method used for this specific data scraping/Data gathering was None ProbabilisticTechnique, Which means that no probability was considered when sampling. To be more specific, a technique called Purposive meaning judgment was used. In the research, this technique is where a website was identified with the top 200 celebrities of 2021[1]. This data has a specific bias toward celebrities who are believed to hold the significance of "fame," so the decision is justifiable. Then, a sample of that sample was taken due to time constrain and convenience. The data collected a second time was also Purposive and convenient as the top 50 celebrities with the best personal belief that the data would represent the overall population well.

3.5 Data Collection

Data collection began with the search for an appropriate data set.

3.5.1 Kaggle

A step taken to search through famous was through data libraries. Kaggle was an appropriate start. Kaggle is an online community of data scientists and machine learning practitioners. Through the search on Kaggle, no dataset met the requirements that satisfied this study. The data that satisfied the study is quantitative data consisting of top x users on the platform and including likes/comments/names/subscribers. Unfortunately, no dataset on the platform met the requirement of any application. The keywords used in searching on Kaggle were (‘TikTok’,’ Instagram’,’ Youtube’,’ Twitter’,’ Facebook’,’ Fame’)

3.5.2 Scraping through Python

Python is a programming language used for many applications. One of the applications of python that it can be used for is web scraping. We have already developed a good understanding of how to correctly Scarpe data through classwork, but an issue arose. The computer's IP was being taken down from scraping, providing no data but only an error 404. This error occurs when a website suspects a device of scraping information and blocks a particular IP. Unfortunately, no reasonable solution through coding was found. Another error arose when scraping from applications such as Instagram required a log in every time their website was used, and the blocked IP after login was bypassed.

3.5.3 Parsehub

Parsehub is an application used for web scraping online documentation, perfect for the required task. An issue began when scarping the data.
Parse hub is primarily a paid service that requires much upfront cost to parse even a tiny amount of information. It would have to be one of the most expensive options to use the application for this specific job of scarping likes/comments. Sadly the quantity of 599$ per month was not within the budget, so the idea was scrapped.

3.5.4 Octoparse - Octopus Data Inc

Fortunately, a solution was indeed found. Octoparse is a Chinese data web scraping application which was free. The only drawback of the application was that the application only operated in Chinese. A web application called ‘Deep L’ was used and validated to be a very accurate translating application on the market. Deep L was used to help appropriately understand the functionality of the application. Happily, Octoparse did not require additional funds to scrape a

large sum of data and did not struggle with scraping when applications required login or opposed an ‘error 404’. Therefore, it was chosen as a primary source of data scraping because it was highly beneficial and not time-consuming.

3.5.5 Choosing what to Scarpe

Upon an application chosen for the task, a challenge struck, How to correctly gather a fair sample for testing. The possible options are choosing famous people at random, gathering from top x on the platform or gathering based on a real-life top list of celebrities. The approach taken for gathering the data could not have been celebrities chosen at random, as the relevance of a celebrity matters in their potential growth. It is not appropriate to compare an A list celebrity such as Tom Holland to a high school kid who recently gained 10k followers, As this would not have been a fair comparison. The second approach could not have been accurate as we have not gathered a sample of celebrities not from the application. The celebrities chosen from the application would have had a superior edge over non-users. Due to these factors, the final stage of data gathering had to come from a source that obtained a list of celebrities both on social media and without social media. This source came from a google search of “200 most famous people in 2021”[10]. From this list, a sample of top 50 celebrities was selected as they ranged in TikTok users and none TikTok users, Which is precisely needed for this experiment. This type of sampling was previously mentioned to be purposive and convenient.

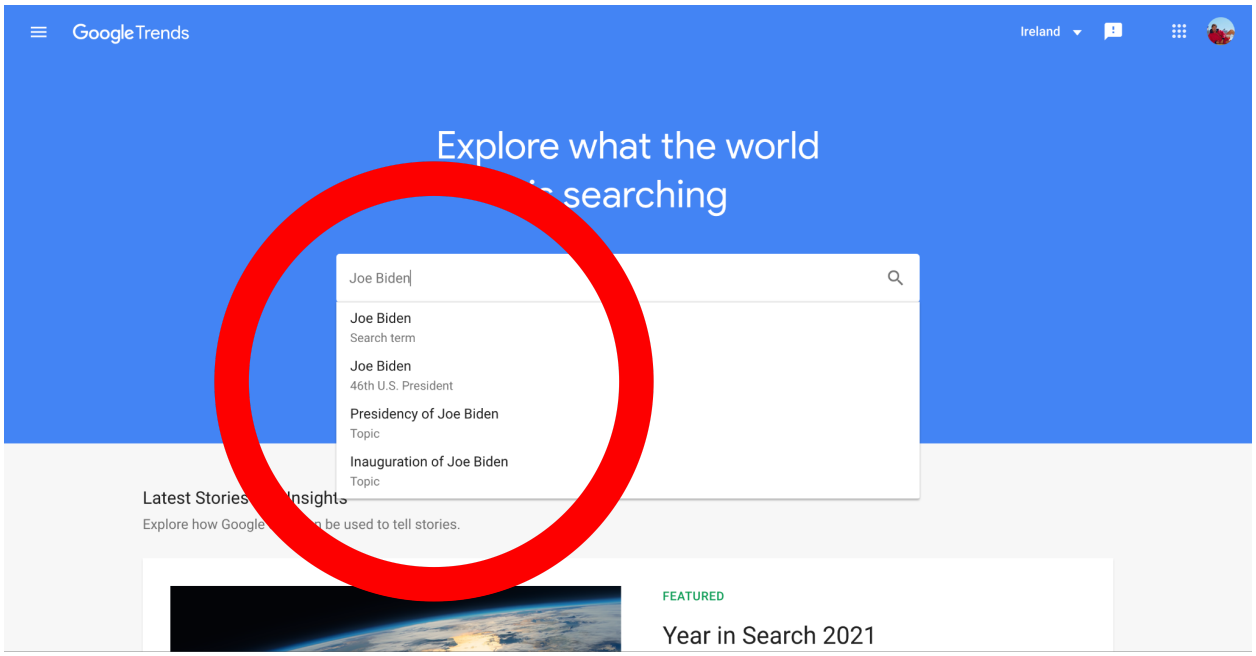
3.5.6 Data verification

Before collecting user data, users had to be identified and grouped into none TikTok users and TikTok users. Selecting users on social media was done by checking all users on TikTok alongside other social media platforms and finding their (‘url’ and ‘@name’). Upon identifying user activity on the site, TikTok users were then checked on other social media platforms such as Instagram, Youtube, and Twitter. Luckily, all users within the TikTok group were identified on all social media applications.

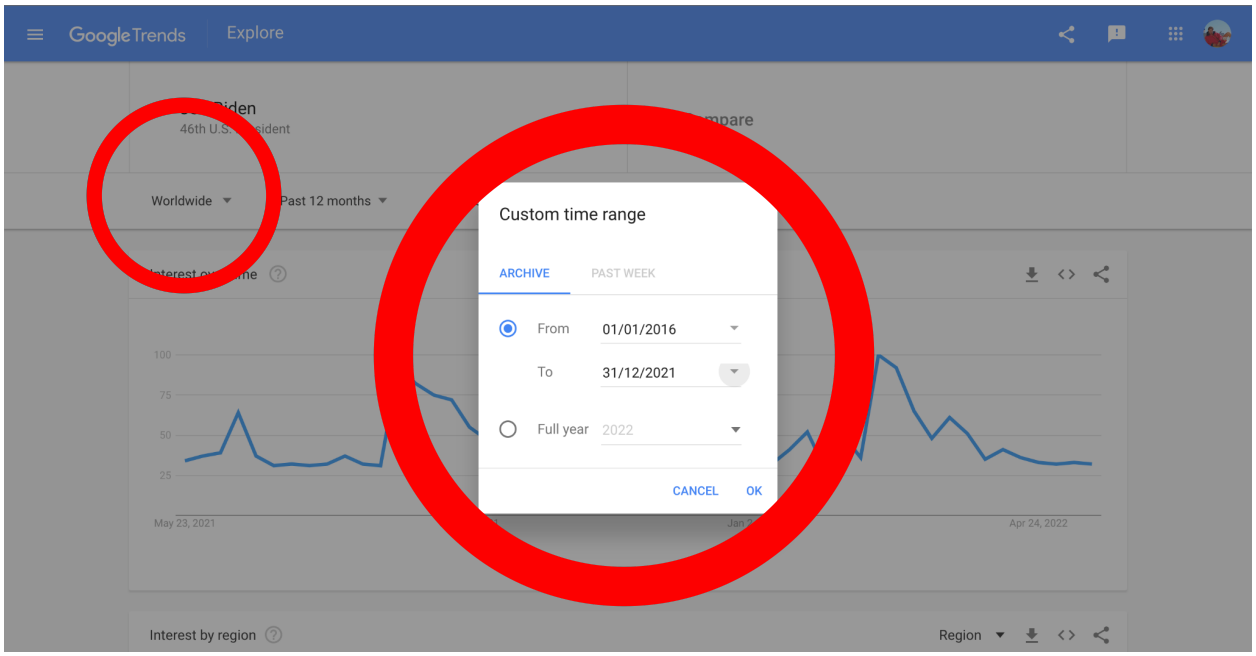
3.5.7 Google Trends

Google trends is a social media information centre that provides quantitive data since google released up to today’s data information as of May 2022.[11] Google also contains pre-programmed google trend prediction software that indicates prediction values up to 1-2 months ahead. The data was calculated by directly placing a celebrity name into the google Trend search bar and selecting the appropriate result for calculation[Fig.2.], In the example presented, the appropriate selection was “Joe Biden (46th U.S. President) ”, as Other search options would provide inaccurate results. A worldwide collection was taken alongside a data range from 2016 to 2021. 2016 was assigned as it is the starting date of when officially ‘TikTok’ launched into the market, and 2021 was assigned as it was the official date of the celebrities on

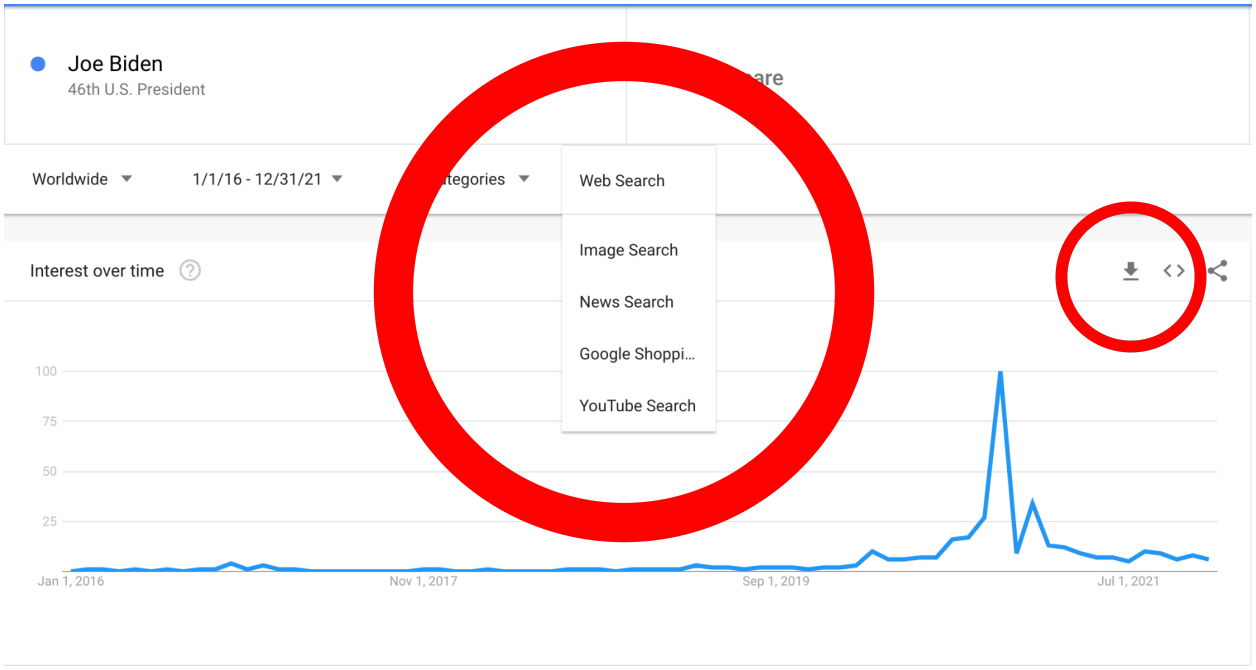
the website that was the focus.[Fig.3.] Finally, the appropriate terms such as youtube information, google search, news search and image search all cumulatively display a popularity of a user, so data was downloaded using a download button [Fig.4.].



[Fig.2.] How to search on Google Trends



[Fig.3.] How to adjust the Time on Google Trends



[Fig.4.] How to Adjust the data and How to Dowload on Google Trends

3.5.8 Excel

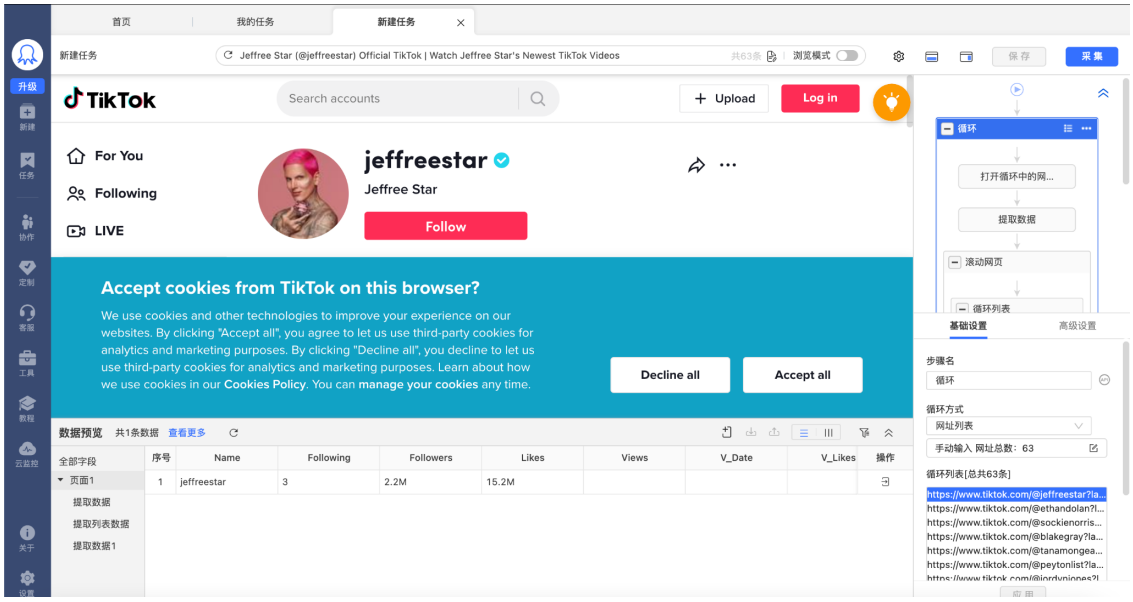
Excel was the primary source to combine all different files downloaded from google trends. Although the application is intuitive, there is no way of downloading a mass quantity of information at once, so a week of combining the files organising the structure and renaming column names was devoted. An additional column for each celebrity was made; this included using the '.average()' function to create a column summarising all individual celebrities' growth as it ranged between 0-100%. An average provided evidence of individual celebrities' overall growth rather than growth in a particular field, leading to an unfair advantage; for example, YouTubers who own TikTok would have an advantage in youtube searches over tiktokers who do not possess an account on youtube. The average was taken from all four fields for individual celebrities to counteract any advantages named "x total'.[Fig.5.] The result was that the dataset had a shape of 72 rows × 246 columns which included all top 50 celebrities scarped from Google Trends.

	A	B	C	D	E	F
1	Month	Charli D'Amelio web	Charli D'Amelio yt	Charli D'Amelio image	Charli D'Amelio news	Charli D'Amelio Total
2	2016-01	5	5	3	1	3.5
3	2016-02	4	5	2	1	3
4	2016-03	4	4	2	0	2.5
5	2016-04	4	4	1	1	2.5
6	2016-05	3	4	1	1	2.25
7	2016-06	3	3	1	0	1.75
8	2016-07	3	4	1	0	2
9	2016-08	3	4	1	1	2.25
10	2016-09	3	3	2	1	2.25
11	2016-10	3	4	1	0	2
12	2016-11	3	4	1	0	2
13	2016-12	2	4	1	1	2
14	2017-01	2	3	1	0	1.5

[Fig.5.] Scraped Google Trend Data Set which Also Includes Total Column for each Celebrity

3.5.9 Octoparse

OctoParse was a Chinese application so to set up. Firstly Application needs to be downloaded from ann official website. The next step is assigning what is needed to parse. Finally, run in bulk. The only requirement from the user was to gather all links into a list, which included looking them up on the Application and copy-pasting a verified account URL into a list of users folder. This method took around 2-7 hours to parse data for each platform, depending on each user's activity.[fig.6.]



[fig.6.] OctoParse Home Page

3.5.10 All data structure

Data from Google Trends can be seen in [Fig.5.]. It is a data that is short with only 72 rows but lengthy in width with more than double columns, sitting at 246 columns, All data were counted

as integers, and the initial row of the month was already date-time data. What was necessary was appropriate grouping into two groups, Tiktok users and None Tiktok users.
Data from the first parsing of TikTok users using octoparse can be seen in [Fig.6.]. It is data consisting of an index, name, followers, following and all attributes that can be seen in a video, including likes, views, comments and even the date posted. Two problems arose, including integers containing a B, M, or T, which stands for Billion, Million and Thousand respectfully. The second issue lay within the V_Date column, which was incorrectly portrayed as an object, contained data outside the range required (2016-2021) and was in daily data rather than monthly. This V_date error would result in data that does not match other than the user's name.
Which had to be also grouped into a singular entity column to match the original dataset as all users of TikTok in the first data set can be found in tiktokers data.
The final Three data were all alike, as seen in [fig.7.] all contain three columns, including date, user name and finally, the number of uploads on that application.
Moreover, since three different data all had the same size and differed by name and upload, it is only suitable to place them under the same branch.

yt_upload

Name	Date	yt_upload
charlidamelio	2021-12-30	
charlidamelio	2021-12-29	
charlidamelio	2021-12-27	
charlidamelio	2021-12-26	1
charlidamelio	2021-12-24	2
charlidamelio	2021-12-23	1
charlidamelio	2021-12-22	1

tw_upload

Name	Date	upload_tweet
charlidamelio	2021-11-15	3
charlidamelio	2021-11-14	12
charlidamelio	2021-11-12	41
charlidamelio	2021-11-09	5
charlidamelio	2021-11-07	3
charlidamelio	2021-11-03	1
charlidamelio	2021-11-02	3
charlidamelio	2021-11-01	8

insta_upload

Name	Date	Uploads
@charlidamelio	2021-10-17	9
@charlidamelio	2021-10-04	27
@charlidamelio	2021-08-16	45
@charlidamelio	2021-05-23	20
@charlidamelio	2021-04-24	5
@charlidamelio	2021-04-15	77
@charlidamelio	2020-11-17	3
@charlidamelio	2020-11-13	

[Fig.7.] All 3 Dataset including YouTube, Instagram and Twitter.

3.6 Data Pre-Processing

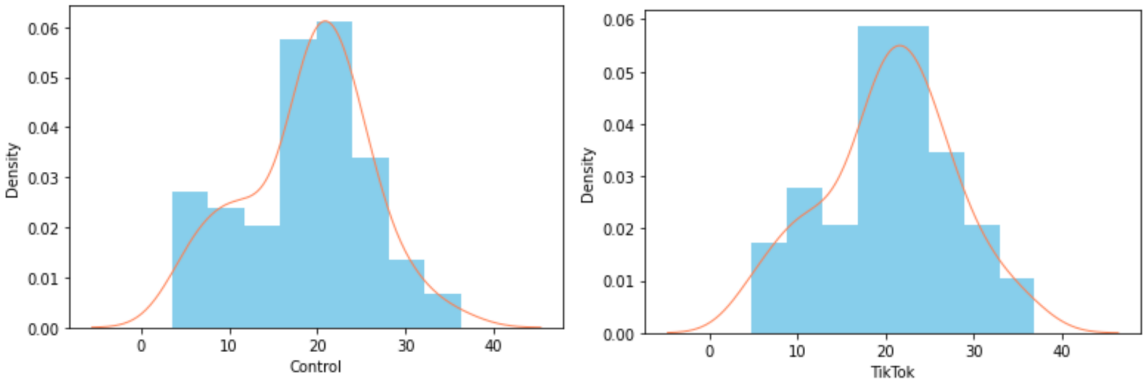
The first step was to decide the language used to decipher this data. Although R studios could have also been a solid choice for analysis, the approach used was ultimately python. Due to the data being divided into three types, so will data pre-processing. Firstly, the prediction dataset will be based on 'TikTok users vs. one TikTok user dataset' followed by the other datasets.

3.6.1 TikTokers vs none TikTok users

This Data set was scraped using Google Trends and combined through Excel was uploaded on python using pandas and read_csv function as a data frame. Pre uploading to python, a new column for each celebrity was created. This total column summarised all columns associated with that name, e.g., charlie youtube, google search, google images, google news, was averaged into charlie total. This new column was done as the total of each celebrity indicates its overall popularity online. If only one method were used, such as youtube searches, it would be unfair to

TikTok users who have a youtube account and vice versa. It would be unjust to users who do not own a youtube account; therefore, an average of all sites was a decision to satisfy all users. This data frame was then checked for type, meaning if it was a string, int, object, time, or boolean and if the shape of the dataset was correct. The entire shape of the data frame upon uploading was 72 rows × 246 columns[Fig.1.] The first step in this dataset was to check for null values and outliers. To fill in null values, fillna was used, a python function that can be filled with anything. In this case, it was filled with 0. The decision to check normalization was made after the grouping of data. Before grouping data, .drop was used to removing anything that was not the total as it was the fairest type of analysis.

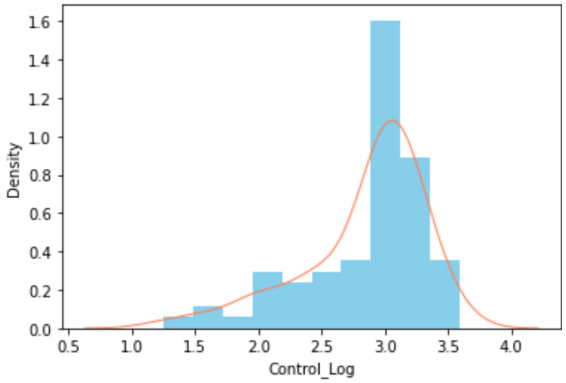
To group data, A list was made of all users who obtained TikTok. This list was then used to check in the data frame containing that person's name using the .contain function in python and finally filtered out by using the .drop function only to have a data set of the control group. Then to receive the group of TikTok users, a concat was made of both datasets and removed duplicate columns and duplicates found in the dataset leaving only TikTok users. The collaborated values were added using the median function, which provided a much more accurate result than the average. Then both Control and TikTok datasets were combined as they both share the same time. Nevertheless, both datasets were checked to see if they were normalized, but they are not perfect[fig.8].



[Fig.8.] Normalisation graph indicating Both TikTok and Control are almost perfect

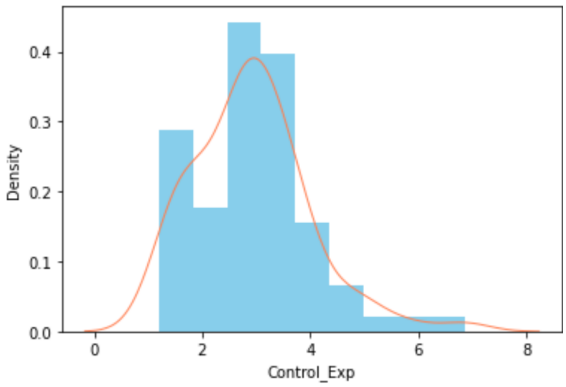
After attempting four normalization methods to control, the results could not get it to normalize to perfection.

The methods used were Logarithmic Transformation [fig.9.], resulting in worse data before the transformation.



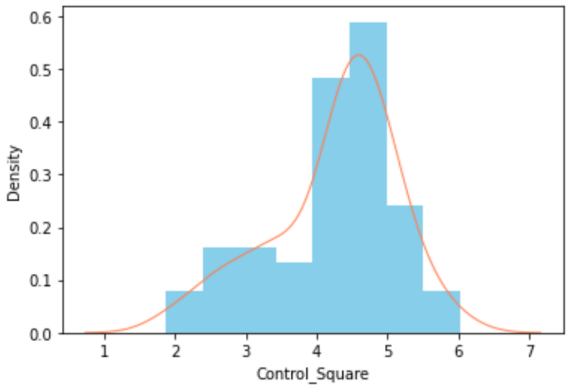
[fig.9.] Logarithmic Transformation

The second method used was Exponential Transformation [fig.10.], which also resulted in worse data before the transformation.



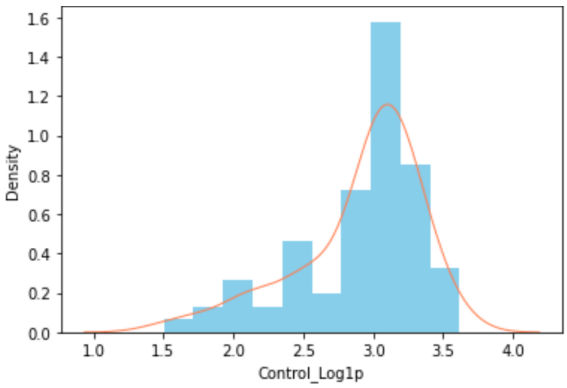
[fig.10.] Exponential Transformation

The Third method used was Power function transformation [fig.11.], resulting in worse data before the transformation.



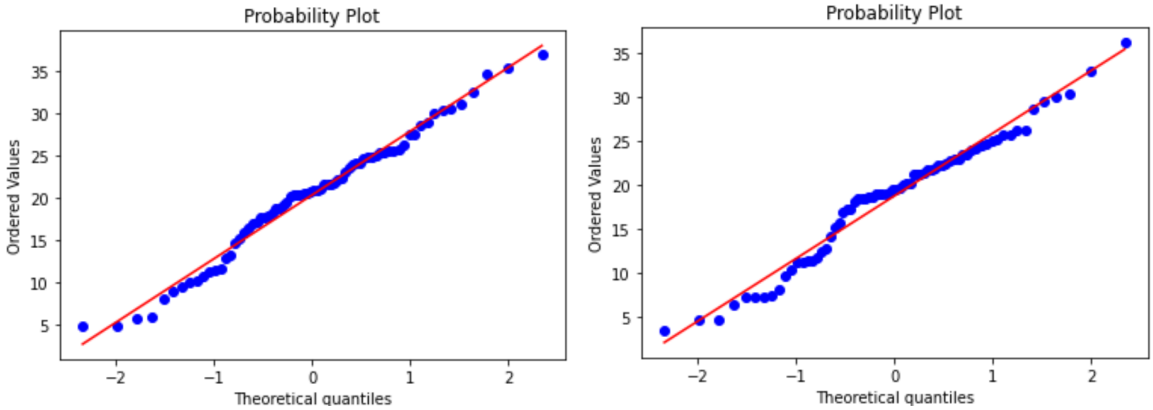
[fig.11.] Power function transformation

The final method used Log1p [Fig.12], without sounding like a broken record, it did not improve the normalization performance, resulting in an overall conclusion to leave this result as it is best left for convenience, as not all normalizations are perfect.



[fig.12.] Log1p

A Probability plot was made to ensure that the data was linear with a straight line. Undoubtedly, both TikTok users and control did indeed have an excellent probability plot. Then the finished product was made.[fig.13.]



[fig.13] probability plot For both Tiktok left and Control Right

A quick decision was made to find the difference between both datasets and analyze the prediction based on those results. This decision did indeed turn out to be a failure. The dataset difference was sporadic and seemed random, and predictions made also were the same, so the idea was scrapped as it was not very informative or helpful to the research.
Finally, a cumulative sum method turned out to be the golden goose of the research. It provided a sum of the percentages, and since no percentage was negative, the growth of both control and TikTok users was exponential, but at different rates. This method was done using a .cumsum function with the addition of .groupby. This sum performed well for analysis and prediction in all models used, such as Arima, Sarima, Lstm, and GRU.[Table.2].

[Table.2.] better representation of transformed TikTok_Uploads data.

Date	TikTok_uploads
2016-01	205
2016–02	185
…	…
2021-11	363
2021-12	381

Extra unused datasets were a result of brainstorming. The first one was a dataset separated into all different aspects found in social media, such as dancers, singers, YouTubers,tiktokers, and actors. This separation was created as it may have been necessary to identify a particular field when forming predictions and analyses. The idea was scrapped as the project focused on a more generalized group of TikTok users.
An additional data set was created with the idea of how early/late a user joined TikTok and if that had any effect on user growth.
Again, this idea was scrapped as the generalization had far greater importance than specification. Since the data type dealt with was quantitative, any future project could implement this type of grouping for personal studies

3.6.2 TikTok parsed data using OctroPrase

The dataset scraped was the original data scraped prior to any other dataset. Therefore it had the most amount of changes in it. As seen before in the above example, it began with importing the dataset into a data frame using pd.read_csv. The dataset was then printed, and the shape was identified [fig.14.]

	Name	Following	Followers	Likes	Views	V_Date	V_Likes	V_Comments
0	charlidamelio	1267	136.7M	10.6B	1.4M	3h ago	342.1K	19K
1	charlidamelio	1267	136.7M	10.6B	4.4M	2d ago	799.3K	55.6K
2	charlidamelio	1267	136.7M	10.6B	27M	4d ago	5.8M	83.1K
3	charlidamelio	1267	136.7M	10.6B	9.8M	4d ago	1.5M	35.4K
4	charlidamelio	1267	136.7M	10.6B	6.5M	4d ago	828.5K	31.4K
...	...	...	...	...	...	...	...	...
18782	gavinmagnus	4020	5.1M	156.8M	2.9M	2021-12-7	317.5K	13.2K
18783	gavinmagnus	4020	5.1M	156.8M	4.8M	2021-12-3	428.6K	6948
18784	gavinmagnus	4020	5.1M	156.8M	9.4M	2021-11-28	1.1M	37.5K
18785	gavinmagnus	4020	5.1M	156.8M	1.9M	2021-11-26	110.7K	3914
18786	gavinmagnus	4020	5.1M	156.8M	2.5M	2021-11-25	184.7K	3061

18787 rows × 8 columns

[fig.14.] Parsed Data of TikTok upload information for each Celebrity with a TT account

The shape of the dataset was 18787 rows × 8 columns, and the result needed for this dataset was a singular column summarising everything while also changing the date from object day to the time frame and month. This data pre-processing began with changing the date from an object to a time frame and setting it as the index. Then the time was selected between 2016-2022, which meant all data was collected from 01-01-2016 to 31-12-2021. Another adjustment was removing any unnecessary string associated with scraping the date, including the words "Ago" that resulted from scraping new data that was posted, e.g., 2 min ago, 2 months ago, and so on. Post-time fixes, the next step necessary was removing T, M, and B in the dataset, which represented t for thousand, m for million, and b for billion. This TMB was due to how the data was represented on the webpage and had to be replaced manually. This transformation was done using the Substitute method, where the associated word provided the sum of the associated word a correct multiplier, essentially creating that 4T into 4,000. This transformation is a little inaccurate as, in actuality, they might have had a larger or smaller sum. However, an approximation was made that it only gave a correct amount of 0 after the letter T, M, and B came up. The next step was to accurately change the time from day to month while at the same time transforming the data. Data were summarized in two ways. One way created a .sum to receive a sum of every day within that month for all users, ad the next was a .count which counted the number of times a video was uploaded. The result was two datasets that included the date in months and video views per month and a second dataset with the same time frame, but this time, the number of uploads were calculated using .count and .groupby.[Table.3.][Table.4.]

[Table.3.]Dataframe of uploads

Date	TikTok_uploads
2016-01	205
2016–02	185
…	…
2021-11	363
2021-12	381

[Table.4.]Dataframe of Views

Date	Views
2016-01	1822612
2016–02	2066889
…	…
2021-11	197879102
2021-12	198717388

This work was then ready for analysis. After completing the data set and comparing this dataset and the user dataset, an error occurred. Some months were missing. To deal with the disappearance of a couple of months. This resolution was made by creating a row for every month missing in months and deciding to fill in the missing values with 0 for simplicity. This resolution helped adjust the datasets to an even 72 rows and 1 column per dataset, including both views and uploads as the datasets specified.

3.6.3 Social media platforms datasets

These datasets were an attempt to solidify the defense for the hypothesis that helped investigation while also exp[anding the horizon of the possibilities made with the newly found datasets. Youtube, Twitter, and Instagram came in a similar format.[fig.15.]

tw_upload

Name	Date	upload_tweet
charlidamelio	2021-11-15	3
charlidamelio	2021-11-14	12
charlidamelio	2021-11-12	41
charlidamelio	2021-11-09	5
charlidamelio	2021-11-07	3
charlidamelio	2021-11-03	1
charlidamelio	2021-11-02	3
charlidamelio	2021-11-01	8

[Fig.15.] Twitter Upload dataset

The only difference is that more data was scraped through Instagram as it was the first scraped, so all unnecessary columns were dropped. All three datasets looked alike, With a user name, date, and upload quantity for each application. The first step was to remove null values and assign them with 0 using fillna. Next was to remove unnecessary text within DateTime and assign it correctly instead of an object to date. Then as previously seen, it was the summary of all users found on the left-hand side and a monthly sum of all uploads made by all users per month. This transformation resulted in a clean dataset, but the shape did not meet the requirements. The exact process of shape readjustment by filling in missing dates was added with 0. An error occurred when the data collected contained only values made by the previous year. The second problem occurred when the realization occurred that tweet data had many outliers ranging from hundreds to thousands. In general, tweets had a much higher frequency of combined monthly tweets than all applications combined. The solution was to divide the tweets dataset as the large sums would overshadow the other datasets while still maintaining the same amount of relevancy to the rest. The original problem of missing values from the previous year was left unresolved as correlation could still be identified with the information.

In the process of data pre-processing, a new CSV was made. This new CSV took results such as the date and TikTok users' data from the first data pre-processing and saving the CSV. The reasoning behind this decision was to use them for predictions. The combination of all Instagram, Twitter, and youtube upload data was transferred to a particular dataset alongside the TikTik user data set.[fig.16.]

	Month	Control	TikTok	TikTok_uploads	Insta_upload	upload_tweet	yt_upload
0	2016-01	3.50	4.750000	205	0	0.00	0
1	2016-02	4.75	4.750000	185	0	0.00	0
2	2016-03	7.25	5.750000	225	0	0.00	0
3	2016-04	9.75	5.916667	205	0	0.00	0
4	2016-05	7.50	10.000000	262	0	0.00	0
...	...	...	...	...	...	...	...
67	2021-08	24.75	20.625000	274	91	34.33	75
68	2021-09	21.25	17.625000	310	4	29.33	33
69	2021-10	19.00	14.625000	397	409	28.10	44
70	2021-11	18.75	15.875000	363	201	12.40	27
71	2021-12	23.00	17.625000	381	75	0.00	121

[fig.16.] Combined dataset

This combination was done to analyze further any similarities or correlations each feature has against one another, which will be explained in the results/findings.

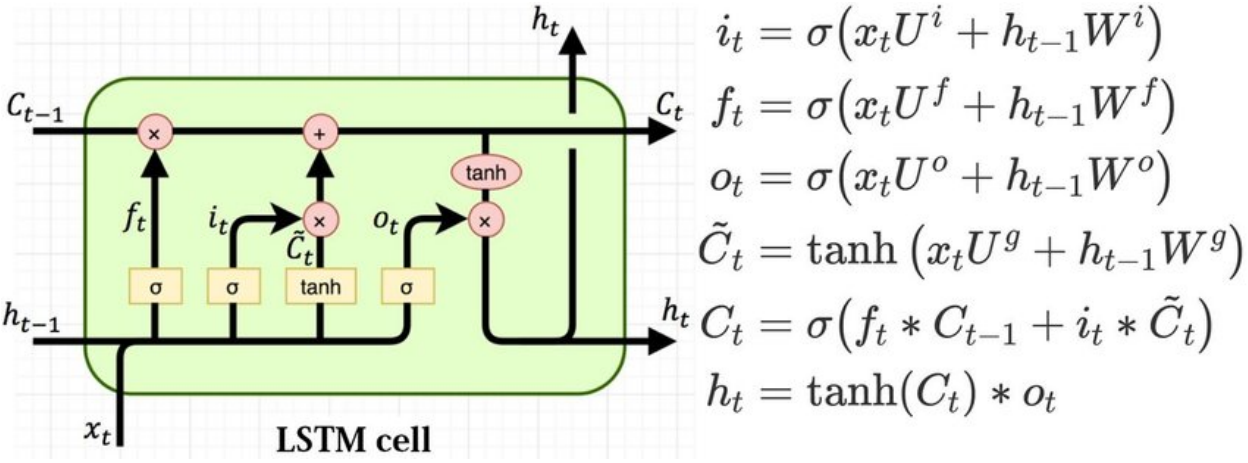
4. Analysis Design

There are multiple options to choose from for Time-series Supervised machine learning models. This Research used Arima, Sarima and Lstm.
Nevertheless, a model named Gru should be considered by anyone more in favour of lstm.

4.1 Lstm

The reason for choosing LSTM was simply that most pieces of literature used it, and it seemed like an innovative approach to a simple prediction.
Long Short-Term Memory, also known as LSTM for short. Lstm network is a type of neural network capable of learning long-term dependencies and is commonly prevalent in sequence prediction problems such as time series problems. This model is usually used to answer a problem when the time step is significant gradient gets either too large or too small, known as a

vanishing gradient problem. Lstm used three gates which are input, forget and output gates. These gates prioritize the data and determine whether the piece of data is allowed to pass. There are also two states in Lstm, which are cell and hidden. These states gather data for processing in the next state. [Fig.17]



[Fig.17] LSTM diagram

4.2 Arima

The reason for choosing Arima was that most kinds of literature did not use it, and it seemed like an innovative and most accurate approach to a simple prediction.
The Auto-Regressive Integrated Moving Average Model is used to fit the data. In simple terms, it helps to analyze the data and form a prediction. Auto-Regression uses a regression equation to use the previous observations in a dataset From the last/previous time steps. Series attempts to stabilize itself using a method known as subtracting an observation value from the previous value. [Eq.1]

$$X_t = \alpha + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_p X_{t-p} + 0_1 \epsilon_{t-1} + 0_2 \epsilon_{t-2} + 0_q \epsilon_{t-q}$$

where X_t is the variable that will be explained in time t [Eq.1]

where α is constant or Intercept

where β is coefficient with each parameter p

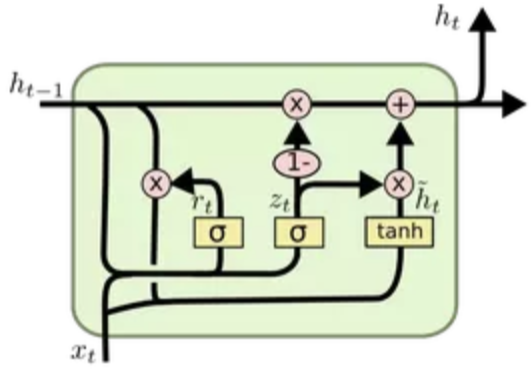
where 0 is coefficient with each parameter q

where ϵ_t is residual or errors in time t

This Formula is the Arima Equation. It is formulated from the average Moving Model combined with the Auto-Regression. The standard Notation of Arima is (p,d and q).
'p' refers to the number of lag observations included within the model, which helps when adjusting the prediction line fitted onto the model, and 'd' refers to the number of transformations that differentiate from one another, which are essential to turning the time series data stationary. Finally, 'q' refers to the overall size of the moving average window.
In this research case, the data is sorted by time and monthly values of the control Group alongside the TT group.

4.3 Gru

Gru in this research was covered because it is not spoken about within the topic of social media and Time-series analysis in literature. For novelty and theory, Gru was tested to find out if it is a reliable prediction model.
GRU Stands for Gated Recurrent Unit. It is very similar to LSTM as it is a recurrent Neural Network. The difference is that it obtained a less complicated structure than Lstm. Instead, it has four gates, an input and forget gate, but lacks an output gate. Instead, it has a gate z, gate r. Gate z is a gate that is an update gate, and gate r is a reset gate. These gates are vectors that pass the information on whether data can pass to the output.[Fig.18]



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

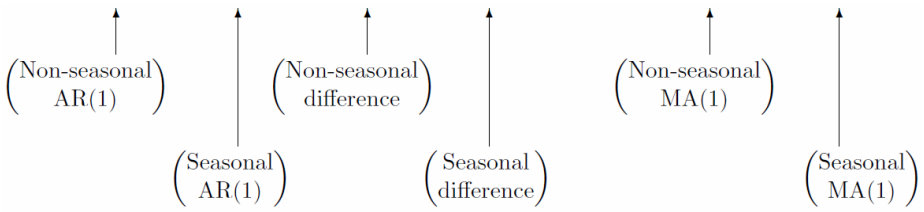
[Fig.18] LSTM diagram

4.4 Sarima

Sarima was involved in this project as it is less spoken about in pieces of literature and is a valuable asset within a researcher's inventory of models with regards to time series analysis.

Seasonal Auto-Regressive Integrated Moving Average Model is the same as Arima, except that instead of the actual data, it is analyzed and predicted in “seasons,” being months. Sarima changes the Original Standard Notations from (p,d,q) to seasonal (P, D, Q)m [Eq.2.]

$$(1 - \phi_1 B)\ (1 - \phi_1 B^4)\ (1 - B)\ (1 - B^4)\ y_t = (1 + 0_1 B)\ (1 + 0_1 B^4) e_t \qquad \text{[Eq.2.]}$$

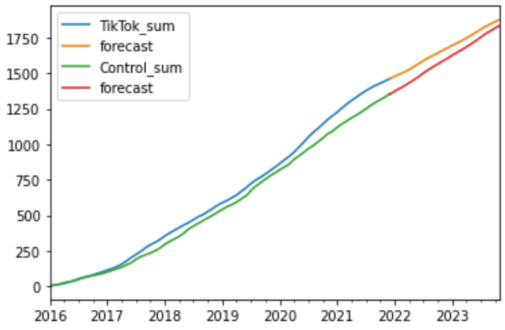


Analysis Design Arima has a superior edge in its prediction power over GRU and lstm. Hence, it was used in the research to predict the potential growth of TT users and none TT users when determining the superiority of one.

5. Results/Findings

In the results and findings, three main questions will be discussed in three parts, Does TikTok hold significance within the short/long term? Does TikTok correlate with the TikTok user's google trends data regarding "Fame"? Compare social media competitors and distinguish the differences they may hold against the TikTok user data from google trends.

- The first objective is to analyse the Significance of TikTok user superiority over none TikTok users. Initially, information came in as "insignificant" and "random", but significance was found upon improving the dataset and the Arima model. This significance was the result of Two predictions. The cumulative sum (meaning the sum of the first value - the previous value/s) of Both the Control Group and TikTok user Group dictated an advantage in owning a TikTok Account [Fig.19].



[Fig.18] Forecast of TikTok users Vs None - TikTok users

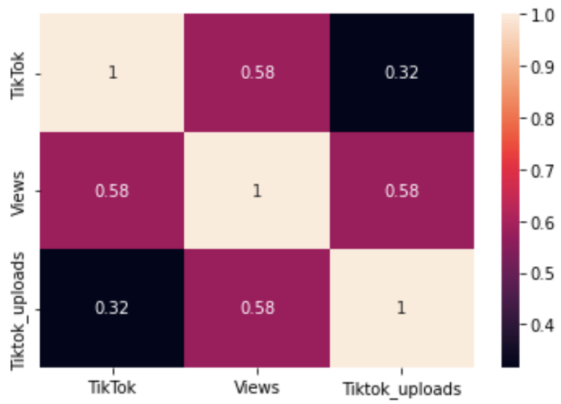
This information is a slight/some significance that looks to be nothing at first sight, But this is what separates the good from the great. The slight difference of what can be seen in [Table.3.]

[Table.3.] Recreation of the Sums, Forecast and Difference results from Arima

	Sum	**Forecast**
TikTok	1462.458333	1879.823332
Control	1355.5	1837.951291
Difference	**106.95833**	**41.87204**

shows a 41.872 difference between the two models. This difference has to be considered a percentage, and 42% in a singular month could change popularity from hypothetically 50% to a respectable 92%. In summary, the hypothesis found some significance as predicting A company's future growth provided evidence of superiority and a lack of evidence of a control group ever surpassing the TT user group in the future. Although this hypothesis was proven, further studies on the matter would be ideal for fully defending the hypothesis.

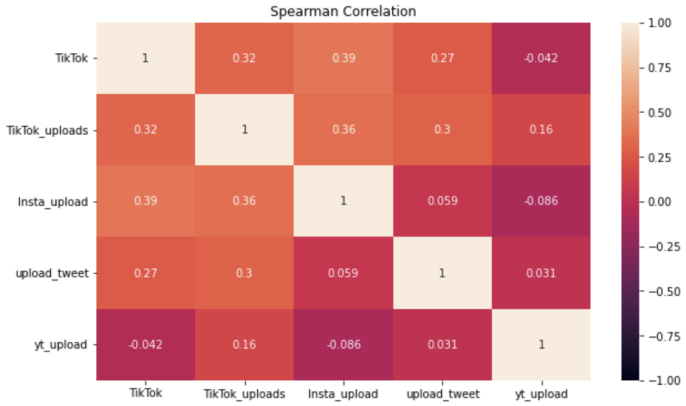
- Secondly, The Defence of TikTok Has a significant correlation with the TikTok users, Identifying the Pragmatism query that TikTok users are growing because of the interaction with content creation on the application. This hypothesis was proven using Both TikTok Upload Quantity and Views Obtained prior to Data pre-processing, leading to the following results.[Fig.20.]



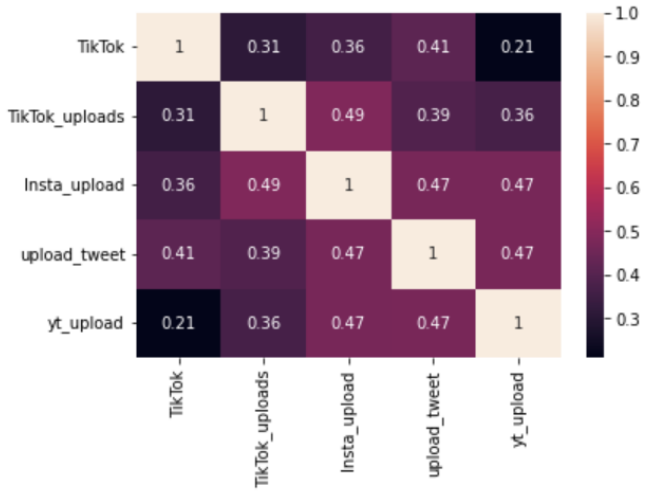
[Fig.20]TikTok Users vsTikTok Upload Quantity and Views Relationship

In the Correlation Heatmap, Clear evidence of superior correlation was found within the content of the information meaning Views compared to the number of uploads. This information is Significant evidence provided by the research that the information taken from the content has a greater significance to the quantity, At least in the prediction of Future Growth. This evidence helps with the later study of All Applications. It provides evidence that anything found to form the analysis can be multiplied through the information of the upload quantity. It is not to say that upload quantity is irrelevant to the matter, but further study with more features/variables such as views, likes and comments would defend this hypothesis.

- Thirdly, The defence of Social media Platforms is significant with TikTok users' growth. Two Correlation heatmaps have been created post data pre-processing to identify the relationship of each feature. [Fig.21.] In the Correlation Matrix Heatmap and [Fig.22.]



[Fig.21.] Correlation matrix (Heatmap) between all features



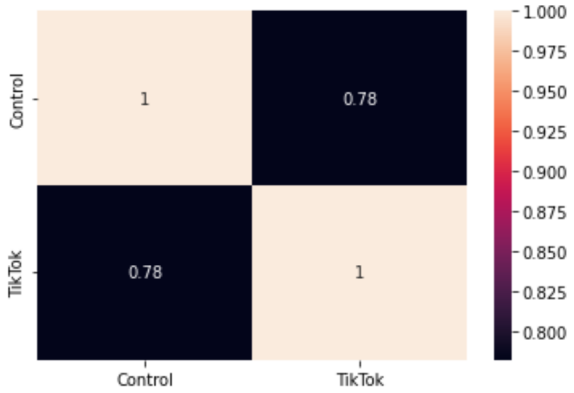
[Fig.22.] Correlation Matrix Spearman between all features

Correlation matrix heatmap including "Spearman", we can firstly identify that in the first [Fig.1.], the correlation between TikTok users and social media applications seemed to be especially low, indicating low overall significance. It yet confirmed the earlier statement made on point number two. It could be justified with the usage of Application information, including Likes/Views and Comments, rather than the number of uploads in that particular application. In figure 2, we can see some correlation between all applications, specifically Instagram, regarding the Speakman Heatmap correlation diagram. This significance displays evidence of cross usage of applications, specifically among users of Instagram. Hypothetically users who post on Instagram frequently upload on other social media sites creating an interrelationship between content creation on social media. This information would require further unsupervised analysis research using correlative models such as KNN, Decision Trees, clustering, and Associations to determine the true significance of the relationship between various applications.

Nevertheless, The information provided indicates some evidence of that interrelationship present in the research. This interrelationship proves that the overall popularity is more complex, meaning it is the combination of multiple factors that lead to the growth of a celebrities growth at least on Google Trends. This Surplus of knowledge is a piece of valuable information within the research studies. It provides clear evidence of focus when determining the root cause of popularity when analysing future growth.

Furthermore, Evidence of Control Group having the highest correlation of growth comparatively to TikTok users was vital information identified by accident. This information indicates that celebrities correlate highly and hold a strong correlation coefficient. This high correlation is

essential information that researchers could use in future studies when attempting to predict an individual / grouped future growth/trend/downfall prediction.[fig.23.]

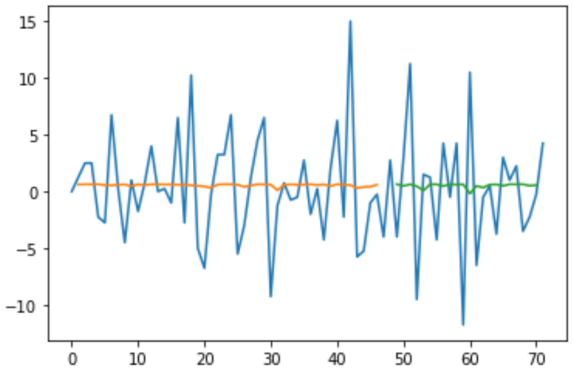


[fig.23.] TikTok vs Control Correlation Heatmap

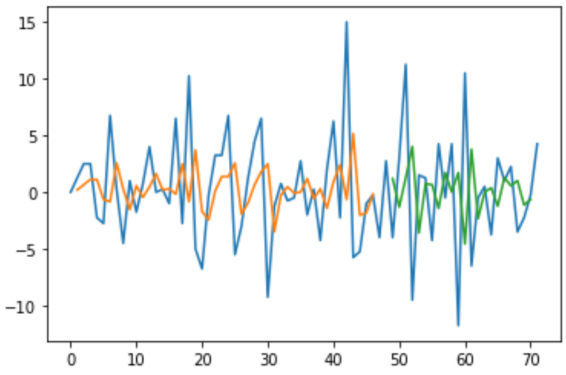
5.1 Improvements

Improving the research came in a couple of different ways such as Improving LSTM Through adjustment. Improving Sarima Model through making the data stationary also from adjustment. Spearman and visualistaion better portrait teh research objectives imporving the overall performance of teh research. Finally the late addition of Gru in teh project provided the Knowledge of a secondary possible model for prediction, This is an extension of LSTM called Biderational LSTM which allowed for a overall better performance in the prediction.

- In LSTM, to increase the accuracy of the performance, epochs were interchanged through trial and error from 120 down to 40 to increase further its overall performance changing form [fig.24.] to [fig.25.]



[fig.24.] pre Changing epoches (120)



[fig.25.] Post changing epochs (40)

- Arima TikTok prediction maintained Non-stationary when attempting to fit the first difference. This evidence is terrible as Arima needs to work with static data. The improvement made was the change of the integer in the shift. This change did indeed

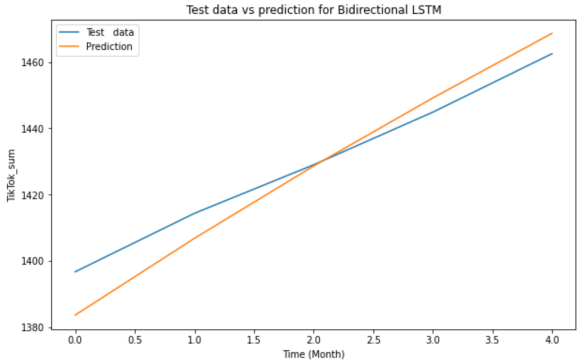
decrease the number of observations. However, it changed the data from non-stationary to stationary. The data applicable to the research resulted in the diagram [Fig.1]. And improvement in the performance.

- The correlation between TikTok. User dataset and uploads were very low. To further analyse this significance shown in the results and findings, analysing the information alongside the upload quantity seemed to benefit the research.

- The correlation between applications and the TikTok user dataset was significantly low, but no information data was presented. A spearman method was applied to the correlation to overcome this obstacle, which presented a new viewpoint and allowed the user to spectate the potential interrelationship between multiple applications displayed in the heatmap.[fig18.]

- Visualisation of data rather than its integer representation provided a more knowledgeable insight into the analysis of the findings.

- The Addition to Gru was not a total Failure as it provided evidence of a Low ME and RMSE in the BiDirectional LSTM this is great news for future researchers looking for a Time-series model that suits social media data and provides good results.[fig.26.][Table.4.]



[Fig.26.] GRU prediction Performance

[Table.4.] ME and RMSE found in GRU and Bidirectional Prediction

	Mean Absolute Error	**Root Mean Squared Error**
GRU	45.6921	6.2804
Bidirectional LSTM	46.0671	7.5337

5.2 Acknowledge Failures

Sadly, not all things worked out as this research attempted to try out new overused models for trend analysis such as gru it failed it do so due to accuracy. Moreover LSTM also underperformed and was therefore left out of the literature.

- Gru was a supervised machine learning model that is not necessarily new but it does provide novelty to this research as it hasn't been done before. Sadly it resulted in a poor prediction of approximately 46% in the Mean Error and for the Mean Absolute Squared Error, it performed at also 46% this poor result led to the discontinuation of the prediction as valid evidence.
- LSTM itself, without the extension of bidirectional LSTM, did not meet the satisfactory prediction accuracy, This sadly resulted in the drop of LSTM and a full focus on the Arima model.

6. Conclusion/Future work

This study revealed "TikTok" was indeed the Platform to achieve greatness, specifically To be of significance within the scraped Celebrities Dataset. This significance was proven using three models, LSTM, ARIMA/SARIMA, and GRU, but only one model proved to be the most accurate, This model being Arima. The results showed that Tiktok users had a more excellent online performance than their non-TT counterparts; this edge over competition allowed this project to be the first to identify the importance of a single application to impact the overall growth of a user.
Moreover, the further analysis identified a significant relation between TikTok users and the information stored on the users' TikTok application contents.
The final piece of information revealed the interrelationship between models. Using spearman on the correlation function displayed a high correlation between Instagram as the core and all other applications.

- The research proved some significance within the project's incentive of analysing the significance of TikTok concerning popularity/Fame. This significance proved a predicted advantage over the control group of 41.872%, and a standard advantage pre prediction analysed from the cumulated sum to be 106.96%. The information provides Strong significance in the studied group that TikTok users outweigh the percentage of none TikTok users by almost 110%.

- Through The literature review found in this research, a few important things have been identified; The history of social media applications, Web2.0, Time series analysis and different strategies, approaches to Trend Analysis, Time series prediction and how literature progressed over the years. Therese pieces of information allowed this research to perform to the fullest of its potential and increased the topic's popularity within the Literature Space, following the current trend of Famous works of literature also gaining interest in this topic space.

This research Contributed To the Research of “TikTok” and “Fame” ad they lack literatures online. This research followed what was already talked about which is Trend analysis but took a unique twist on the matter and a specific combinations of tasks including the research of fiding teh significance of an application in relation to user growth which was never doen before in any literature that was to made to my best knowledge.

This study Lacked in further informative data from the applications such as Likes, Views and commets,Futhermore research lacked a mix of qualitative and quantitative data. This limited the perform as unsupervised learning models were not used in identifying relationship between features.

As previously mentioned, Future research must overcome the challenges left in this research if the research were presented with an opportunity to continue the research at hand. It would begin by defending the hypothesis made in this research and finding new pieces of information that provide both Academics and Businesses with the information, whether a Social media platform that Positively or Negatively impacts their Growth online. The following hypothesis would be Defended as "Quality over Quantity", a hypothesis made on the number of uploads compared to the content contained within those uploads, such as likes, comments, shares, and views. The second hypothesis would identify using a more considerable sum of users. To "Verification of All social media platforms performance regarding their popularity and Their Strengths/weaknesses".This defence would continue to take More Google Trends data and More Data scraping further to analyse the potential information and flaws in the research. Finally, To "prove the relationship between a single Application and User Growth outside the application", this hypothesis defends the one made on TikTok vs TikTok users. It was clarified that it holds significance. This sign needs to be verified for this careful research to be reliable. It would be a personal burden to continue developing this paper with the solving of the Pragmatism query as the end goal.

This study persuaded me into enrolling into a data science course in Tudublin where hopefully a further/deeping study can be made on the analysis of Social media influence with relation to “Fame”.

References

A, R. M., & A, A. (2018). Research methodology in business: A starter's guide. *Management and organizational studies*, *5*(1), 1-14.

Battisby, A. (n.d.). *An In-Depth Look at Marketing on TikTok | Blog | Online Digital Marketing Courses*. Digital Marketing Institute. Retrieved May 24, 2022, from https://digitalmarketinginstitute.com/blog/an-in-depth-look-at-marketing-on-tiktok

C, M., H, Y., & D, E. J. (2021). On the psychology of TikTok use: A first glimpse from empirical findings. *Frontiers in public health*, *9*, 62.

Carta, S., Podda, S. A., Recupero, R. D., & Usai, G. (2020). Popularity prediction of instagram posts. *Information*, *11*(9), 453. https://www.mdpi.com/2078-2489/11/9/453

Google Trends. (n.d.). Google Trends. Retrieved May 23, 2022, from https://trends.google.com/trends/?geo=IE

Gundecha, Pritam, & Liu, H. (2012). Mining social media: a brief introduction. *New directions in informatics, optimization, logistics, and production*, 1-17. https://doi.org/10.1287/educ.1120.0105

Hou, L. (2018). Study on the perceived popularity of TikTok. extension://pjmlamaidnkoemaaofddboidllnogmhe/http://dspace.bu.ac.th/bitstream/123456789/3649/1/Hou%20Liqian.pdf

Introduction - Basics of Octoparse. (n.d.). Octoparse. Retrieved May 24, 2022, from https://www.octoparse.com/doc-wf/introduction

Jain, R. (n.d.). Rajat Jain - Top Digital Marketing Consultant. Retrieved May 23, 2022, from https://www.speakrj.com/

Jeon, S., Hong, B., Kim, J., & Lee, H. J. (2016). Stock price prediction based on stock big data and pattern graph analysis. *SCITEPRESS*, *2*, 223-231. 10.5220/0005876102230231

M., K. A., & M., H. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business horizons. *53*(1), 59-68. https://doi.org/10.1016/j.bushor.2009.09.003

Mathioudakis, Michael, & Koudas, N. (2010, June). Twittermonitor: trend detection over the twitter stream. *The Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 1155-1158.

Mudinas, A., Zhang, D., & Levene, M. (2019). Market trend prediction using sentiment analysis: lessons learned and paths forward. *arXiv:1903.05440v1*. extension://pjmlamaidnkoemaaofddboidllnogmhe/https://arxiv.org/pdf/1903.05440.pdf

Pavlyshenko, M. B. (2019). Machine-learning models for sales time series forecasting. Data. *4*(1), 15. extension://pjmlamaidnkoemaaofddboidllnogmhe/https://mdpi-res.com/d_attachment/data/data-04-00015/article_deploy/data-04-00015-v2.pdf?version=1548153423

S., A., & A., H. B. (2010). Predicting the Future with Social Media. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 492-499. 10.1109/WI-IAT.2010.63

Shams,, B. M., Hossain, J. M., & Noori, S. R. H. (2020, July). Time Series Analysis of Trends With Twitter Hashtags Using LSTM. *In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1-6. 10.1109/ICCCNT49239.2020.9225349

200 Most Famous People 2021. (n.d.). List Challenges. Retrieved May 23, 2022, from https://www.listchallenges.com/200-most-famous-people-2021

Yamak, P. T., Yujian, L., & Gadosey, P. K. (2019). A Comparison between ARIMA, LSTM, and GRU for Time Series Forecasting. *In Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, 49-55. https://doi.org/10.1145/3377713.3377722