Enhancing Sentiment Analysis of Twitter Data: Comparative Study and Fine-Tuning machine learning Model

Supervised by:

Afsana Begum

Associate Professor & Coordinator M.Sc

Department of Software Engineering

Submitted by:

Zarif Ahmed

ID: 221-44-239

Department of Software Engineering

Daffodil International University

A thesis submitted in partial fulfillment of the requirement for the degree of

Master of Science in Software Engineering

Department of Software Engineering

Daffodil International University Dhaka, Bangladesh

Semester-Spring-2023

APPROVAL

This thesis/project/internship titled on ""Enhancing Sentiment Analysis of Twitter Data: Comparative Study and Fine-Tuning machine learning Model", submitted by Zarif Ahmed, ID: 221-44-239 to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Masters of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMI

-- Chairman

Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Daffodil International University

Fazla Elah

-- Internal Examiner 1

Dr. Md. Fazla Elahe
Assistant Professor and Associate Head
Department of Software Engineering
Daffodil International University

-- Internal Examiner 2

Afsana Begum
Assistant Professor
Department of Software Engineering
Daffodil International University

-- External Examiner

Dr. Md. Sazzadur Rahman,
Associate Professor
Institute of Information Technology
Jahangirnagar University

Declaration

I hereby declare that this thesis has been done by Zarif Ahmed, ID: 221-44-239 under the supervision of Afsana Begum, Associate Professor & Coordinator M.Sc ,Department of Software Engineering, Daffodil International University. It also declare that neither this thesis nor any part of this has been submitted elsewhere for award of any degree.

Zarif Ahmed

....
Zarif Ahmed
ID:221-44-239
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Certified by:

....
Afsana Begum,
Associate Professor & Coordinator M.Sc
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

This thesis titled “Enhancing Sentiment Analysis of Twitter Data: Comparative Study and Fine-Tuning machine learning Model.”, submitted by Zarif Ahmed, ID: 221-44-239 and to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Master of Science in Software Engineering and approval as to its style and contents.

Acknowledgement

In this very special moment, first and foremost I would like to express my heartiest gratitude to the almighty Allah for allowing me to accomplish this M.SC study successfully. I am really thankful for the enormous blessings that the Almighty has bestowed upon me not only during my study period but also throughout my life. In achieving the gigantic goal, I have gone through the interactions with and help from other people, and would like to extend my deepest appreciation to those who have contributed to this dissertation itself in an essential way. I would like to take the opportunity to express my gratitude to my respected supervisor Afsana Begum madam. She has been a great influence throughout and I am grateful to him for not only her intellectual support, providing me with vast knowledge, but also for the moral support and proper guidelines that he gave me, encouraging me always to work harder, raise the bar and his honest quest to extract the best out of us. It is needless to say without her influence and support this dissertation would never be successful.

In the finishing note, I am thankful to my family and friends for their help and immense support in different aspects to make this dissertation a success.

Abstract

Because so much work is being put into data mining and information closure, sentiment is a typical technique for people to communicate the current trends in internet-based lives. Using sentiment analysis, a number of potential events can be found, including human trafficking, break-ins, and unclear feelings, and so on. Rapid Miner is a social media mining application we use to get unrefined dataset about cricket betting and matchfixing using hashtags. After cleaning the dataset via stemming, lemmatizing, analysing of sentiment etc we get a refined dataset. This study's primary objective is to use sentiment analysis on texts, evaluating the effectiveness of machine learning algorithms such as KNN, Naïve Bayes, Logistic regression and then checking which model has the highest accuracy and fine tune the model into further increasing the accuracy.

Keywords: Rapid miner, Python, Cloud mining, Sentiment, LR, KNN, Mapping, New model

v

Table of contents

Figures

Chapter 1 Introduction

1.1 Overview

Similar to grid computing, cloud computing, often referred to as distributed computing, is a web-based computing platform that enables the on-demand sharing of resources among devices. Data is saved on servers rather than on local workstations because it uses a web paradigm . One of the main benefits of cloud computing is that it removes the need for local servers by enabling on-demand file sharing through the internet. Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) are the three basic methodologies that make up cloud computing. Services that are commonly accessed through software platforms, like web browsers, are referred to as SaaS. A platform-based service called PaaS offers a hosting environment for the deployment of cloud applications. The platform provided by IaaS, on the other hand, can be used to deploy applications across numerous locations[1]. The potential of cloud computing to lower maintenance costs for infrastructure and hardware is one of its fundamental benefits. Organizations can transfer the responsibility for operating and maintaining physical servers and hardware components by utilizing cloud services. This enables companies to more effectively allocate resources and concentrate on their core capabilities.

A wide range of architectural styles, such as hybrid, public, and private architectures, are also present in cloud computing. Hybrid clouds give enterprises more flexibility and scalability by combining the usage of both public and private cloud infrastructures. On the other hand, public clouds are open to all users and are run by external service providers.

Contrarily, private clouds are devoted to a particular company or entity and are often controlled internally.[2] In conclusion, cloud computing is a web-based computing platform that enables on-demand resource sharing among devices. It provides a variety of service models, including SaaS, PaaS, and IaaS, to meet varied demands and specifications. Utilizing cloud services allows businesses to save money on hardware and infrastructure while also taking advantage of the flexibility and scalability that various cloud architecture types offer.

The idea of "cloud mining" encourages the use of cloud computing platforms for mining. Mining synchronization can be accomplished in cloud systems by employing databases, programming tools, servers, and storage resources. Cloud mining includes using remote server facilities to share computer power, allowing users to mine bitcoin or other digital currencies without having to keep track of different types of gear. Because it functions as a Software as a Service (SaaS) system, the administration has to pay certain expenses while providing the client with reduced returns. Opinion mining, a branch of sentiment analysis, is essential to this procedure since it uses textual data to produce emotional outputs. We can successfully distinguish between the assumptions held by various people by using impact score analysis. Data is gathered for this methodology-driven sentiment analysis from online social media sites like Twitter and Facebook. Mining algorithms help by segmenting the sentences or content once the data has been stored. The computers then assign tags to the sentences and determine whether they are positive, negative, or neutral. The technique uncovers the depth of a person's sentimental assumptions by using polarity schemes. The approach then outputs an analysis of the result after estimating the intensity. This study proposes a novel method for sentiment analysis that relies on approximation sentiment prediction algorithms. This study compares the effectiveness of the various algorithms and describes the workflow of sentiment analysis under specified settings. The study also considers the difficulties in locating more accurate indicator factors within individuals' cognitive processes. This comparison is being done in order to identify which algorithm performs better when used for sentiment analysis.

1.2 Contribution

This study's original approach of using sentiment analysis on cricket betting. As a result, we focused on the betting of international cricket and gathered the essential information from our corresponding Twitter accounts and Twitter cloud repositories. The use of continuous real-time Twitter cloud data collected from cricket matches in 2023 was then evaluated.

Using text mining, we appended a comment to the text to describe the expected attitudes of the readers. The evaluation and configuration of classification data mining algorithms including K-Nearest Neighbors (KNN), Logistic Regression (LR), Naive Bayes, and Random Forest etc. key goals of this study is to check the highest accuracy of all the models and create a new model which give higher accuracy of the said model. There are various sections in the paper's structure. The "Introduction" part of the paper gives a general summary of the goals and setting of the study. The second section presents a thorough examination of the pertinent literature, outlining the state of the art in sentiment analysis, data mining, and classification approaches. We go into the methods used in the third section, "Methodology" outlining the procedures required to gather and examine the Twitter data. The results of our investigation are presented in the fourth section, "Results and discussions," which also shows how accurate the naïve bayes algorithm was at categorizing relationships. We also discuss the advantages and disadvantages of each method while contrasting these outcomes with the performance. Then we create a new model based on the naïve bayes algorithm using a parameter called ‘alpha’. And it the fifth section, "Limitations and Future Works," describes the difficulties encountered during the study and suggests directions for further investigation in order to overcome them and improve sentiment analysis methods as they apply to cricket players.

The final part, "Conclusion and Final Comments," highlights the main conclusions and contributions of the study while highlighting the value of the KNN algorithm for sentiment analysis and its potential to be used to understand how the general public feels about cricket players.

2 Literature review

[1] A paper was studied, the paper concentrates on utilizing Twitter, a leading microblogging platform, to conduct sentiment analysis. The authors outline their process of automatically gathering a corpus for sentiment analysis and opinion mining. Through linguistic analysis of the collected corpus, they uncover noteworthy phenomena. They develop a sentiment classifier using this corpus, capable of discerning positive, negative, and neutral sentiments in a document. The paper's main contribution lies in demonstrating the efficacy of their techniques through experimental evaluations, which surpass the performance of prior methods. An additional highlight is the applicability of their approach to languages beyond English. However, a limitation of the study could be its exclusive focus on Twitter data, potentially overlooking sentiments expressed in other contexts or platforms [1].

[2] describes an evaluation process that assesses the effectiveness of current lexical resources and features designed to capture the informal and creative language prevalent in microblogging. The approach involves utilizing a supervised method, while also incorporating existing Twitter hashtags to construct training data. The primary contribution lies in this combined evaluation strategy. However, a limitation could be the reliance solely on hashtags for building training data, potentially neglecting other valuable linguistic cues present in microblogging content [2]. [3] Twitter's significance as a platform for political opinions and its connection to electoral events. It highlights the rapid response of Twitter users to emerging news or events, allowing for real-time analysis of the relationship between public sentiment and political occurrences. The main contribution is a system that performs sentiment analysis on the entire Twitter conversation regarding an election, providing immediate and continuous

results. This offers a novel and timely perspective for the public, media, politicians, and scholars to understand the dynamics of the electoral process and public opinion. However, a potential limitation could be the system's reliance on sentiment analysis, which might not fully capture the complexity and nuances of political discourse on the platform [3]. [4] An automated sentiment detection approach for tweets is introduced. The approach involves analyzing tweet writing characteristics and meta-information of words in the messages. It leverages noisy labels obtained from sentiment detection websites as training data. The primary contribution is an improved solution that captures abstract tweet representations, making it more effective and robust compared to previous methods. It specifically addresses biased and noisy data inherent in the provided sources. However, a limitation could be the reliance on noisy labels, which might introduce inaccuracies and impact the model's performance in certain cases [4]. [5] the challenges presented by Twitter data streams, particularly in the context of classification problems. It highlights the application of these streams for opinion mining and sentiment analysis. To address the issue of unbalanced classes in streaming data, the authors introduce a sliding window Kappa statistic for evaluating changing data streams over time. The main contribution is the proposed Kappa statistic, enabling evaluation of evolving data streams. The authors conduct a study on Twitter data using machine learning algorithms designed for data streams. A limitation could be that the focus on the Kappa statistic may not fully address all challenges related to unbalanced classes or evolving data streams, potentially requiring additional techniques for comprehensive analysis [5].

[6] A sentiment prediction approach applied to three Twitter datasets. The approach demonstrates an average increase in the F harmonic accuracy score for identifying negative and positive sentiments, with approximately 6.5% and 4.8% improvements over baseline methods using unigrams and part-of-speech features, respectively. The main contribution is the introduction of semantic features, leading to enhanced sentiment classification results. The method is also compared to sentiment-bearing topic analysis, revealing that semantic features offer better Recall and F score for negative sentiment classification, and improved Precision with lower Recall and F score for positive sentiment classification. A limitation might be the

lack of detailed explanation about the semantic features used and potential challenges they may introduce, such as potential overfitting or complex feature extraction[6]. [7] a fundamental aspect of sentiment analysis. It introduces a comprehensive process for this categorization, specifically targeting online product reviews from Amazon.com. The process details are provided. The study conducts experiments for both sentence-level and review-level categorization, yielding positive results. The main contribution lies in proposing a structured process and conducting successful categorization experiments. However, the limitation is that the paper doesn't elaborate on potential challenges faced during the process or the limitations of their approach. The conclusion hints at future work but doesn't provide specific details about what areas will be explored [7].
[8] The task consists of five subtasks, with three being novel compared to previous editions. The first two subtasks involve predicting overall sentiment and sentiment towards a topic in a tweet, while the new subtasks are variations of the basic sentiment classification in Twitter. The first variant employs a five-point scale, adding an ordinal dimension to the classification. The second variant centers on accurately estimating the prevalence of each class, akin to quantification in supervised learning. The main contribution is the introduction of these new subtasks, which attracted significant participation. A limitation could be the lack of details regarding the specific challenges faced by participants or the potential drawbacks of these new subtasks [8]. [9] training a model using supervised training data from the Semeval-2015 Twitter Sentiment Analysis evaluation campaign. By comparing the approach's results with other participating systems on official test sets, the authors find that their model could potentially rank within the top two positions for both phrase-level subtask A (among 11 teams) and message-level subtask B (among 40 teams). This outcome signifies the practical efficacy of their solution. A limitation could be the absence of specific details about the model's architecture or the factors contributing to its superior performance, which might limit the replicability or understanding of their approach [9].

[10] It highlights a comparison between the outcomes of the authors' approach and those of systems partaking in an official challenge's test sets. The comparison indicates that the authors' model could potentially secure a position within the top two ranks in both the phrase-level subtask A (among 11 teams) and the message-level subtask B (among 40 teams). This result signifies the practical significance of the authors' solution. However, the paragraph does not

elaborate on the specific details of the approach, the nature of the challenge, the metrics used for ranking, or the potential reasons behind the model's high performance[10]. [11] Personality serves as a foundational driver of human behavior, impacting interactions and preferences. Personality tests are employed to determine individual traits. Social media provides a platform for self-expression, and users' posts can offer insights into their personalities. This study employs text classification to predict personality traits using Twitter text. English and Indonesian languages are considered, employing classification techniques like Naive Bayes, K-Nearest Neighbors, and Support Vector Machine. Experimental results indicate that Naive Bayes exhibited slightly better performance compared to the other methods. However, the paragraph lacks details about the specific personality traits targeted, the size and diversity of the dataset used, potential biases in the predictions, and the significance of the findings in broader contexts [11]. [12]The challenge in sentiment analysis stems from a shortage of labeled data in the realm of Natural Language Processing (NLP). To address this, sentiment analysis has incorporated deep learning techniques, which excel in automatic learning. This review paper explores recent research where deep learning models like deep neural networks and convolutional neural networks have been applied to various sentiment analysis issues. These include sentiment classification, cross-lingual challenges, textual and visual analysis, and product review analysis. However, the paragraph doesn't elaborate on the specific insights or findings from the highlighted studies, potential limitations of deep learning models, or the generalizability of the techniques across different domains [12]. [13]The paragraph discusses the utilization of data mining and analytical techniques in the context of disaster management. These techniques are employed for three main purposes: (i) prediction, (ii) detection, and (iii) the creation of effective disaster management strategies. The collected data from various disasters are used as inputs for these techniques. However, the paragraph lacks specific details about the types of data mining methods, analytical techniques, and real-world examples of their application in disaster management scenarios [13]. [14] The Konstanz Information Miner (KNIME), a robust data mining tool equipped with various features and visualization tools, was employed to analyze Twitter data. The study utilized a dataset of ten thousand Twitter entries. This research focused on sentiment analysis, essentially a classification task, using machine learning algorithms on Twitter data. The outcomes were evaluated through accuracy analysis. The integration of KNIME's extensive visualization tools is expected to enhance the ease and reliability of sentiment analysis studies, suggesting its broader adoption in the field. However, the paragraph does not provide details about the specific machine learning algorithms used, the sentiment analysis task's specific goals, or the key findings or insights obtained from the study [14].

[15] The study examines two well-known sentiment analysis models, the Netflix and Stanford models. Additionally, the researchers train their own sentiment models using escort review data obtained from the internet. The individual model performances and preliminary analysis lead to the creation of two ensemble sentiment models. These ensemble models function as feature proxies and demonstrate an ability to identify human trafficking in 53% of cases. This evaluation is based on a dataset of 38,563 ads provided by the DARPA MEMEX project. The paragraph lacks details about the specific techniques used to train the sentiment models, the characteristics of the ensemble models, and the limitations or challenges faced during the process [15]. [16] For over a decade, investment banks like Goldman Sachs, Lehman Brothers, and Salomon Brothers held sway in financial advice. However, the rise of the Internet and financial social networks like StockTwits and SeekingAlpha has empowered global investors to share experiences. While individual experts can predict stock market movements on these platforms with reasonable accuracy, understanding the collective sentiment of these expert authors towards various stocks remains a question. This paper explores whether Deep Learning models can enhance sentiment analysis performance for StockTwits. The researchers applied neural network models including long short-term memory, doc2vec, and convolutional neural networks to opinions shared on StockTwits about the stock market. The findings indicate that Deep Learning models can effectively enhance financial sentiment analysis, with the convolutional neural network emerging as the best model to predict sentiment among StockTwits authors. However, the paragraph lacks details about the dataset size, specific performance improvements, potential limitations of the approach, or the broader implications of the findings [16]. [17] Sentiment Analysis (SA) is an active research area within text mining, focusing on the computational analysis of opinions, sentiments, and subjectivity in text. This survey paper provides a comprehensive overview of recent developments in this field. It explores the advancements made in algorithms and various applications of SA. The survey categorizes articles based on their contributions to different SA techniques. However, the

paragraph does not provide specific examples of the enhancements or applications discussed, the key findings, or potential limitations addressed in the survey [17].

[18] Many organizations are increasingly interested in understanding customer opinions and tweets about their products. This enables effective planning of marketing campaigns and harnessing the positive effects of Word-of-Mouth. The study's results are thoroughly analyzed, particularly in terms of polarity and emotion classifications. This analysis demonstrates how sentiment analysis impacts organizational decision-making. Notably, in the emotions classification, "joy" is more prevalent for Microsoft Azure compared to Amazon, while "sadness" is more pronounced for Amazon than Microsoft Azure. In terms of polarity, Microsoft Azure exhibits a higher percentage of positive tweets (65%) than Amazon (45%), while Amazon has a higher negative polarity (50%) compared to Microsoft Azure (25%). These findings provide insights into the sentiment surrounding these products, aiding organizations in making informed decisions. However, the paragraph lacks context on the specific methods used for sentiment analysis, the size and diversity of the dataset, potential limitations of the analysis, and the implications of the findings for marketing strategies [18]. [19] The study delves into the utilization of Multiple Criteria Decision Analysis (MCDA) in selecting services in cloud computing (CC). It consolidates various MCDA techniques, offering a thorough evaluation of this technology for a broader audience. A taxonomy is constructed based on a literature survey, and practical aspects of MCDA implementation in CC service selection are emphasized. The study's contributions are fourfold: (a) focusing on the latest MCDA techniques, (b) offering a comparative analysis and suitability assessment of diverse MCDA methods, (c) presenting a taxonomy derived from extensive literature review, and (d) analyzing and summarizing CC service selections across different scenarios. However, the paragraph lacks specific details about the MCDA techniques discussed, the dimensions used in the taxonomy, and the practical implications or findings drawn from the analysis of cloud computing service selections [19]. [20] The study explores state-of-the-art service provisioning objectives in cloud computing. It covers crucial aspects such as essential services, topologies, user requirements, metrics, and pricing mechanisms. By conducting an in-depth literature review, the authors compile and condense various provision techniques, approaches, and models. A thematic taxonomy of cloud service provisioning is introduced based on the systematic review. Additionally, the study outlines potential paths for future research and

identifies unresolved issues within this field. However, the paragraph does not provide specifics about the types of service provisioning techniques discussed, the details of the thematic taxonomy, or the identified future research directions and open issues [20].

[21] The text describes a concise case study showcasing the effectiveness of a prototype in detecting both agitated and calm states of patients. The study also assesses the usability of wearable devices across various age groups of Alzheimer's patients, as the final assistive system is intended to be packaged as a wearable device. Furthermore, the study delves into exploring the electrodermal activity of patients aged 55-60 and 60-70 years to evaluate their health condition. However, the paragraph lacks specific details about the methodology used in the case study, the performance of the prototype, the outcomes of the usability assessment, the findings from the electrodermal activity exploration, and the potential implications of these results for the development of the assistive system [21].

3.0 Research methodology

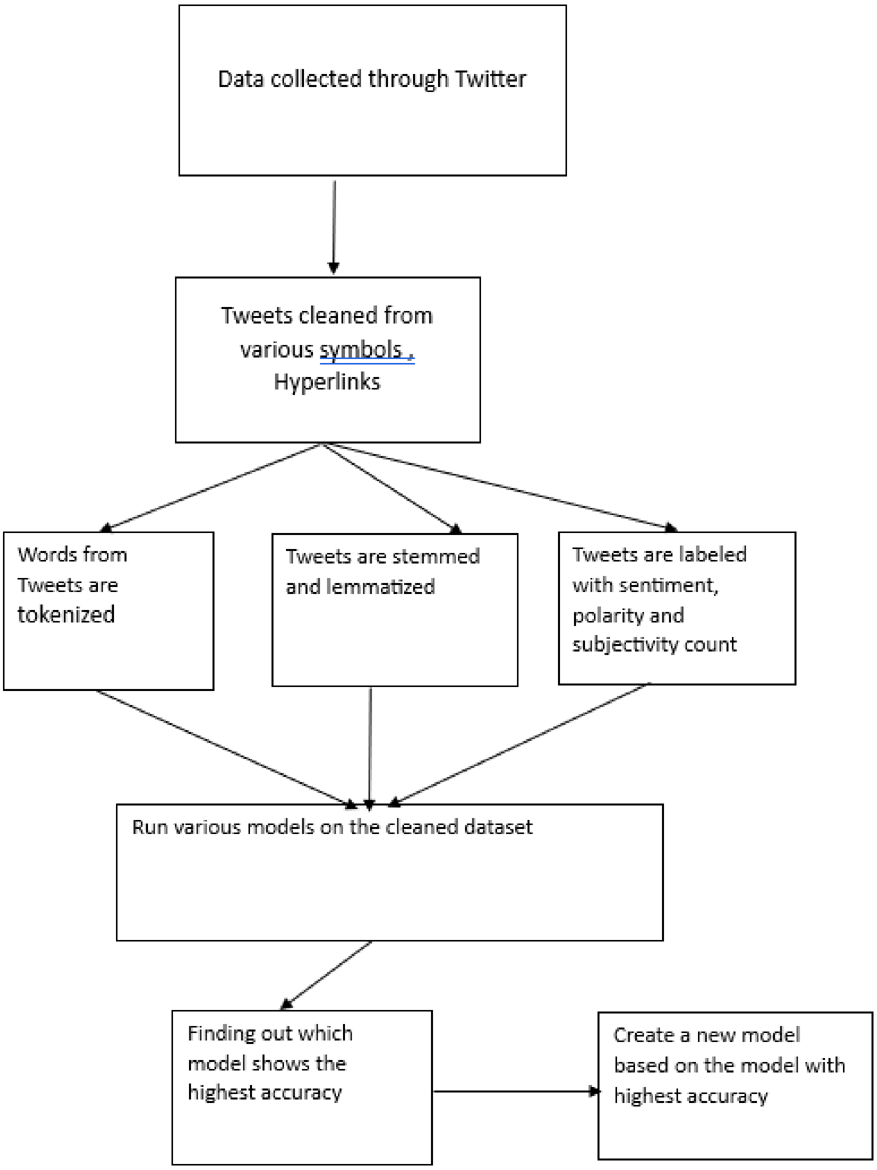

Data collected through Twitter
Tweets cleaned from various symbols , Hyperlinks
Words from Tweets are tokenized
Tweets are stemmed and lemmatized
Tweets are labeled with sentiment, polarity and subjectivity count
Run various models on the cleaned dataset
Finding out which model shows the highest accuracy
Create a new model based on the model with highest accuracy

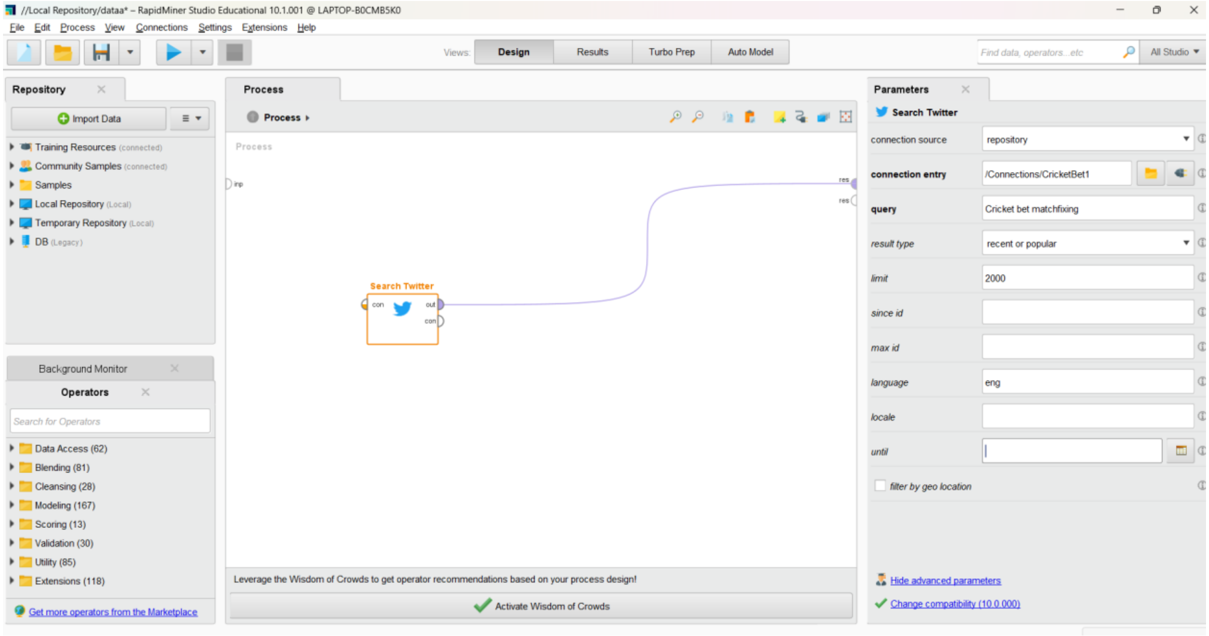
Figure 1: Workflow Relationships in Sentiment Analysis Approach

3.1 Data Collection and Preparation:

3.1.1 Extract tweets related to cricket betting from Twitter using the Twitter API or web scraping:

RapidMiner, a research-oriented instructional version, was used in the context of this study to speed up the sentiment analysis process by allowing the extraction of significant terms from cloud databases. The fundamental data was initially gathered using the text mining tools built into RapidMiner and then combined with Twitter cloud accounts before being processed using the content toolbox. The technology known as Trends Map was used to collect predicted terms useful for sentiment analysis. The website Trends Map, which serves as a database of Twitter hashtags and terms from various geographical locations, was used to incorporate widely used or well-liked words and hashtags into the analysis. This technological method made it possible to include widely used terms in the study. Trends Map provided the distinct advantage of obtaining real-time data and insights by actively tracking the frequency and application of hashtags within particular geographical domains. For instance, the examination's focus could be narrowed to particular geographic areas, such as Bangladesh with its capital city of Dhaka, or to specific states, such as Washington in the United States. Insights on the prevalent views and discussion themes among users in certain areas were gathered by careful observation of the emergence and use of hashtags within these specific locales.

By using this strategy, we were able to compile pertinent information and terms that were frequently used in the context of the sentiment analysis they were conducting. They made sure that their analysis was current and representative of the current state of the field by using the data from Trends Map trends and conversations happening on Twitter. we were able to glean valuable information and feelings from the enormous repository of Twitter data by utilizing RapidMiner's power and Trends Map's real-time insights. This combination of tools and methods allowed for a more thorough and precise analysis, illuminating the prevailing attitudes toward particular hashtags and terms within various fields of interest.

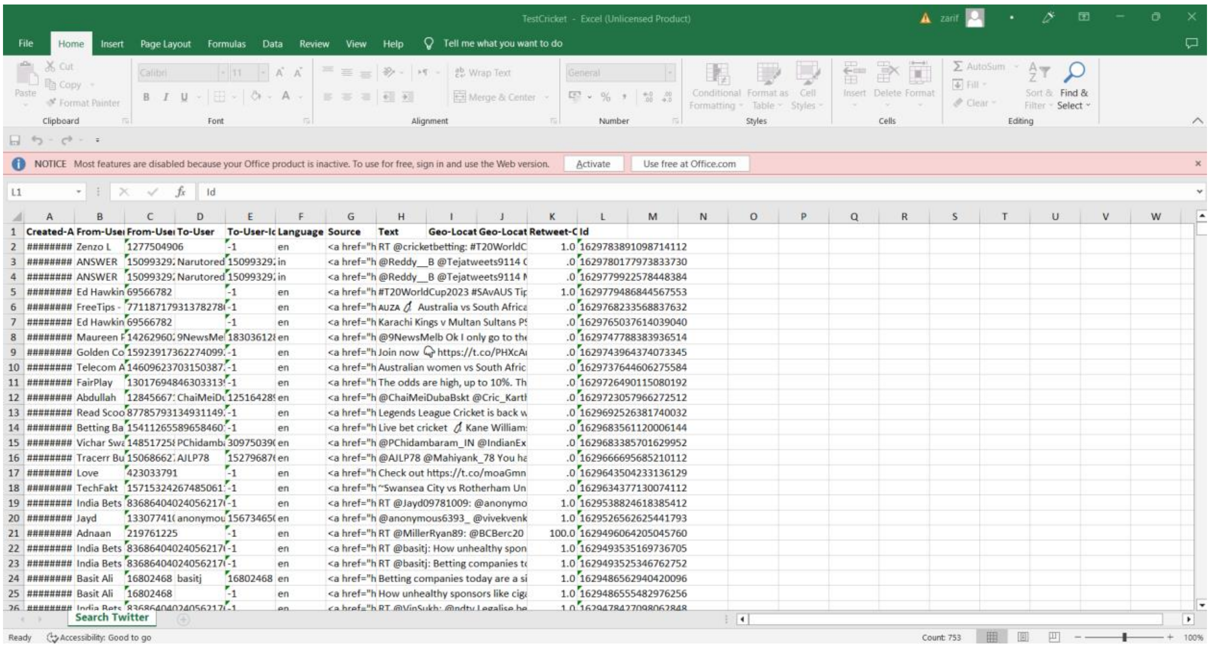

//Local Repository/dataa* – RapidMiner Studio Educational 10.1.001 @ LAPTOP-B0CMB5K0
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Community Samples (connected)
Samples
Local Repository (Local)
Temporary Repository (Local)
DB (Legacy)
Background Monitor
Operators
Search for Operators
Data Access (62)
Blending (81)
Cleansing (28)
Modeling (167)
Scoring (13)
Validation (30)
Utility (85)
Extensions (118)
Get more operators from the Marketplace
Process
Process
inp
res
res
Search Twitter
con
out
con
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Search Twitter
Hide advanced parameters
Change compatibility (10.0.000)

connection source	repository
connection entry	/Connections/CricketBet1
query	Cricket bet matchfixing
result type	recent or popular
limit	2000
since id	
max id	
language	eng
locale	
until	

□ filter by geo location

Figure 2: Searching for twitter data using Rapidminer

Overall, we were given a strong framework for doing sentiment analysis by the combination of RapidMiner and Trends Map, allowing them to delve further into the ideas, attitudes, and feelings shared by Twitter users in various geographic locations. Utilizing these techniques demonstrates our dedication to utilizing cutting-edge methodologies and technologies to improve the comprehension and interpretation of emotion in social media.



NOTICE Most features are disabled because your Office product is inactive. To use for free, sign in and use the Web version. Activate Use free at Office.com

L1 Id

	A	B	C	D	E	F	G	H	I	J	K	L
1	Created-A	From-User	From-User	To-User	To-User-Id	Language	Source	Text	Geo-Locat	Geo-Locat	Retweet-C	Id
2	########	Zenzo L	1277504906		-1	en	<a href="h	RT @cricketbetting: #T20WorldC			1.0	1629783891098714112
3	########	ANSWER	15099329	Narutored	15099329	in	<a href="h	@Reddy__B @Tejatweets9114 (			.0	1629780177973833730
4	########	ANSWER	15099329	Narutored	15099329	in	<a href="h	@Reddy__B @Tejatweets9114 N			.0	1629779922578448384
5	########	Ed Hawkin	69566782		-1	en	<a href="h	#T20WorldCup2023 #SAvAUS Tip			1.0	1629779486844567553
6	########	FreeTips -	771187179313782780		-1	en	<a href="h	AUZA 🏏 Australia vs South Africa			.0	1629768233568837632
7	########	Ed Hawkin	69566782		-1	en	<a href="h	Karachi Kings v Multan Sultans PS			.0	1629765037614039040
8	########	Maureen F	14262960	9NewsMel	18303612	en	<a href="h	@9NewsMelb Ok I only go to the			.0	1629747788383936514
9	########	Golden Co	15923917362274099		-1	en	<a href="h	Join now 👇 https://t.co/PHXcAı			.0	1629743964374073345
10	########	Telecom A	14609623703150387		-1	en	<a href="h	Australian women vs South Afric			.0	1629737644606275584
11	########	FairPlay	13017694846303313		-1	en	<a href="h	The odds are high, up to 10%. Th			.0	1629726490115080192
12	########	Abdullah	12845667	ChaiMeiDu	12516428	en	<a href="h	@ChaiMeiDubaBskt @Cric_Kartl			.0	1629723057966272512
13	########	Read Scoo	87785793134931149		-1	en	<a href="h	Legends League Cricket is back w			.0	1629692526381740032
14	########	Betting Ba	15411265589658460		-1	en	<a href="h	Live bet cricket 🏏 Kane William:			.0	1629683561120006144
15	########	Vichar Swa	14851725	PChidamba	30975039	en	<a href="h	@PChidambaram_IN @IndianEx			.0	1629683385701629952
16	########	Tracerr Bu	15068662	AJLP78	15279687	en	<a href="h	@AJLP78 @Mahiyank_78 You ha			.0	1629666695685210112
17	########	Love	423033791		-1	en	<a href="h	Check out https://t.co/moaGmn			.0	1629643504233136129
18	########	TechFakt	15715324267485061		-1	en	<a href="h	~Swansea City vs Rotherham Un			.0	1629634377130074112
19	########	India Bets	83686404024056217		-1	en	<a href="h	RT @Jayd09781009: @anonymo			1.0	1629538824618385412
20	########	Jayd	13307741	anonymou	15673465	en	<a href="h	@anonymous6393_ @vivekvenk			1.0	1629526562625441793
21	########	Adnaan	219761225		-1	en	<a href="h	RT @MillerRyan89: @BCBerc20			100.0	1629496064205045760
22	########	India Bets	83686404024056217		-1	en	<a href="h	RT @basitj: How unhealthy spon			1.0	1629493535169736705
23	########	India Bets	83686404024056217		-1	en	<a href="h	RT @basitj: Betting companies to			1.0	1629493525346762752
24	########	Basit Ali	16802468	basitj	16802468	en	<a href="h	Betting companies today are a si			1.0	1629486562940420096
25	########	Basit Ali	16802468		-1	en	<a href="h	How unhealthy sponsors like ciga			1.0	1629486555482976256
26	########	India Bets	83686404024056217		-1	en	<a href="h	RT @VinSukh: @ndtv Legalise be			1.0	1629478427098062848

Figure 3: Extracting unrefined data using rapidminer

We also carefully examine the row- and column-wise insightful connections, which helps them better comprehend how different words and hashtags relate to one another. This thorough investigation of connections results in a deeper and more complex interpretation of the sentiment data. Finally, we link the output string of the Excel box within the RapidMiner pane to the answer string. In order to ensure the correctness and dependability of the sentiment analysis results, they might extensively review and validate the output parameters in this last phase.

After collecting the data, here is our roadmap for the thesis.

3.2 Clean and preprocess the data:

The special characteristics built into the dynamic structure of the Twitter language model serve as the cornerstone of our feature reduction technique. These distinguishing characteristics serve as the cornerstone of a streamlined approach created to improve information extraction and computational efficiency. The following is a definition of the feature reduction paradigm, which is characterized by the use of Twitter's inherent qualities.

3.2.1 Utilizing Usernames: The usage of usernames, sometimes denoted by the "@" symbol (e.g., @alecmgo), is anotable linguistic phenomena seen inside Twitter. A well-known technique is applied as part of our strategic feature reduction method, which entails the establishment of an equivalence class token (USERNAME) to repeatedly replace all occurrences that start with the "@" symbol. This creative tactic drastically shrinks the feature space, accelerating following analytical steps.

3.2.2 Navigating Hyperlinks: A complex feature reduction method is required due to the frequent use of hyperlinks in tweets. An equivalence class called "URL" that includes all URLs is introduced in order to reduce this complexity. This process requires converting complex URLs, like "http://tinyurl.com/cvvg9a," into this short token. This tactical move demonstrates our dedication to maintaining data integrity throughout the reduction process while also accelerating the process

In conclusion, our method for feature reduction is supported by a cautious balance between computational effectiveness and memory preservation, meticulously aligned with Twitter's unique linguistic characteristics. We precisely shape the feature space by sparingly using equivalence class tokens for usernames and URLs together with a delicate handling of lexical duplication. This clever combination successfully navigates the linguistic nuances of sentiment analysis while capturing the subtleties of Twitter. The resulting reduced and refined analytical framework is well-suited for improved computational effectiveness and interpretability.

3.2.3 Tokenize the text into words or subword units:

The following is a thorough description in passive form of the process of text tokenization into words or subword units for conducting sentiment analysis within the Colab environment: The text tokenization process must be started within the Colab platform as the first stage. Text tokenization, a crucial pre-processing step for sentiment analysis, involves breaking down the provided text into discrete linguistic units, such as individual words or subword units, to make the analysis process easier. The text is initially segmented into its individual parts in a systematic manner while conforming to the chosen segmentation unit. Word-level tokenization divides the text into separate words, each of which stands in for a separate semantic unit. In contrast, when subword-level tokenization is used,the text is fragmented into subword elements, which may encompass partial word forms or linguistic units. The text can now be subjected to sentiment analysis thanks to the tokenization procedure. By dividing the text into manageable chunks, the subsequent analysis can successfully assess the sentiment connected to each unit, fostering a thorough grasp of the text's underlying emotional tone.

In conclusion, the passive text tokenization procedure within the Colab environment is meticulously and precisely carried out, entailing the systematic segmentation of the given text into either words or subword units. This preliminary phase is crucial in setting the stage for sentiment analysis to come, allowing for a nuanced analysis of sentiment dynamics in the text.

3.2.4 Removing stop words:

Following is a description of the thorough procedure for eliminating stopwords from the Colab environment in passive form: The stopwords elimination procedure is starting in the

Colab platform as part of the first phase. Stopwords are targeted for deletion to improve the quality and relevance of textual material for further analysis. Stopwords are frequently occurring words that lack significant semantic meaning. The text data is initially carefully examined to find any instances of stopwords. These words include commonly used ones like "the," "and," "is," and others that add little to the text's overall meaning.

Each occurrence of a stopword in the text is precisely discovered by methodical analysis. Following identification, a deliberate removal process is implemented, successfully removing the stopwords from the text.

Stopword removal is essential for preparing the text data for subsequent analysis. By eliminating these superfluous aspects, the ensuing analysis can concentrate on the important and significant parts of the text, enabling a more precise evaluation of language trends, attitudes, or other crucial insights.

To sum up, the passive activity of eliminating stopwords inside the Colab environment is carried out with meticulous attention to detail, entailing the systematic identification and subsequent elimination of frequently recurring, semantically unimportant words.

3.2.5 Remove special character from tweets:

Following is an explanation of the thorough procedure for eliminating special characters from tweets in order to prepare them for sentiment analysis in passive form:

As a first step in preparing tweets for sentiment analysis, the removal of special characters from tweets is started. Punctuation marks, symbols, and other non-alphanumeric characters are considered special characters because they don't directly affect the text's emotional tone. To find instances of special characters, the content of each tweet is thoroughly reviewed. The characters that are contained in the tweet text must be carefully examined during this process. Following their identification, special characters are removed. The removal is carried out in a manner that preserves the integrity of the linguistic content while excluding extra elements

The quality of the tweet data for sentiment analysis is improved by the elimination of special characters. By removing these non-semantic components, the analysis that follows can concentrate on the text's main emotional themes. Subsequent sentiment analysis is more accurate when special characters are not there. Because the analysis is based on the most basic linguistic components, researchers can more precisely determine the sentiment that is communicated in the tweet text. Special characters are precisely removed from tweets using a passive technique in the context of sentiment analysis. The detection, careful removal, and eventual disappearance of special characters result in a refined dataset that is better suited for sentiment analysis.

In conclusion, the passive process of deleting special characters from tweets, carried out within the framework of sentiment analysis, entails the methodical detection and eradication of non-semantic characters. This procedure improves the quality of the tweet data and makes it possible to interpret sentiment dynamics within the examined text in a way that is more precise and insightful.

3.2.6 Perform stemming or lemmatization:

Lemmatization and stemming are linguistic strategies that standardize words by reducing them to their root or base forms, improving analytical accuracy in the process. The text data is carefully examined to find terms that call for lemmatization or stemming. By simplifying variant or inflected word representations, these strategies are used to streamline word forms, enabling more efficient analysis.

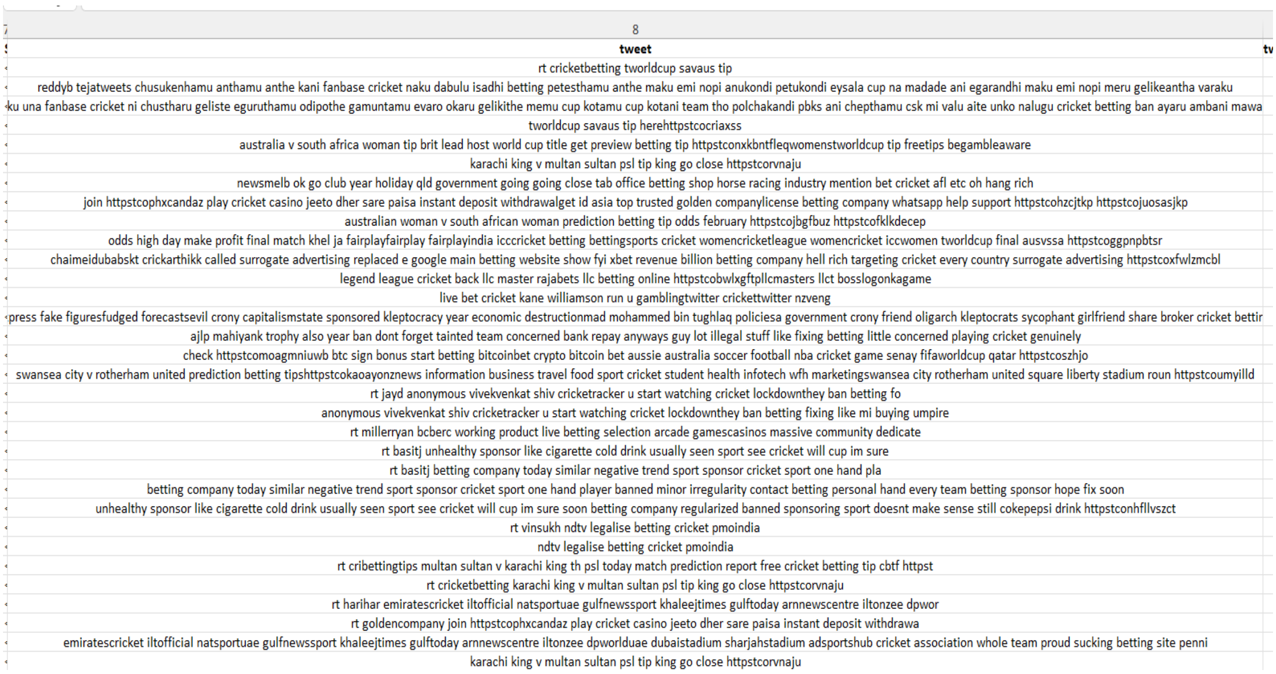
Each detected word is subjected to a number of specified linguistic rules and algorithms when stemming is applied. These guidelines govern the elimination of prefixes, suffixes, and other linguistic elements that result in the creation of a word that has been shortened or stemmed. By combining word forms, this approach makes it easier to identify similar linguistic ancestry. A more thorough linguistic analysis is conducted in the case of lemmatization. Each word is associated with the lemma that represents its canonical or dictionary form. This mapping ensures appropriate normalization by taking into account the word's grammatical context and part of speech.

Application of stemming or lemmatization has an impact on normalization. Text data is normalized as a result of these processes. By breaking down words into their most basic forms, analysis can focus on essential semantic patterns and substance rather than grammatical nuances or inflections.

3.2.7 Conclusion of the Process: The passive process of executing stemming or lemmatization is meticulously carried out within the Colab environment. Words are found, used, and then normalized to provide a refined dataset that can be subjected to thorough linguistic analysis and the extraction of illuminating patterns.

The lemmatization or stemming process passively involves the meticulous application of linguistic rules and algorithms to normalize words within the Colab context. This procedure produces a refined dataset that makes it easier to analyze the text data accurately and meaningfully, revealing important insights within the context.

After cleaning the data the tweet section looks like this



8
tweet
rt cricketbetting tworldcup savaus tip
reddyb tejatweets chusukenhamu anthamu anthe kani fanbase cricket naku dabulu isadhi betting petesthamu anthe maku emi nopi anukondi petukondi eysala cup na madade ani egarandhi maku emi nopi meru gelikeantha varaku
ku una fanbase cricket ni chustharu geliste eguruthamu odipothe gamuntamu evaro okaru gelikithe memu cup kotamu cup kotani team tho polchakandi pbks ani chepthamu csk mi valu aite unko nalugu cricket betting ban ayaru ambani mawa
tworldcup savaus tip herehttpstcocriaxss
australia v south africa woman tip brit lead host world cup title get preview betting tip httpstconxkbntfleqwomenstworldcup tip freetips begambleaware
karachi king v multan sultan psl tip king go close httpstcorvnaju
newsmelb ok go club year holiday qld government going going close tab office betting shop horse racing industry mention bet cricket afl etc oh hang rich
join httpstcophxcandaz play cricket casino jeeto dher sare paisa instant deposit withdrawalget id asia top trusted golden companylicense betting company whatsapp help support httpstcohzcjtkp httpstcojuosasjkp
australian woman v south african woman prediction betting tip odds february httpstcojbgfbuz httpstcofklkdecep
odds high day make profit final match khel ja fairplayfairplay fairplayindia icccricket betting bettingsports cricket womencricketleague womencricket iccwomen tworldcup final ausvssa httpstcoggpnpbtsr
chaimeidubabskt crickarthikk called surrogate advertising replaced e google main betting website show fyi xbet revenue billion betting company hell rich targeting cricket every country surrogate advertising httpstcoxfwlzmcbl
legend league cricket back llc master rajabets llc betting online httpstcobwlxgftpllcmasters llct bosslogonkagame
live bet cricket kane williamson run u gamblingtwitter crickettwitter nzveng
press fake figuresfudged forecastsevil crony capitalismstate sponsored kleptocracy year economic destructionmad mohammed bin tughlaq policiesa government crony friend oligarch kleptocrats sycophant girlfriend share broker cricket bettir
ajlp mahiyank trophy also year ban dont forget tainted team concerned bank repay anyways guy lot illegal stuff like fixing betting little concerned playing cricket genuinely
check httpstcomoagmniuwb btc sign bonus start betting bitcoinbet crypto bitcoin bet aussie australia soccer football nba cricket game senay fifaworldcup qatar httpstcoszhjo
swansea city v rotherham united prediction betting tipshttpstcokaoayonznews information business travel food sport cricket student health infotech wfh marketingswansea city rotherham united square liberty stadium roun httpstcoumyilld
rt jayd anonymous vivekvenkat shiv cricketracker u start watching cricket lockdownthey ban betting fo
anonymous vivekvenkat shiv cricketracker u start watching cricket lockdownthey ban betting fixing like mi buying umpire
rt millerryan bcberc working product live betting selection arcade gamescasinos massive community dedicate
rt basitj unhealthy sponsor like cigarette cold drink usually seen sport see cricket will cup im sure
rt basitj betting company today similar negative trend sport sponsor cricket sport one hand pla
betting company today similar negative trend sport sponsor cricket sport one hand player banned minor irregularity contact betting personal hand every team betting sponsor hope fix soon
unhealthy sponsor like cigarette cold drink usually seen sport see cricket will cup im sure soon betting company regularized banned sponsoring sport doesnt make sense still cokepepsi drink httpstconhfllvszct
rt vinsukh ndtv legalise betting cricket pmoindia
ndtv legalise betting cricket pmoindia
rt cribettingtips multan sultan v karachi king th psl today match prediction report free cricket betting tip cbtf httpst
rt cricketbetting karachi king v multan sultan psl tip king go close httpstcorvnaju
rt harihar emiratescricket iltofficial natsportuae gulfnewssport khaleejtimes gulftoday arnnewscentre iltonzee dpwor
rt goldencompany join httpstcophxcandaz play cricket casino jeeto dher sare paisa instant deposit withdrawa
emiratescricket iltofficial natsportuae gulfnewssport khaleejtimes gulftoday arnnewscentre iltonzee dpworlduae dubaistadium sharjahstadium adsportshub cricket association whole team proud sucking betting site penni
karachi king v multan sultan psl tip king go close httpstcorvnaju

Figure 4: Clean data after tweet

3.2.8 Sentiment analysis based on the context of the tweets:

The detailed procedure of assigning sentiment labels to the data based on the context of the tweets within the Colab environment (such as positive, negative, or neutral) is described in the passive form as follows:

3.2.8.1 Labeling Process: To assign sentiment labels to the data, the labeling process is started within the Colab platform. Using these descriptors, the tweets can be grouped according to their underlying sentiment, which can be either favorable, negative, or neutral.

3.2.8.2 Contextual Analysis: To establish the predominant sentiment, each tweet's text content is methodically examined. The goal of this research is to identify the sentiment orientation of the tweets by closely examining their linguistic phrases, emotional tones, and semantic context.

3.2.8.3 Application of Labels: Sentiment labels are carefully assigned to each tweet based on the findings of the contextual analysis. The identified sentiment is labeled as either positive, negative, or neutral, which helps to divide the tweets into various sentiment classes.

3.2.8.4 Impact of Labeling: The labeling procedure gives the data an organized category, helping further analysis and interpretation. The twitter collection can be organized well thanks to these sentiment labels, which also help researchers identify sentiment trends and patterns.

3.2.8.5 Enhancement of Analysis: The labeled data lays the foundation for a deeper sentiment analysis. To gain understanding and draw conclusions about the sentiment landscape inside the tweets, researchers can carry out additional studies, such as sentiment distribution analysis.

3.2.8.6 Finishing the Labeling Process: In the Colab environment, the passive process of meticulously applying sentiment labels to the data is carried out. A dataset that is sentimentally categorized thanks to the analysis, determination, and subsequent application of labels paves the basis for thorough sentiment analysis and interpretation.

In conclusion, assigning sentiment labels (positive, negative, or neutral) is the process of labeling the data with sentiment. This process begins with a methodical investigation of the context of tweets. This labeling procedure improves the data's organization and readability, enabling more insightful sentiment analysis to be drawn from the tweet collection.

3.2.9 polarity and subjectivity:

The Colab platform is used to begin the assessment of polarity and subjectivity, with the goal of determining the sentimental features of textual material. Subjectivity concerns the amount of personal opinion or factual content in the text, whereas polarity relates to the emotional orientation of the text, such as positive, negative, or neutral.

3.2.9.1 Systematic Analysis: The polarity and subjectivity of the text data are ascertained by a systematic analysis. To ascertain the dominant sentiment orientation, polarity analysis analyzes linguistic clues, emotional expressions, and linguistic patterns. On the other hand, subjectivity assessment comprises locating components that point to subjective ideas, opinions, or emotional content in the text.

3.2.9.2 Polarity is quantified: Based on the analysis, polarity is evaluated quantitatively. The polarity of texts is indicated by numerical values that range from strongly negative to highly positive, or even neutral. These numbers act as proxies for the text's underlying feeling.

3.2.9.3 Measurement of Subjectivity: The amount of subjectivity present in the text is also evaluated. On a scale from highly objective to highly subjective, texts are evaluated. Understanding how much of the material is influenced by individual interpretations or viewpoints is made easier with the use of this quantitative measurement.

3.2.9.4 Integration into Analysis: The assessment's results for polarity and subjectivity are carefully incorporated into the larger sentiment analysis. These metrics help characterize the sentiment landscape on a variety of levels, making it possible to understand the text data in a more complex way.

3.2.9.5 Enhanced Interpretation: An enhanced interpretation of the sentiment analysis data is made possible by the quantifiable assessments of polarity and subjectivity. Researchers learn more about the text's emotional tone and if it contains factual information or subjective perspectives.

Analysis's main finding is that the we carefully carries out the passive process of evaluating polarity and subjectivity. An extensive sentiment analysis is made possible by the analysis, quantification, and subsequent integration of polarity and subjectivity values. This results in a deeper comprehension of the sentiment patterns present in the analyzed text.

In order to measure the emotional orientation and subjective content, text data must be systematically analyzed. This is the passive method of analyzing polarity and subjectivity within the Colab environment. With the aid of these quantified values, sentiment analysis is improved, allowing for a more thorough and accurate interpretation of sentiment in the examined text collection.

3.3 Accuracy, Precision, Recall and f1 score

Before we proceed further it is important to know about accuracy, precision and recall and f1 score.

3.3.1 Accuracy:

The accuracy of sentiment analysis in the context of analysis of Twitter data is a crucial performance statistic that calls for in-depth justification.

A fundamental evaluation metric known as accuracy evaluates how accurate overall predictions made by a sentiment analysis model are in relation to the total dataset. Accuracy provides important information about the model's ability to correctly classify tweets into their associated sentiment classes in the specific context of sentiment analysis applied to Twitter data within the Colab environment.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Here, True Positives (TP) denote the proportion of tweets that were accurately classified as having a positive sentiment, while True Negatives (TN) denote the proportion of tweets that were accurately classified as having a negative emotion. The dataset's true positives, true negatives, false positives (sentiment that was wrongly classified as positive), and false negatives (sentiment that was incorrectly classed as negative) are all included in the denominator.

In conclusion, accuracy plays a crucial role in the field of sentiment analysis on Twitter data in Colab. It offers a thorough evaluation of the model's prediction power while taking into account classifications of both positive and negative sentiment. Monitoring accuracy, along with other pertinent metrics, helps to improve the efficiency of sentiment analysis models in capturing and deciphering sentiment patterns in the dynamic and expressive data.

3.3.2 Precision:

Precision in sentiment analysis is a critical performance parameter that calls for in-depth explanation in the context of Colab's analysis of Twitter data.

A sentiment analysis model's accuracy in making positive predictions in relation to the actual positive instances in the dataset is measured by precision, a fundamental evaluation metric. Precision is an important measure of the model's ability to correctly identify tweets as having a positive sentiment in the context of sentiment analysis applied to Twitter data within the Colab environment.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Here, True Positives (TP) represent the number of tweets correctly classified with their actual positive sentiment, and True Negatives (TN) represent the number of tweets correctly classified with their actual negative sentiment. The denominator encompasses all instances in the dataset, including true positives, true negatives, false positives (incorrectly classified as positive sentiment), and false negatives (incorrectly classified as negative sentiment).

3.3.3 Recall:

Recall, sometimes referred to as the true positive rate or sensitivity, measures how well a sentiment analysis model can identify every instance of a specific sentiment class (for example, positive sentiment) out of all the instances that actually fall into that class. Recall provides important information about the model's ability to record and recover pertinent sentiment instances from the dataset in the context of sentiment analysis on Twitter data.

The ratio of true positives (positive sentiment instances correctly detected) to the total of true positives and false negatives (negative sentiment instances mistakenly classified as positive) is used to calculate recall. Recall is denoted mathematically as:

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

True positives (TP) represent the number of tweets correctly classified with their actual positive sentiment, and false negatives (FN) represent the number of tweets with actual positive sentiment that were incorrectly classified as negative sentiment by the model.

3.3.4 F1 Score:

The performance of a sentiment analysis model can be fully evaluated using the F1 score, a composite statistic that strikes a compromise between precision and recall. The F1 score provides insightful information regarding the model's capability to effectively categorize feelings while taking into account both false positives and false negatives in the specific context of sentiment analysis on Twitter data using Python.

The harmonic mean of recall and precision is used to generate the F1 score:

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Here, recall is the ratio of true positives to the total of true positives and false positives, or instances of negative sentiment that were mistakenly classified as positive, and precision is the ratio of true positives to the total of true positives and false negatives, or instances of positive sentiment that were mistakenly classified as negative.

Recall and precision in balance Precision (positive predictive value) and recall (true positive rate) are balanced by the F1 score. The trade-off between reducing false positives and false negatives is captured, giving a comprehensive picture of model performance. Consideration of Misclassification: The F1 score provides a thorough assessment of the model's capability to handle various sentiment occurrences by accounting for both types of misclassification (false positives and false negatives). A higher F1 score shows that the sentiment analysis model does a good job of correctly recognizing positive sentiment instances and minimizes the misclassification of negative sentiment instances. Depending on the application and environment, one can choose to optimize the model for either greater precision or greater recall, or for the best combination of the two. When sentiment classes are skewed in the dataset, the F1 score is especially instructive. It aids in evaluating how well the model performs across various sentiment categories.

3.4 Model Evaluation and application

Now that we have successfully got the cleaned data with sentiment analysis and now we evaluate the performance of models— Logistic Regression(LR) K-Nearest Neighbors (KNN), Decision Tree , Naïve bayes and random forest —is then evaluated after the pre-processing stage

3.4.1 KNN model

The K-Nearest-Neighbors (KNN) algorithm is a non-parametric supervised classification method that excels in a variety of situations due to its simplicity. This classifier is well-known for its skillful performance and effective results, holding a significant position in the field of pattern recognition. Its simplicity makes it useful across a variety of industries, with uses in domains including pattern recognition, machine learning, text classification, data mining, and object recognition, among others. The class that boasts the smallest average distance to the unidentified pattern is accorded precedence, imparting a logical basis for decision-making. This paradigm contributes to tasks requiring categorization and prediction across a wide range of areas. The KNN technique enables accurate predictions without the need for involved model training by using the collective characteristics of nearby data points. Its adaptability, which derives from its underlying principles, has made it a fundamental tool in the toolbox of many fields.

The k-NN model determines the separation between feature vectors of data points. When comparing twitter representations, Euclidean distance or other distance metrics are used to gauge similarity. According to their proximity in the feature space, the k-nearest neighbors of each tweet are determined. The number of adjacent data points that are taken into account while reaching a conclusion depends on the analyst's choice of k.

The discovered neighbors' sentiment labels are analyzed. The emotion label that appears the most frequently among neighbors becomes the predicted sentiment label for the target tweet using the k-NN model, which uses a majority voting process. The accuracy for KNN model is 63%

3.4.2 Naïve Bayes

Naive Bayes, when applied to sentiment analysis, operates as a probabilistic machine learning algorithm that yields valuable insights into sentiment classification. Naive Bayes is used to begin the process of categorizing text data, such as tweets, into different sentiment categories based on their underlying emotional tones. To construct predictions regarding sentiment labels, Naive Bayes uses a probabilistic framework. Using conditional probabilities, it calculates the likelihood that a given text belongs to a particular sentiment class. Text data is subjected to feature extraction, which identifies and converts pertinent language components into numerical representations. common methods include bag-of-words or TF-IDF, which enable Naive Bayes to process and analyze the textual content. Naive Bayes operates under the conditional independence among features assumption, which states that each feature's presence is taken into account independently of all other features. Although it makes calculations easier, this assumption may oversimplify complex relationships found in the text. Naive Bayes determines the likelihood that each sentiment class in the text will be represented by the text. Given the sentiment class and the prior probability of the sentiment class happening, it determines the likelihood of detecting the features. Using the observed features, Naive Bayes applies the Bayes' Theorem to get the posterior probability of the sentiment class. To determine the probability distribution, multiply the prior probability by the likelihood before normalizing. Using the Bayes theorem and the premise of conditional independence, resulting in the probabilistic estimation of sentiment class probabilities based on observed features. A richer knowledge of sentiment dynamics within the investigated context is made possible by this technique, which enables reliable sentiment categorization of text data, such as tweets. The accuracy for naïve bayes is 84%

3.4.3 Random forest model

The Random Forest model, a potent and well-known machine learning algorithm, is used in sentiment analysis to categorize text into distinct sentiment categories, such as positive, negative, or neutral. It is a member of the ensemble learning family, which mixes several different decision trees to generate predictions, and has shown promise in a range of NLP tasks, including sentiment analysis. Preprocessing the text data, which includes actions like tokenization, deleting stop words, and changing text's case, is the initial step. The preprocessed text is then represented numerically using methods like word embeddings, TF-IDF,. The Random Forest model is made up of numerous distinct decision trees, each trained on various subsets of the data (bootstrapped samples) and with various subsets of features (randomly picked). These decision trees are referred to as "weak learners" because, despite their individual shortcomings, they may work together to create a powerful model. Each decision tree in the Random Forest is trained using a separate feature subset and a random subset of the dataset. To develop decision rules that as correctly separate the various sentiment classes as feasible, the decision tree continually divides the data based on the feature values during the training phase. Accuracy of random forest model is 77%.

3.4.4 Decision Tree model

To categorize text data, such as tweets, into certain sentiment categories based on their emotional context, the Decision Tree model is being integrated into sentiment analysis. Decision Tree model builds a hierarchical structure of decisions, with each internal node representing a choice made in response to a certain feature and each leaf node denoting a sentiment class. Text data is subjected to feature selection, where pertinent linguistic characteristics are found and employed as standards for the decision-making procedure inside the Decision Tree model. The Decision Tree model ranks the importance of features according to how well they help the model forecast the future. This understanding of feature importance helps determine which linguistic components have a major impact on sentiment. To assess the purity of each decision node, the Decision Tree model uses metrics like Gini Impurity or Entropy. These metrics direct the branching of the tree by giving preference to features that provide more homogeneous sentiment classes. The decision tree model goes through training, where it picks up decision rules from the training data and sentiment labels associated with them. The ingrained sentiment patterns in the text data are reflected by the learnt rules. Also, Pruning, which involves cutting off branches that don't add much to the performance of the model, is used to avoid overfitting. Generalization to fresh, untested data is improved by pruned decision trees. The accuracy of decision tree model is 78%.

After the data has been cleaned and models have been applied, the assessment will be carried out to determine which model demonstrates the highest accuracy. Subsequently, a new model will be created with the aim of enhancing the overall accuracy.

3.4.5 Logistic Regression

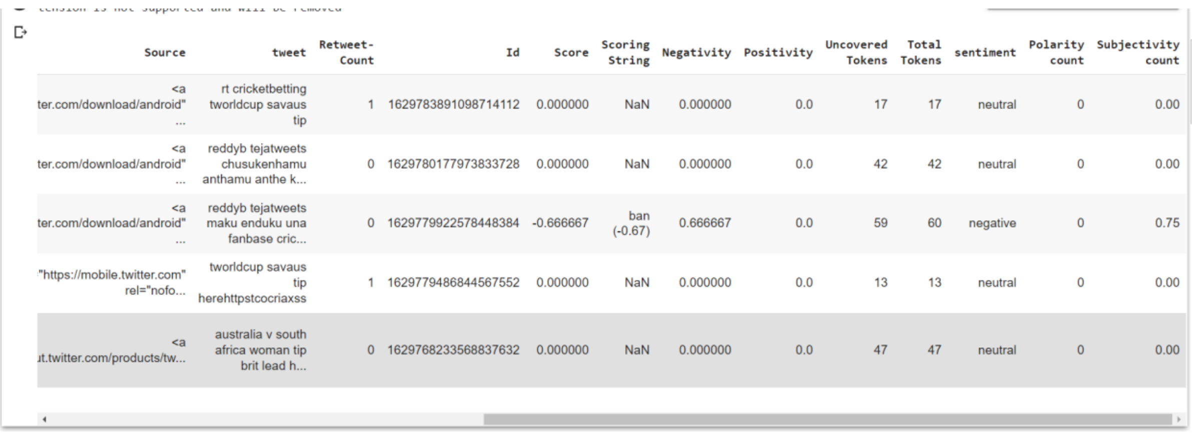
Binary classification is handled using the statistical and machine learning technique of logistic regression. In situations when the dependent variable is categorical and binary, or has just two possible outcomes, it is especially well suited. Logistic regression, despite its name, is utilized for classification rather than regression. The objective of logistic regression is to model the likelihood that a given input belongs to a specific class. The result of logistic regression is a probability score, which is modified with the use of a logistic function (also called the sigmoid function) to make sure the result is between 0 and 1. The binary classification choice is then based on this altered output. Data from Twitter is gathered, and preprocessing is done including text cleaning, tokenization, lowercase conversion, special character removal, and stopword removal. By transforming text data into numerical features using methods like TF-IDF, logistic regression may be applied to the data. To classify the data into sentiment groups like positive, negative, or neutral, sentiment labels are applied. Training data are used to build the logistic regression model, and testing data are used to assess the model's effectiveness. The data are then separated into training and testing subsets. The effectiveness of the sentiment analysis model is evaluated using measures including accuracy, precision, recall, and F1-score. To improve the model's performance in sentiment analysis, the procedure can be improved by experimentation, including changes to preprocessing, feature extraction, and hyperparameters. The accuracy of logistic regression is 83%.

3.4.6 Fine tuning a model

After running all the models on the cleaned dataset we check which model gives the highest accuracy, precision, f1 score and recall. After checking the results of the models that will be naïve bayes algorithm. We try to create a new model based on naïve bayes algorithm and will try to increase the accuracy of it as much as possible. We will use a parameter called 'alpha'. "Alpha" is a smoothing parameter used to address situations in which some feature values might not be present in the training data.

4 Results and discussions

Now we put into use of the cleaned dataset we got from preprocessing, sentiment analysing and counting subjectivity and polarity of the data.



```
tension is not supported and will be removed
```

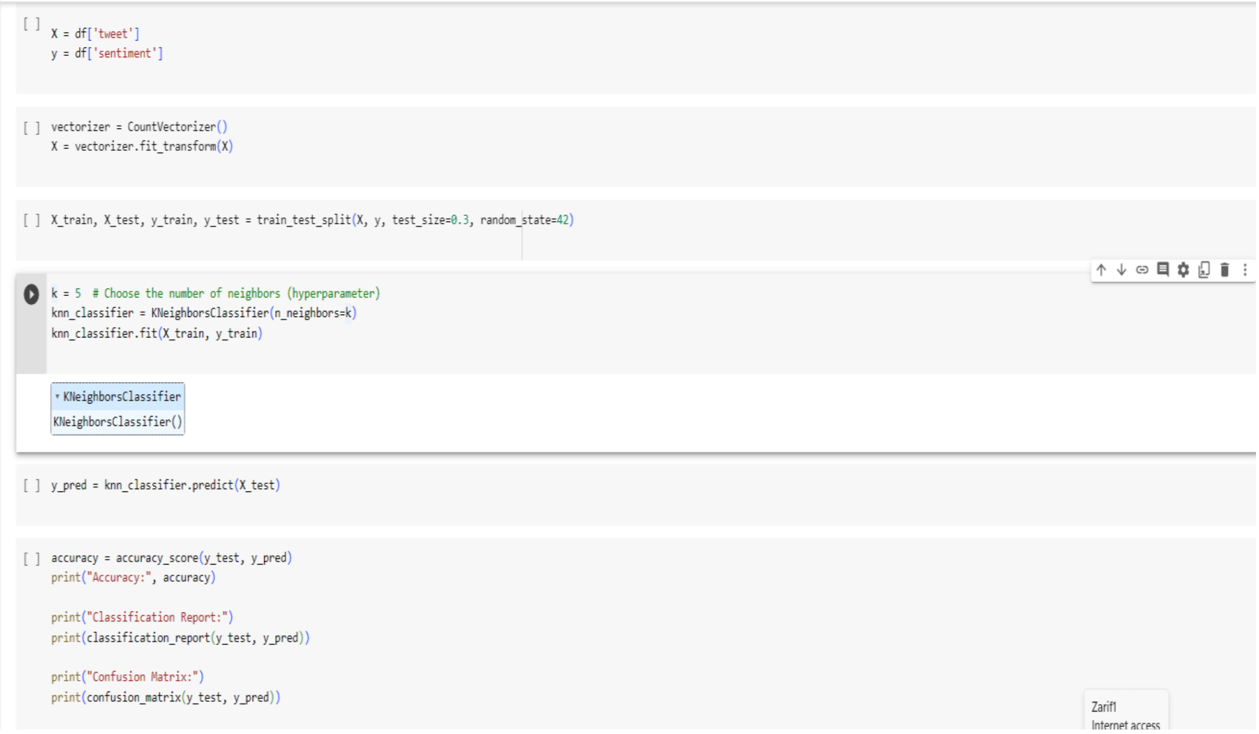
	Source	tweet	Retweet-Count	Id	Score	Scoring String	Negativity	Positivity	Uncovered Tokens	Total Tokens	sentiment	Polarity count	Subjectivity count
	<a ter.com/download/android" ...	rt cricketbetting tworldcup savaus tip	1	1629783891098714112	0.000000	NaN	0.000000	0.0	17	17	neutral	0	0.00
	<a ter.com/download/android" ...	reddyb tejatweets chusukenhamu anthamu anthe k...	0	1629780177973833728	0.000000	NaN	0.000000	0.0	42	42	neutral	0	0.00
	<a ter.com/download/android" ...	reddyb tejatweets maku enduku una fanbase cric...	0	1629779922578448384	-0.666667	ban (-0.67)	0.666667	0.0	59	60	negative	0	0.75
	"https://mobile.twitter.com" rel="nofo...	tworldcup savaus tip herehttpstcocriaxss	1	1629779486844567552	0.000000	NaN	0.000000	0.0	13	13	neutral	0	0.00
	<a ıt.twitter.com/products/tw...	australia v south africa woman tip brit lead h...	0	1629768233568837632	0.000000	NaN	0.000000	0.0	47	47	neutral	0	0.00

Figure 5: Data after pre processing and sentiment analysis

Now we utilize the models to get the accuracy, precision, recall and f1 score to determine which one has the highest percentage out of all the models.

4.1 KNN model

After applying the KNN model in the dataset



```
X = df['tweet']
y = df['sentiment']
```

```
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(X)
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

```
k = 5  # Choose the number of neighbors (hyperparameter)
knn_classifier = KNeighborsClassifier(n_neighbors=k)
knn_classifier.fit(X_train, y_train)
```

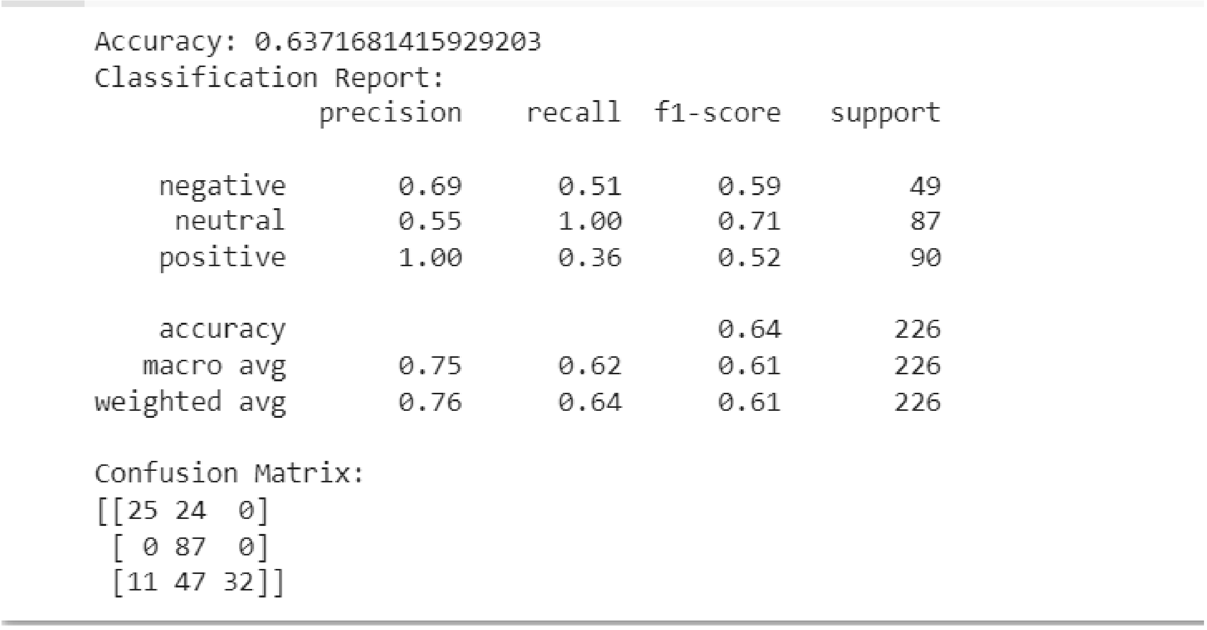
```
▾ KNeighborsClassifier
KNeighborsClassifier()
```

```
y_pred = knn_classifier.predict(X_test)
```

```
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

print("Classification Report:")
print(classification_report(y_test, y_pred))

print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))
```

Figure 6: Code for KNN model

We get the accuracy, recall, precision and f1 score of

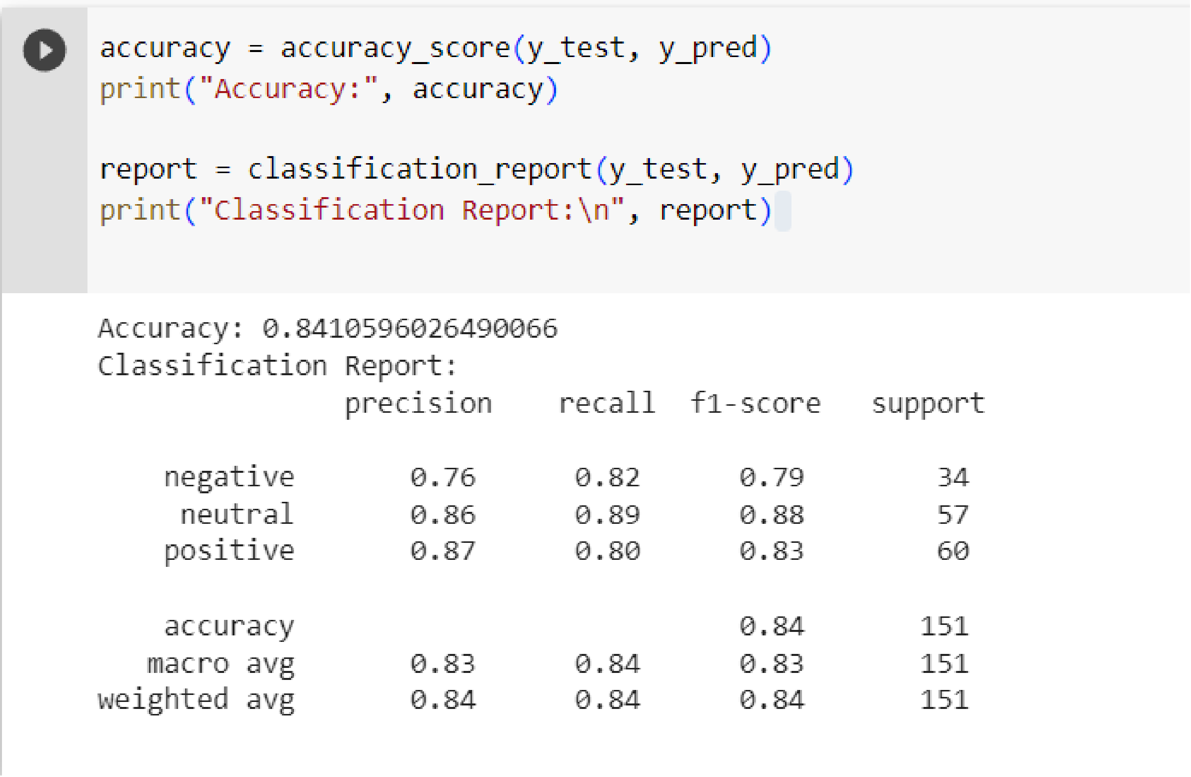


```
Accuracy: 0.6371681415929203
Classification Report:
              precision    recall  f1-score   support

    negative       0.69      0.51      0.59        49
     neutral       0.55      1.00      0.71        87
    positive       1.00      0.36      0.52        90

    accuracy                           0.64       226
   macro avg       0.75      0.62      0.61       226
weighted avg       0.76      0.64      0.61       226

Confusion Matrix:
[[25 24  0]
 [ 0 87  0]
 [11 47 32]]
```

Figure 7: Accuracy, Precision, Recall and f1 score of KNN model

4.2 Naïve Bayes

After applying the Naïve Bayes in the dataset We get the accuracy, recall, precision and f1 score of



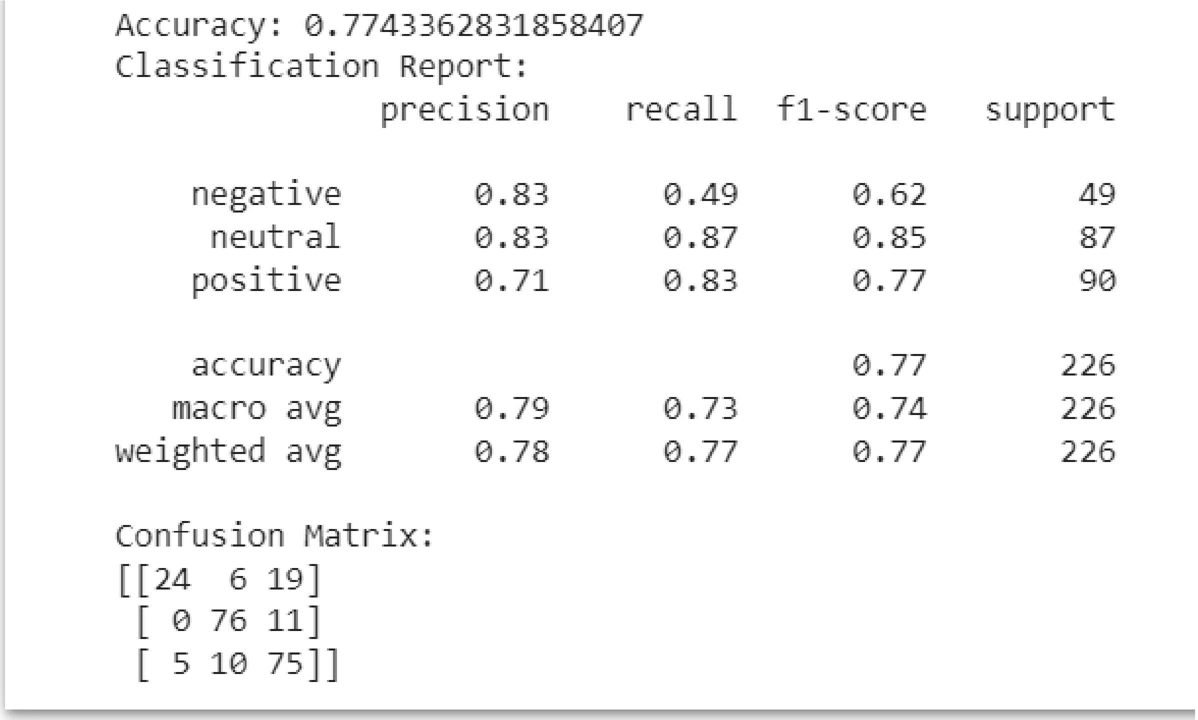
```
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

report = classification_report(y_test, y_pred)
print("Classification Report:\n", report)
```

```
Accuracy: 0.8410596026490066
Classification Report:
               precision    recall  f1-score   support

    negative       0.76      0.82      0.79        34
     neutral       0.86      0.89      0.88        57
    positive       0.87      0.80      0.83        60

    accuracy                           0.84       151
   macro avg       0.83      0.84      0.83       151
weighted avg       0.84      0.84      0.84       151
```

Figure 8: Accuracy, Precision, Recall and f1 score of Naïve Bayes

4.3 Random forest model

After applying the random forest model in the dataset We get the accuracy, recall, precision and f1 score of

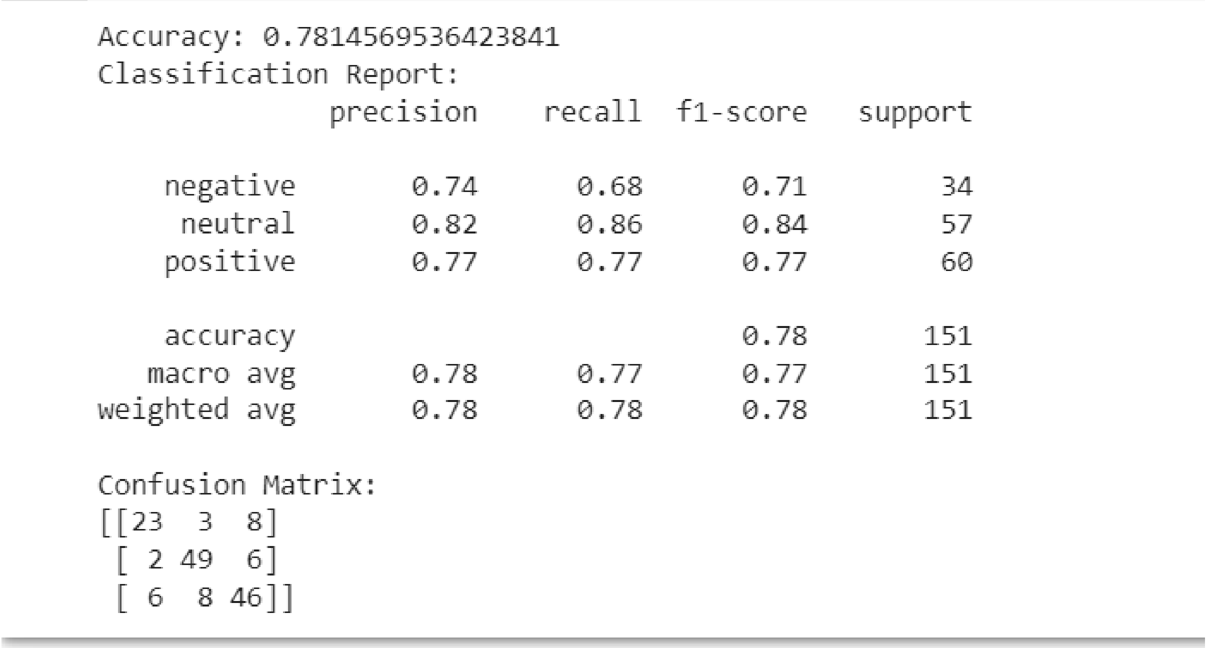


```
Accuracy: 0.7743362831858407
Classification Report:
              precision    recall  f1-score   support

    negative       0.83      0.49      0.62        49
     neutral       0.83      0.87      0.85        87
    positive       0.71      0.83      0.77        90

    accuracy                           0.77       226
   macro avg       0.79      0.73      0.74       226
weighted avg       0.78      0.77      0.77       226

Confusion Matrix:
[[24  6 19]
 [ 0 76 11]
 [ 5 10 75]]
```

Figure 9: Accuracy, Precision, Recall and f1 score of Randon forest model

4.4 Decision Tree model

After applying the Decision Tree model in the dataset We get the accuracy, recall, precision and f1 score of

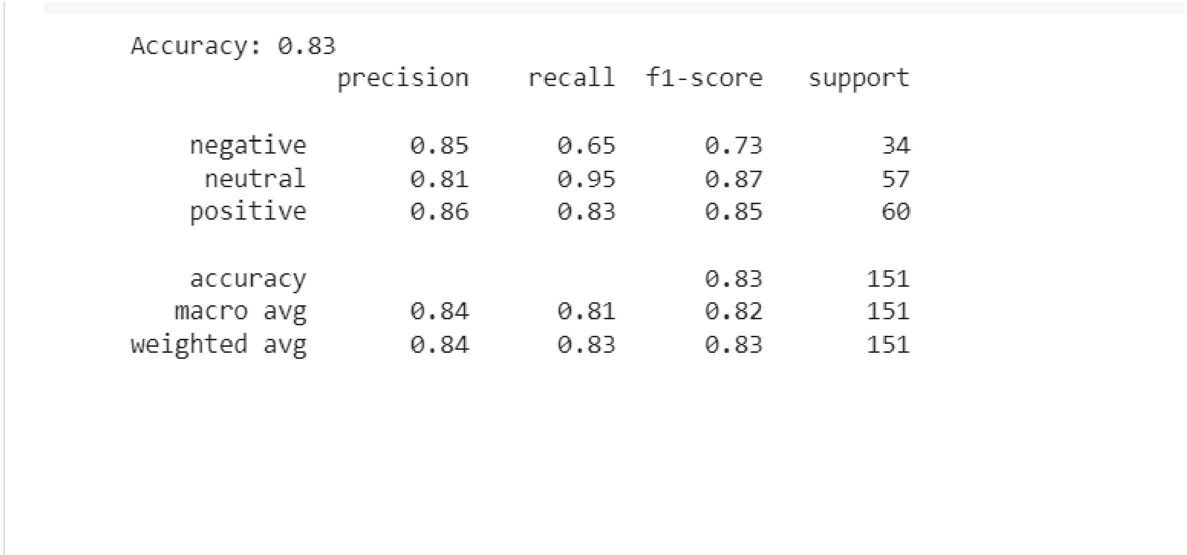


```
Accuracy: 0.7814569536423841
Classification Report:
              precision    recall  f1-score   support

    negative       0.74      0.68      0.71        34
     neutral       0.82      0.86      0.84        57
    positive       0.77      0.77      0.77        60

    accuracy                           0.78       151
   macro avg       0.78      0.77      0.77       151
weighted avg       0.78      0.78      0.78       151

Confusion Matrix:
[[23  3  8]
 [ 2 49  6]
 [ 6  8 46]]
```

Figure 10: Accuracy, Precision, Recall and f1 score of Decision Tree model

4.5 Logistic Regression

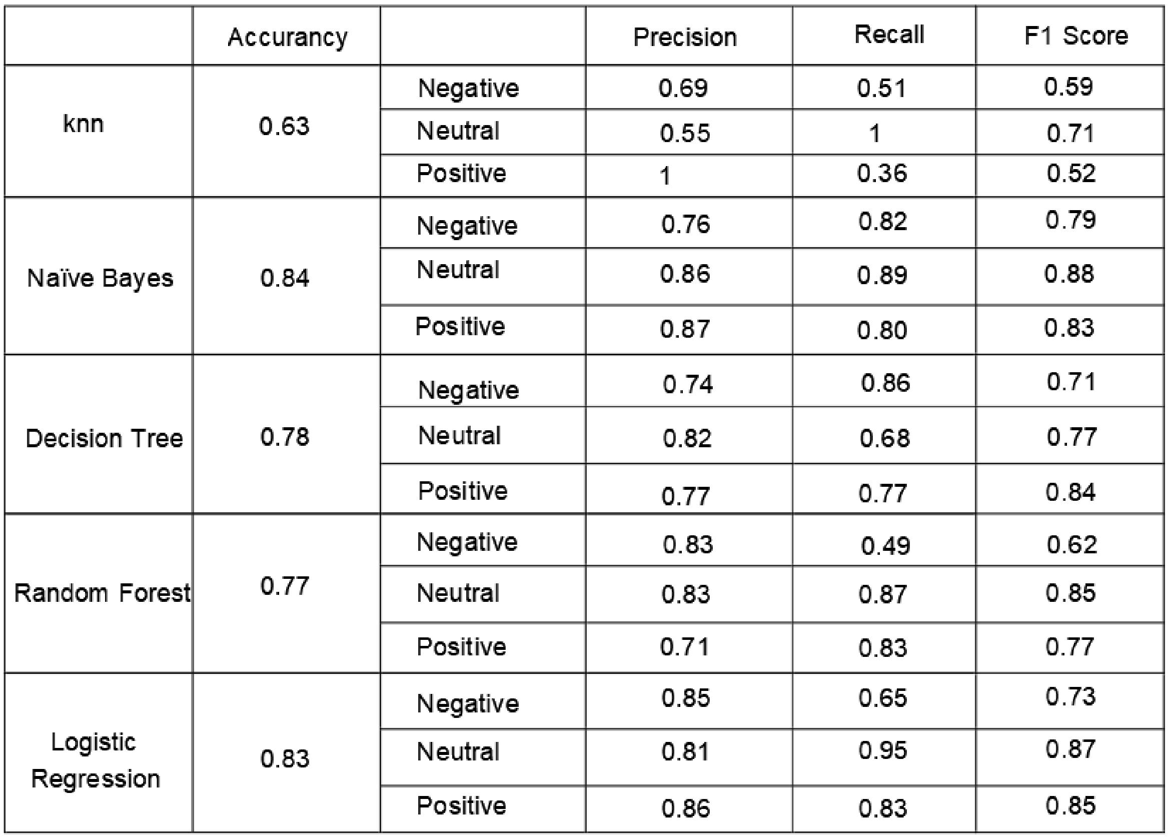
After applying the Decision Tree model in the dataset We get the accuracy, recall, precision and f1 score of



```
Accuracy: 0.83
              precision    recall  f1-score   support

    negative       0.85      0.65      0.73        34
     neutral       0.81      0.95      0.87        57
    positive       0.86      0.83      0.85        60

    accuracy                           0.83       151
   macro avg       0.84      0.81      0.82       151
weighted avg       0.84      0.83      0.83       151
```

Figure 11: Accuracy, Precision, Recall and f1 score of Logistic Regression

So, to sum it up, The result of all the models are



	Accurancy		Precision	Recall	F1 Score
knn	0.63	Negative	0.69	0.51	0.59
		Neutral	0.55	1	0.71
		Positive	1	0.36	0.52
Naïve Bayes	0.84	Negative	0.76	0.82	0.79
		Neutral	0.86	0.89	0.88
		Positive	0.87	0.80	0.83
Decision Tree	0.78	Negative	0.74	0.86	0.71
		Neutral	0.82	0.68	0.77
		Positive	0.77	0.77	0.84
Random Forest	0.77	Negative	0.83	0.49	0.62
		Neutral	0.83	0.87	0.85
		Positive	0.71	0.83	0.77
Logistic Regression	0.83	Negative	0.85	0.65	0.73
		Neutral	0.81	0.95	0.87
		Positive	0.86	0.83	0.85

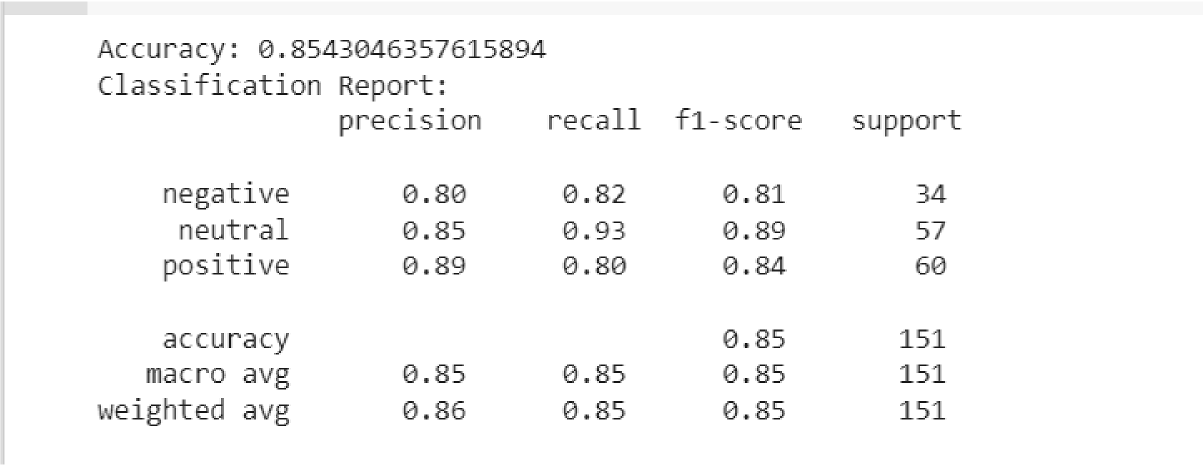
Figure 12: Accuracy, Precision, Recall and f1 score of all models

4.6 Fine tuning Naïve bayes

We can clearly see that Naïve Bayes is the model which shows the highest amount of accuracy out of all the model we tested. Now we further increase the accuracy of the model using Alpha parameter. The Multinomial Naive Bayes algorithm is used to build a classifier using a set of hyperparameters. The "alpha" value, which is the hyperparameter being set, is given a value of 0.1.

The Multinomial Naive Bayes algorithm, which mathematically determines the likelihood of a particular class given the observed attributes, is used to categorize data based on Bayes' theorem. A smoothing parameter called "alpha" is used to address situations in which some feature values might not be present in the training data. The Multinomial Naive Bayes classifier is mathematically represented by computing the conditional probabilities of features given a class, and then choosing the class with the highest probability for the classification outcome.

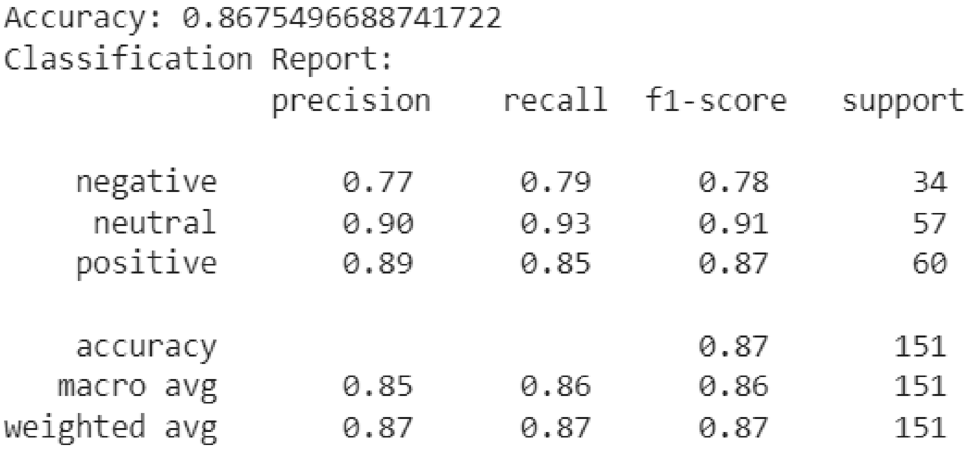
A Multinomial Naive Bayes classifier instance is initialized with the stated "alpha" value of 0.1 in the specific line of code "classifier = MultinomialNB(alpha=0.1)". This alpha value affects how much smoothing is used when calculating probabilities, which might affect how well the classifier performs and handles hidden data point. Now we the accuracy of naïve bayes algorithm increases from 84% to 85.43%.



```
Accuracy: 0.8543046357615894
Classification Report:
              precision    recall  f1-score   support

    negative       0.80      0.82      0.81        34
     neutral       0.85      0.93      0.89        57
    positive       0.89      0.80      0.84        60

    accuracy                           0.85       151
   macro avg       0.85      0.85      0.85       151
weighted avg       0.86      0.85      0.85       151
```

Figure 13: New naïve bayes model using alpha=0.1

After creating a new model we decided if we can further increase the accuracy of the new model. By lowering the value of alpha to 0.01 in the specific line of code "classifier = MultinomialNB(alpha=0.01) we can increase the accuracy.



```
Accuracy: 0.8675496688741722
Classification Report:
              precision    recall  f1-score   support

    negative       0.77      0.79      0.78        34
     neutral       0.90      0.93      0.91        57
    positive       0.89      0.85      0.87        60

    accuracy                           0.87       151
   macro avg       0.85      0.86      0.86       151
weighted avg       0.87      0.87      0.87       151
```

Figure 14: New naïve bayes model using alpha=0.01

5 Limitations and Future Work

There are more work to be done on alpha parameter to get the best possible outcome. Optimal output is necessary to get higher accuracy but the value of alpha differentiates on different datasets. We are aiming to work on that.

We also plan to use this fine tuned naïve bayes model on other social media platforms like facebook, Instagram, reddit to see how much more accuracy it shows on larger, high volume datasets.

6 Conclusion

Sentiment analysis has become a famous and effective method in the modern digital world, enabling a wide range of applications in numerous fields. This methodology sheds light on unusual and occasionally surprising events that continuously take place over time, such as instances of burglary, sneaking, human trafficking, and other noteworthy occurrences. It does this by analyzing sentiments expressed in text data. By recommending an investigation-based methodology in this study, the authors have adopted a proactive approach that will help people learn more about the inner workings of sentiment analysis approaches. In order to enable researchers, analysts, and practitioners to use this knowledge, it is important to understand the fundamental mechanisms and elements that drive sentiment analysis. Although the performance of the existing algorithmic platform has demonstrated its worth, the authors agree that there is still room for improvement and growth. As a result, they have a positive outlook on the future and picture a unique setup that includes an advanced algorithmic process. This proposed improvement could serve as a helpful aid, enhancing and raising the field of text mining approaches to new levels of precision and effectiveness.

The area of sentiment analysis is positioned to grow even more powerful and significant by embracing ongoing inquiry, innovation, and improvement, revealing previously unattainable insights and knowledge from enormous amounts of text data. The ability to interpret the complex web of human emotions and views via sentiment analysis will definitely thrive, changing how we view and engage with the digital world.

7 References

References

1. Pak, A. and Paroubek, P., 2010, May. Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).

2. Kouloumpis, E., Wilson, T. and Moore, J.D., 2011. Twitter sentiment analysis: The good the bad and the omg!. Icwsm, 11(538-541), p.164.

3. Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S., 2012, July. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of

the ACL 2012 System Demonstrations (pp. 115-120). Association for Computational Linguistics.

4. Barbosa, L. and Feng, J., 2010, August. Robust sentiment detection on twitter from biased and noisy data. In Proceedings of the 23rd international conference on computational linguistics: posters (pp. 36-44). Association for Computational Linguistics.

5. Bifet, A. and Frank, E., 2010, October. Sentiment knowledge discovery in twitter streaming

data. In International conference on discovery science (pp. 1-15). Springer, Berlin, Heidelberg.

6. Saif, H., He, Y. and Alani, H., 2012, November. Semantic sentiment analysis of twitter. In International semantic web conference (pp. 508-524). Springer, Berlin, Heidelberg.

7. Fang, X. and Zhan, J., 2015. Sentiment analysis using product review data. Journal of Big Data, 2(1), p.5.

8. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F. and Stoyanov, V., 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In Proceedings of the 10th international workshop on semantic evaluation (semeval-2016) (pp. 1-18).

9. Severyn, A. and Moschitti, A., 2015, August. Twitter sentiment analysis with deep convolutional neural networks. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 959-962). ACM.

10. Saif, H., He, Y., Fernandez, M. and Alani, H., 2016. Contextual semantics for sentiment analysis of Twitter. Information Processing & Management, 52(1), pp.5-19.

11. Pratama, B.Y. and Sarno, R., 2015, November. Personality classification based on Twitter

text using Naive Bayes, KNN and SVM. In Data and Software Engineering (ICoDSE), 2015 International Conference on (pp. 170-174). IEEE.

12. Ain, Q.T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B. and Rehman, A., 2017. Sentiment analysis using deep learning techniques: a review. Int J Adv Comput Sci Appl, 8(6), p.424.

13. Goswami, S., Chakraborty, S., Ghosh, S., Chakrabarti, A. and Chakraborty, B., 2018. A review on application of data mining techniques to combat natural disasters. Ain Shams Engineering Journal, 9(3), pp.365-378.

14. Baydogan, C. and Alatas, B., 2018, March. Sentiment analysis using Konstanz Information Miner in social networks. In Digital Forensic and Security (ISDFS), 2018 6th International Symposium on (pp. 1-5). IEEE.

15. Mensikova, A. and Mattmann, C.A., 2018. Ensemble Sentiment Analysis to Identify Human Trafficking in Web Data.

16. Sohangir, S., Wang, D., Pomeranets, A. and Khoshgoftaar, T.M., 2018. Big Data: deep learning for financial sentiment analysis. Journal of Big Data, 5(1), p.3.

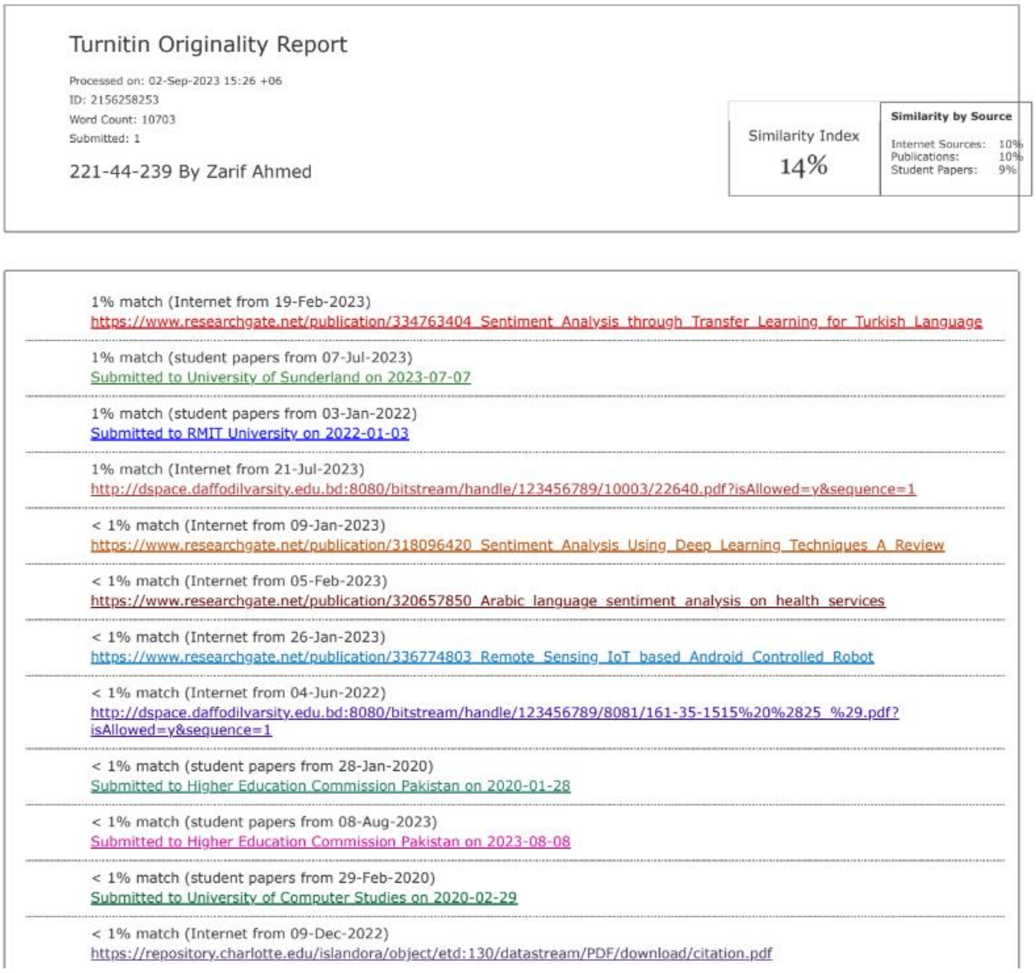
17. Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), pp.1093-1113.

18. Qaisi, L.M. and Aljarah, I., 2016, July. A twitter sentiment analysis for cloud providers: a case study of Azure vs. AWS. In Computer Science and Information Technology (CSIT), 2016 7th International Conference on (pp. 1-6). IEEE.

19. Whaiduzzaman, M., Gani, A., Anuar, N.B., Shiraz, M., Haque, M.N. and Haque, I.T., 2014. Cloud service selection using multicriteria decision analysis. The Scientific World Journal, 2014.

20. Whaiduzzaman, M., Haque, M.N., Rejaul Karim Chowdhury, M. and Gani, A., 2014. A study on strategic provisioning of cloud computing services. The Scientific World Journal, 2014.

21. Akhund, T.M.N.U., Mahi, M.J.N., Tanvir, A.H., Mahmud, M. and Kaiser, M.S., 2018, December. ADEPTNESS: Alzheimer's Disease Patient Management System Using Pervasive Sensors-Early Prototype and Preliminary Results. In International Conference on Brain Informatics (pp. 413-422). Springer, Cham.



Turnitin Originality Report

Processed on: 02-Sep-2023 15:26 +06
ID: 2156258253
Word Count: 10703
Submitted: 1

221-44-239 By Zarif Ahmed

Similarity Index
14%

Similarity by Source

Internet Sources:	10%
Publications:	10%
Student Papers:	9%

1% match (Internet from 19-Feb-2023)
https://www.researchgate.net/publication/334763404_Sentiment_Analysis_through_Transfer_Learning_for_Turkish_Language

1% match (student papers from 07-Jul-2023)
Submitted to University of Sunderland on 2023-07-07

1% match (student papers from 03-Jan-2022)
Submitted to RMIT University on 2022-01-03

1% match (Internet from 21-Jul-2023)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/10003/22640.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 09-Jan-2023)
https://www.researchgate.net/publication/318096420_Sentiment_Analysis_Using_Deep_Learning_Techniques_A_Review

< 1% match (Internet from 05-Feb-2023)
https://www.researchgate.net/publication/320657850_Arabic_language_sentiment_analysis_on_health_services

< 1% match (Internet from 26-Jan-2023)
https://www.researchgate.net/publication/336774803_Remote_Sensing_IoT_based_Android_Controlled_Robot

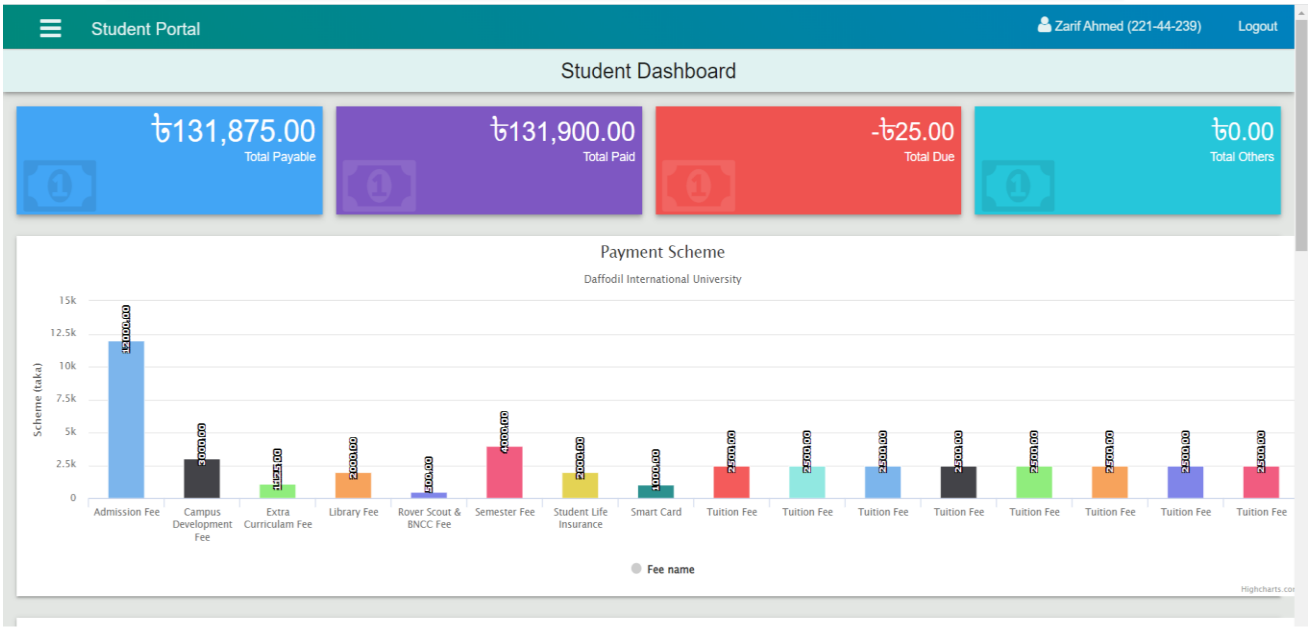
< 1% match (Internet from 04-Jun-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/8081/161-35-1515%20%2825_%29.pdf?isAllowed=y&sequence=1

< 1% match (student papers from 28-Jan-2020)
Submitted to Higher Education Commission Pakistan on 2020-01-28

< 1% match (student papers from 08-Aug-2023)
Submitted to Higher Education Commission Pakistan on 2023-08-08

< 1% match (student papers from 29-Feb-2020)
Submitted to University of Computer Studies on 2020-02-29

< 1% match (Internet from 09-Dec-2022)
https://repository.charlotte.edu/islandora/object/etd:130/datastream/PDF/download/citation.pdf



Student Dashboard

৳131,875.00
Total Payable

৳131,900.00
Total Paid

-৳25.00
Total Due

৳0.00
Total Others