



UNIVERSITI KUALA LUMPUR
ASSESSMENT BRIEF

COURSE DETAILS		
INSTITUTE	UniKL BRITISH MALAYSIAN INSTITUTE	
COURSE NAME	BIG DATA ANALYTICS	
COURSE CODE	BEB43403	
COURSE LEADER	MUHD KHAIRULZAMAN ABDUL KADIR	
LECTURER	MUHD KHAIRULZAMAN ABDUL KADIR	
SEMESTER & YEAR	OCTOBER 2023	
ASSESSMENT DETAILS		
TITLE/NAME	MINI PROJECT	
WEIGHTING	30%	
DATE/DEADLINE	2/2/2024, 5.00PM	
COURSE LEARNING OUTCOME(S)	CLO2: Prepare raw data to adjust missing values, perform normalization and make it useful for processing and effective presentation. (A4, PLO10).	
INSTRUCTIONS	Perform the following tasks: 1. Read the instructions given in the assessment sheet CAREFULLY. 2. Answer all tasks. 3. Answers must be in English.	
<hr/>		
Student Name: MUHAMMAD IKHWAN SYAFIQ BIN NORSHAM MUHAMMAD WAIZ BIN NOR KAMAL AHMAD SYAHMI BIN AHMAD FAUZI	ID: 51221221125 51221221053 51221221003	Group: L01-B02
Assessor's Comment:		Marks:
<hr/>		
Verified by: Course Leader [MUZA] Prepared by: [MUZA] I hereby declare that all my team members have agreed with this assessment. All team members are certain that this assessment complies with the Course Syllabus. <div style="text-align: center; margin-top: 20px;"> Signature: _____ Date : ____8/01/2024_____</div>	QSC format verification VERIFIED	PC/HOS content validation Dr. Nor Amalia Binti Sapiee @ Hamdan Program Coordinator Bachelor of Electronic Engineering Technology with Honours Universiti Kuala Lumpur British Malaysian Institute 9/1/2024

TASK NO	CLO	MARKING SCHEME	MARKS
1	2	Propose one simple project on object detection with working project using Python and complete framework of the project.	15
2	2	Give problem statement and scope of the project propose.	10
3	2	List the literature review related to your proposal.	10
4	2	Show methodology used with the block diagram or flowchart.	20
5	2	Show the result output of the project with the coding and comment for each line if applicable.	10
6	2	Discuss the result attain and future work can be done to improve the proposed project.	15
7	2	Conclusion	10
8	2	Show each team member task delegation in the report and slide presentation.	5
9	2	Present ideas clearly and effectively with Q&A from both parties (Team presenter & member of the floor)	5
		TOTAL	100

INFORMATION ON SK_SP-TA FOR COURSE

Course Code & Name	:	BEB43403 & BIG DATA ANALYTICS
PLOs	:	10

Please tick (☒) in the box provided.

Knowledge Profiles (SK) A programme that builds this type of knowledge and develops the attributes listed below is typically achieved in 4 years of study		
SK1	A systematic, theory-based understanding of the natural sciences applicable to the sub-discipline	
SK2	Conceptually-based mathematics, numerical analysis, statistics and aspects of computer and information science to support analysis and use of models applicable to the sub-discipline	
SK3	A systematic, theory-based formulation of engineering fundamentals required in an accepted subdiscipline	
SK4	Engineering specialist knowledge that provides theoretical frameworks and bodies of knowledge for an accepted sub-discipline	
SK5	Knowledge that supports engineering design using the technologies of a practice area	
SK6	Knowledge of engineering technologies applicable in the sub-discipline	
SK7	Comprehension of the role of technology in society and identified issues in applying engineering technology: ethics and impacts: economic, social, environmental and sustainability	
SK8	Engagement with the technological literature of the discipline	

Definition of Broadly-Defined Problem Solving (SP)			
No.	Attribute	Broadly-defined Engineering Problems have characteristic SP1 and some or all of SP2 to SP7:	
SP1	Depth of Knowledge Required	Cannot be resolved without engineering knowledge at the level of one or more of SK 4, SK5, and SK6 supported by SK3 with a strong emphasis on the application of developed technology	<input checked="" type="checkbox"/>
SP2	Range of conflicting requirements	Involve a variety of factors which may impose conflicting constraints.	
SP3	Depth of analysis required	Can be solved by application of well-proven analysis techniques	
SP4	Familiarity of issues	Belong to families of familiar problems which are solved in well-accepted ways	
SP5	Extent of applicable codes	May be partially outside those encompassed by standards or codes of practice	
SP6	Extent of stakeholder involvement and level of conflicting requirements	Involve several groups of stakeholders with differing and occasionally conflicting needs	
SP7	Interdependence	Are parts of, or systems within complex engineering problems	

Range of Engineering Activities (TA)			
No.	Attribute	Broadly-defined activities	
TA1	Range of resources	Involve a variety of resources (and for this purposes resources includes people, money, equipment, materials, information and technologies)	<input checked="" type="checkbox"/>
TA2	Level of interactions	Require resolution of occasional interactions between technical, engineering and other issues, of which few are conflicting	
TA3	Innovation	Involve the use of new materials, techniques or processes in non-standard ways	
TA4	Consequences to society and the environment	Have reasonably predictable consequences that are most important locally, but may extend more widely	
TA5	Familiarity	Require a knowledge of normal operating procedures and processes	

Data Science Mini Project

Big data refers to the vast amount of data available in various formats, including both structured and unstructured data. It presents opportunities for organizations and individuals to derive valuable insights and make informed decisions. Big data analytics plays a crucial role in unlocking the potential of this data by enabling manipulation and analysis for various purposes.

In this project, a group of experts will be formed with **4 members in a group**. Each member needs to have different tasks in performing and ensuring the successful of the project. Data used must be big enough to be considered as big data. **(Min having 4 features)**

Your detail given task as below: -

TOPICS RELATED TO BIG DATA

- 1- **Marketing and Customer Analytics:** This area focuses on leveraging big data analytics to understand customer behavior, preferences, and sentiments. It involves analyzing customer data, purchase history, online interactions, and social media data to segment customers, personalize marketing campaigns, and optimize customer experiences.
- 2- **Healthcare Analytics:** Big data analytics has significant applications in healthcare, such as analyzing electronic health records, medical imaging data, genomics data, and wearable device data. It can be used for disease prediction, patient monitoring, drug discovery, precision medicine, and healthcare resource optimization.
- 3- **Financial Analytics:** This area involves using big data analytics to analyze financial data, including stock market data, transaction records, customer behavior, and fraud detection. It can be applied for risk assessment, credit scoring, fraud prevention, algorithmic trading, and financial market analysis.
- 4- **Supply Chain and Logistics Analytics:** Big data analytics can be used to optimize supply chain operations, inventory management, and logistics. It involves analyzing data from various sources, such as production data, transportation data, and demand forecasts, to improve efficiency, reduce costs, and enhance customer satisfaction.
- 5- **Social Media Analytics:** Big data analytics is extensively used for analyzing social media data, including text data, user interactions, and network connections. It can help understand customer sentiment, identify trends, detect influencers, and measure the impact of social media marketing campaigns.
- 6- **Energy and Utilities Analytics:** This area focuses on using big data analytics to optimize energy consumption, predict power demand, and improve energy efficiency. It involves analyzing data from smart meters, sensors, weather forecasts, and energy grids to optimize resource allocation and reduce waste.
- 7- **Cybersecurity Analytics:** Big data analytics is crucial for detecting and preventing cyber threats. It involves analyzing large volumes of network traffic, log data, and security alerts to identify patterns and anomalies that indicate potential attacks or vulnerabilities.

TASK

1. Write a full report and present the project consisting of the following: -
 - **Propose one simple project related to the given topics** with working project using Python or any related software with complete framework of the project.
 - **Give** problem statement and scope of the project propose.
 - **List** the literature review related to your proposal.
 - **Show** methodology used with the block diagram or flowchart for the project. (TA2)
 - **Show** the result output of the project with the coding and comment for each line if applicable. This need to be demonstrated during presentation using Python or any related software. (SP3)
 - **Discuss** the result attain and future work can be done to improve the proposed project. (TA2)
 - **Conclusion**
 - **Show** each team member task delegation in the report and slide presentation.
2. The report **must not exceed 40 pages**.
3. Please **include the references used with complete citations** in the report.
4. **Minimum reference on the citation is 10 and preferably more than 10 citations.**
5. All works need to be presented.
6. The maximum duration of the presentation is **20 minutes**.
7. Maximum team size: **4 people**.
8. Slides must clearly show **the name of group member with the student ID**.
9. **All team members must be present** (name and student ID must be stated prior to presentation).

MINI PROJECT RUBRIC**Group Member:**

No	Criteria on mini project report and presentation	Null	Very Poor	Poor	Sufficient	Exceed Expectation	TOTAL
1.0	Propose one simple project related to the given topics with working project using Python or any related software with complete framework of the project.	0	4	7	10	15	
2.0	Give problem statement and scope of the project propose.	0	2.5	5	7.5	10	
3.0	List the literature review related to your proposal.	0	2.5	5	7.5	10	
4.0	Show methodology used with the block diagram or flowchart.	0	5	10	15	20	
5.0	Show the result output of the project with the coding and comment for each line if applicable. This need to be demonstrated during presentation using Python or any related software.	0	2.5	5	7.5	10	
6.0	Discuss the result attain and future work can be done to improve the proposed project.	0	4	7	10	15	
7.0	Conclusion	0	2.5	5	7.5	10	
8.0	Show each team member task delegation in the report and slide presentation.	1	2	3	4	5	
9.0	Present ideas clearly and effectively with Q&A from both parties (Team presenter & member of the floor)	1	2	3	4	5	
		TOTAL					

Assessor comments:

TABLE OF CONTENT

CHAPTER 1.....	3
1.1 PROPOSED PROJECT.....	3
1.2 PROBLEM STATEMENT.....	3
1.3 OBJECTIVES	3
1.4 SCOPE OF THE PROJECT	4
1.5 PHASES OF THE PROJECT	5
1.6 EVALUATION CRITERIA OF THE PROJECT	6
CHAPTER 2.....	7
2.1 LITERATURE REVIEW	7
2.1.1 Tracking the Imagined Audience: A Case Study on Nike's Use of Twitter for B2C Interaction	7
2.1.2 Text Analytics of Customers on Twitter: Brand Sentiments in Customer Support.....	8
2.1.3 Sentiment analysis on tweets. Measuring the correctness of sentiment analysis on brand attitude	10
2.1.4 Twitter Sentiment Analysis in Real-Time.....	11
2.1.5 Analyse Twitter Data with MAXQDA: Social Media Analysis ..	12
2.1.6 A Comparative Analysis of Sentiment Analysis Methods on Twitter Data.....	14
2.1.7 A Comparative Study of Sentiment Analysis Methods on Twitter Data	16
2.1.8 A Hybrid Approach for Sentiment Analysis of Twitter Data	17
2.1.9 Identifying Influential Users on Twitter	19
2.1.10 A Survey on the Use of social media for Marketing Research	20
2.2 CHAPTER SUMMARY	21
CHAPTER 3.....	25

3.1	METHODOLOGY	25
3.1.1	Block Diagram	25
3.1.2	Flowchart.....	26
CHAPTER 4	27
4.1	RESULT	27
4.1.1	Output of the Project.....	27
4.1.2	Source code	33
4.2	DISCUSSION	33
CHAPTER 5	34
5.1	CONCLUSION.....	34
5.2	FUTURE WORK.....	35
REFERENCES	36
APPENDIX	38

CHAPTER 1

1.1 PROPOSED PROJECT

The suggested project is a social media analytics tool that tracks user influence and does sentiment analysis on Twitter. This project aims to track the sentiment of tweets regarding a certain brand or product and identify the most influential individuals discussing the brand or product. The Twitter API will be used by the project to gather tweets regarding the brand or product. After the sentiment analysis of the tweets is complete, the most influential people will be determined by looking at the quantity of followers and retweets of their posts.

1.2 PROBLEM STATEMENT

In the era of rapid information dissemination through social media platforms, understanding public sentiment is crucial for businesses to maintain a positive brand image and address potential issues promptly. The purpose of this research is to construct a robust sentiment analysis model specifically tailored to recognize and categorize the emotional sentiment expressed in tweets related to the Tesla Corporation.

1.3 OBJECTIVES

The purpose of the research is to construct a sentiment analysis model that can recognise the emotional sentiment of tweets regarding the Tesla corporation. The model will be used to monitor public opinion towards Tesla and detect possible issues with its products or services.

1.4 SCOPE OF THE PROJECT

The project will include the following tasks:

- Creating a dataset of tweets about Tesla.
- Cleaning and per-processing the data.
- Perform exploratory data analysis (EDA) on the dataset.
- Training a machine learning model to recognise sentiment.
- Evaluating the model's performance.
- Implementing the model in a production setting.

Data collection

Gathering a dataset of tweets on Tesla will be the project's initial phase. The datasets must be substantial enough to adequately capture the spectrum of opinions that individuals have regarding Tesla. Either Twitter scraping or the Twitter API can be used to gather the data.

Data cleaning and per-processing

The data will then be cleaned and subjected to further processing. In order to do this, the data must be cleaned up of any mistakes or noise and formatted so that the machine learning model can use it.

Exploratory data analysis

The data will next be subjected to exploratory data analysis (EDA) following cleaning and per-processing. To understand the distribution of sentiment and other aspects, it will be necessary to explore the data.

Model training

Training a machine learning model to recognise sentiment will be the next stage. The cleaned datasets from the data collection stage will be used to train the model. The words used in the tweet, the author's emotion, and the time of day the tweet was written are just a few of the characteristics the model will utilise to determine sentiment.

Model evaluation

Evaluating the model's performance will come next once it has been trained. This will include testing the model with a dataset that is held out. A collection of tweets that were not utilised to train the algorithm will be the held-out datasets. For every emotion, we will compute the model's F1 score, accuracy, and recall.

Model deployment

The model will be deployed to a production environment when it has been assessed and found to be successful. The algorithm is going to be used to automatically extract sentiment from Tesla-related tweets.

1.5 PHASES OF THE PROJECT

The project will be divided into the following phases:

Phase 1: Data collection and labelling

The project's initial stage will entail gathering datasets of tweets regarding Tesla and annotating each one with the author's emotional response. The datasets must be substantial enough to adequately capture the spectrum of opinions that individuals have regarding Tesla. Either Twitter scraping or the Twitter API can be used to gather the data. After then, a group of human annotators will label the data.

Phase 2: Machine learning model training

In the project's second phase, a machine learning model will be trained to recognise sentiment. The Phase 1 labelled datasets will be used to train the model. The words used in the tweet, the author's emotion, and the time of day the tweet was written are just a few of the characteristics the model will utilise to determine sentiment.

Phase 3: Model deployment

The model will be deployed to a production environment during the third phase of the project. The algorithm is going to be used to automatically extract

sentiment from Tesla-related tweets. Tesla plans to utilise the model to monitor public opinion of the corporation and spot possible issues with its goods or services.

1.6 EVALUATION CRITERIA OF THE PROJECT

The project will be evaluated based on the following criteria.

Accuracy: The proportion of tweets that the machine accurately recognises as conveying a specific mood is known as accuracy. For instance, the model's accuracy for pleasure is 90% if it accurately detects 90% of the tweets that convey happiness.

Recall: The percentage of tweets that the algorithm accurately detects as expressing a specific mood is known as recall. For instance, the model's recall for pleasure is 80% if it accurately recognises 80% of the tweets that convey happiness.

F1 score: Accuracy and recall are measured by the F1 score. The harmonic mean of accuracy and recall is used to compute it. An accurate and highly recallable model is indicated by a high F1 score. By calculating the model's accuracy, recall, and F1 score using held-out datasets, the project will be assessed. A collection of tweets that were not utilised to train the algorithm will be the held-out datasets. For every feeling, the F1 score, accuracy, and recall will be determined. If the model obtains high scores for each sentiment in terms of accuracy, recall, and F1 score, the project will be deemed successful. The project team will decide on the precise success criteria.

CHAPTER 2

2.1 LITERATURE REVIEW

2.1.1 Tracking the Imagined Audience: A Case Study on Nike's Use of Twitter for B2C Interaction

Nike uses Twitter to interact with its consumers in a variety of ways. The company uses hashtags to connect with consumers who are interested in the same topics, retweets user-generated content (UGC) to show that it is listening to its consumers and responds to customer inquiries in a timely and helpful manner. These strategies help to create an "imagined audience" of consumers who are not necessarily following Nike's official account, but who are still interested in the brand and its products. The use of hashtags can help to create a sense of community around a brand. When consumers see that other people are using the same hashtags, it creates a sense of shared interest and belonging. This can make consumers more likely to interact with the brand's content, and to feel like they are part of a larger community.



Figure 1: Sample Twitter conversation. This study only collected tweets

Retweeting UGC is another way to show that a brand is listening to its consumers. When Nike retweets UGC, it shows that the company is interested in what its consumers are saying. This can help to build trust and goodwill with consumers, and it can also help to amplify UGC, which gives other consumers the opportunity to see it. Responding to customer inquiries in a timely and helpful manner is also important for

building trust and goodwill with consumers. When consumers have a problem, they want to know that the brand will be there to help them. By responding to customer inquiries in a timely and helpful manner, Nike shows that it is committed to providing excellent customer service. By using the strategies outlined in this paper, brands can create an imagined audience of consumers who are interested in the brand and its products.

2.1.2 Text Analytics of Customers on Twitter: Brand Sentiments in Customer Support

Sentiment analysis can be used to track the sentiment of customer support tweets. This can be a valuable tool for brands to understand how customers are feeling about their products and services, and to identify areas where customer service can be improved. A study of customer support tweets from ten international enterprises found that the sentiment of customer support tweets was positive. However, there were some areas where customer service could be improved. For example, customers were often frustrated by long wait times and by the lack of timely responses from customer service representatives.

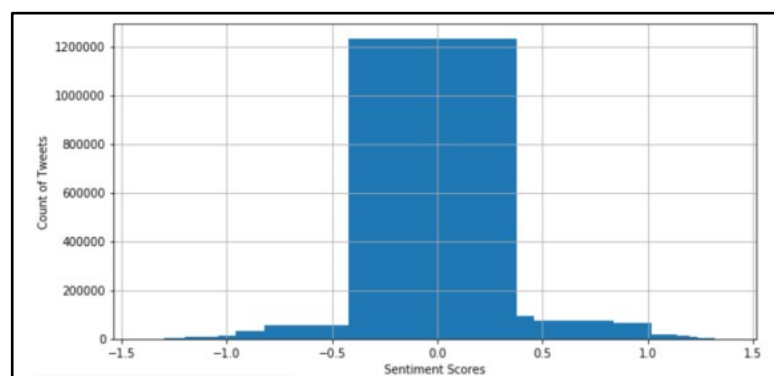


Figure 2: Distribution of Sentiment Scores

By tracking the sentiment of customer support tweets, brands can identify areas where customer service can be improved and take steps to address these issues. For example, brands can shorten wait times by hiring more customer service representatives or by using chatbots to answer simple questions. Brands can also improve the timeliness of their responses by using social media monitoring tools to track customer support tweets and respond to them in a timely manner. Sentiment analysis is a valuable tool that can help brands improve their customer service. By tracking the sentiment of customer support tweets, brands can identify areas where customer service can be improved and take steps to address these issues. This can lead to a more positive customer experience and increased customer satisfaction.

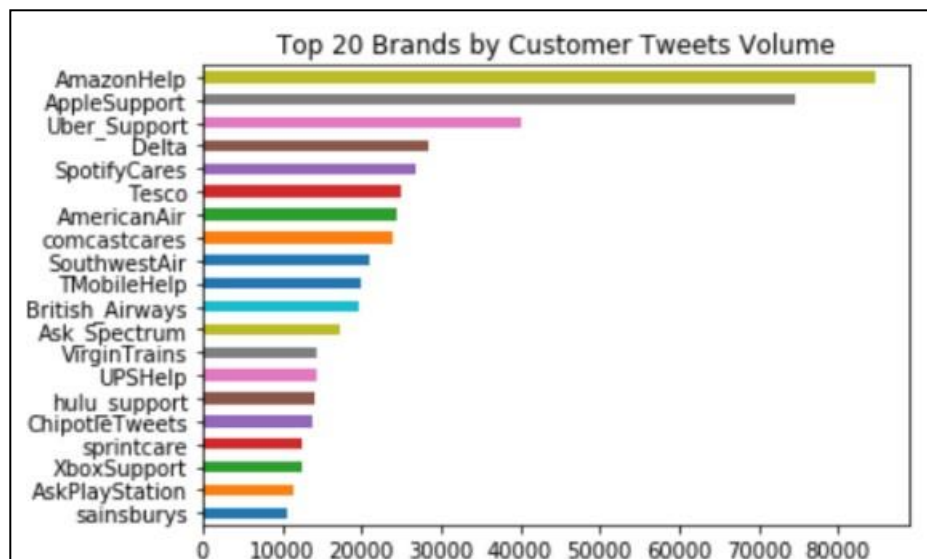


Figure 3: Customer Tweet Volumes for the top 20 brands

2.1.3 Sentiment analysis on tweets. Measuring the correctness of sentiment analysis on brand attitude

Sentiment analysis is a technique that can be used to understand how customers are feeling about a brand. This can be done by analysing the sentiment of tweets that mention the brand. The accuracy of sentiment analysis can vary depending on the method used. Lexicon-based methods are generally more accurate than machine learning methods. This is because lexicon-based methods use a list of words that are associated with positive and negative sentiment. Machine learning methods, on the other hand, are trained on a dataset of tweets that have been manually labelled as positive or negative.

When making decisions about brand strategy, it is important to carefully consider the accuracy of sentiment analysis. If the accuracy of sentiment analysis is low, then the results of the analysis may not be reliable. In this case, it may be better to use other methods to understand how customers are feeling about the brand. Overall, sentiment analysis can be a valuable tool for brands to understand how customers are feeling about their products and services. However, it is important to carefully consider the accuracy of sentiment analysis before making decisions about brand strategy.

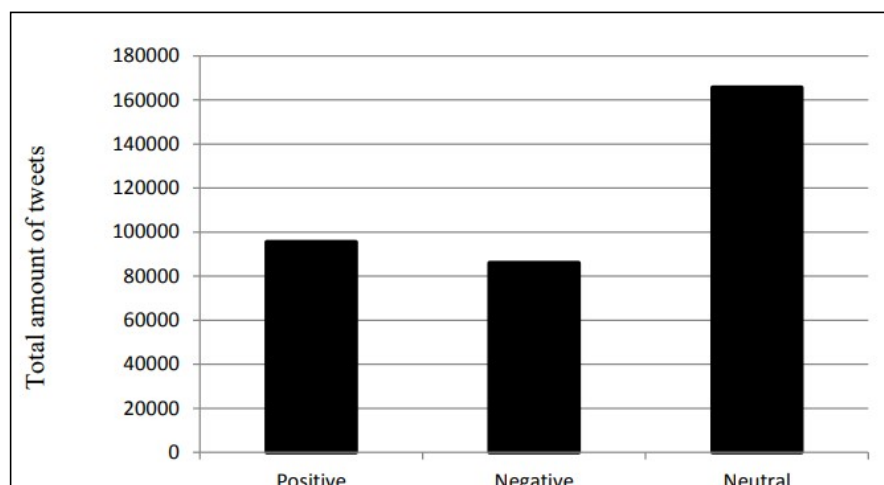


Figure 4: The distribution of number of tweets, by positive, negative, and neutral

2.1.4 Twitter Sentiment Analysis in Real-Time

Sentiment analysis is a technique that can be used to understand how people are feeling about a product, service, or brand. This can be done by analysing the sentiment of tweets that mention the product, service, or brand. Sentiment analysis can be performed in real-time, which means that brands can track the sentiment of tweets as they are being posted. This can be a valuable tool for brands to understand how customers are feeling about their products and services, and to identify potential problems before they escalate.

There are a number of challenges to performing sentiment analysis in real-time. One challenge is the high volume of data that is generated on Twitter. Another challenge is the informal nature of Twitter language. Finally, the sentiment analysis model needs to be updated regularly to keep up with changes in language. Despite the challenges, sentiment analysis can be a valuable tool for brands that want to understand how customers are feeling about their products and services. However, it is important to remember that sentiment analysis is not a perfect science, and that brands should not rely on sentiment analysis alone to make decisions about their products and services.

Here are some tips for brands that are considering using sentiment analysis:

- Use a reliable sentiment analysis tool.
- Set up alerts to notify you of negative sentiment.
- Respond to negative sentiment in a timely manner.
- Use sentiment analysis to improve your products and services.

2.1.5 Analyse Twitter Data with MAXQDA: Social Media Analysis

Twitter data can be analysed using MAXQDA, a qualitative data analysis software. MAXQDA allows users to import Twitter data, code tweets, and perform a variety of analyses. To import Twitter data into MAXQDA, users can connect their Twitter account or import tweets from a public Twitter archive. Once the tweets have been imported, users can code them using MAXQDA's coding tools. The process of social media analysis is usually divided into three stages: capture – understand – present. These phases follow one another, but the process of analysis is not linear, but can be better imagined as a cycle:

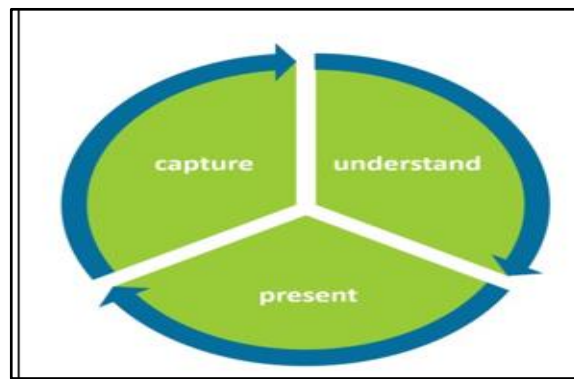


Figure 5: Stages of social media analysis

There are a variety of analyses that can be performed on Twitter data with MAXQDA, including sentiment analysis, topic modelling, and network analysis. Sentiment analysis can be used to determine the sentiment of tweets, such as whether they are positive, negative, or neutral. Topic modelling can be used to identify the topics that are being discussed in tweets. Network analysis can be used to identify the relationships between users on Twitter. MAXQDA is a powerful and versatile tool that can be used to analyse a variety of data types. It is easy to use and provides a variety of tools for coding and analysing data. MAXQDA is also compatible with a variety of other software, such as Microsoft Excel and SPSS. The benefits of using MAXQDA to analyse Twitter data include its power, versatility, ease of use, and compatibility with other software. MAXQDA has 6 steps for analysing twitter data:

- Step 1: Importing twitter data into MAXQDA
- Step 2: Auto code the imported tweets
- Step 3: Filter the data
- Step 4: Analyse sentiments for your twitter data
- Step 5: Create frequency table and charts
- Step 6: Visualize the data with word cloud

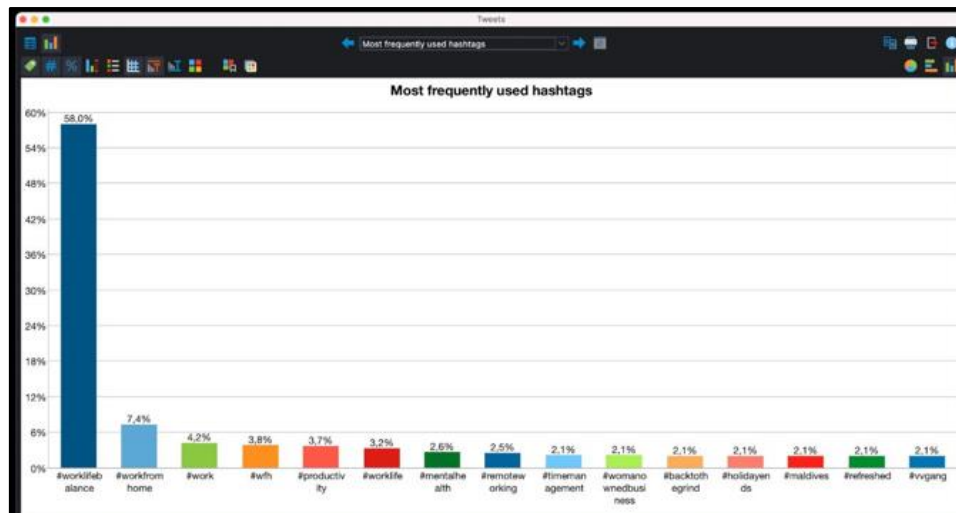


Figure 6: Create frequency tables and charts directly in MAXQDA

Type	Tweet	Retweets	Likes	Followers	Sentiment	Words	Difference
Tweet	"A great read, full of useful information." CBC Bookshelves Page: https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter . https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter	0	2	42,492	Positive	7	0
Tweet	"Be loyal to your values and priorities." I forget who said this, but it is also one of the simplest, but empowering career tips I follow. It doesn't make feel bad if I choose https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter	0	0	1,239	Slightly Positive	7	7
Tweet	"Create a Mentally Healthy Home Environment to Reduce Stress: Strategies for creating a peaceful, mentally healthy home." Very important in these times of working from home! https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter	0	0	321	Positive	15	2
Tweet	"During a year of major life changes, my corner at S&P, remained a constant source of stability for my family and me." - Heide O'Leary, Senior Recruiter at S&P, https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter	0	0	67	Positive	6	0
Tweet	"Empathetic leadership requires three things: acknowledging and overcoming any personal biases and prejudices you might have, actively listening to your people, and being action." https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter	0	0	725	Slightly Positive	7	5
Tweet	"Freizeit isn't just a German word for https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter . Rather than attempting to reconcile the two, the disconnection that comes with https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter	1	4	8,008	Neutral	3	3
Tweet	"Happy Weekend" https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter	0	0	58	Slightly Positive	1	1
Tweet	"Has working from home left you struggling to maintain a work/life balance? Here's how to switch off and reclaim your evening." https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter	0	0	1,009	Neutral	6	3
Tweet	"However, without a flux capacitor to travel back to the future, future you gets no say in the https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter https://www.addinginfo.org/2020/04/01/Heather-Wilde-on-Twitter						

Figure 7: Label your Twitter data with sentiment labels

2.1.6 A Comparative Analysis of Sentiment Analysis Methods on Twitter Data

Sentiment analysis is a challenging task, as Twitter data is often informal and noisy. There are two main approaches to sentiment analysis: lexicon-based methods and machine learning methods. Lexicon-based methods use a lexicon of words that are associated with positive and negative sentiment. These methods are simple to implement, but they can be inaccurate if the lexicon is not up to-date or if the tweets contain slang or informal language.

Machine learning methods learn to classify tweets as positive, negative, or neutral based on a training dataset of manually labelled tweets. These methods are more accurate than lexicon-based methods, but they require a larger training dataset, and they can be more computationally expensive. A study by Pak and Paroubek (2015) found that machine learning methods outperform lexicon-based methods in terms of accuracy on a dataset of Twitter data. However, the authors also found that machine learning methods are more sensitive to the size and quality of the training dataset.

The authors conclude that machine learning methods are the most accurate approach to sentiment analysis on Twitter data. However, they also recommend that lexicon-based methods be used as a first step in sentiment analysis, as they are more efficient and can be used to identify tweets that are likely to be positive or negative.

Query	Positive	negative	Neutral
Movie	53	11.1	35.8
politics	26.6	12.2	61.1
fashion	38.8	13.3	47.7
fake news	16.3	72.1	11.4
Justice	35.2	15.9	48.8
Humanity	36.9	33.3	29.7

Figure 8: sentiment analysis results

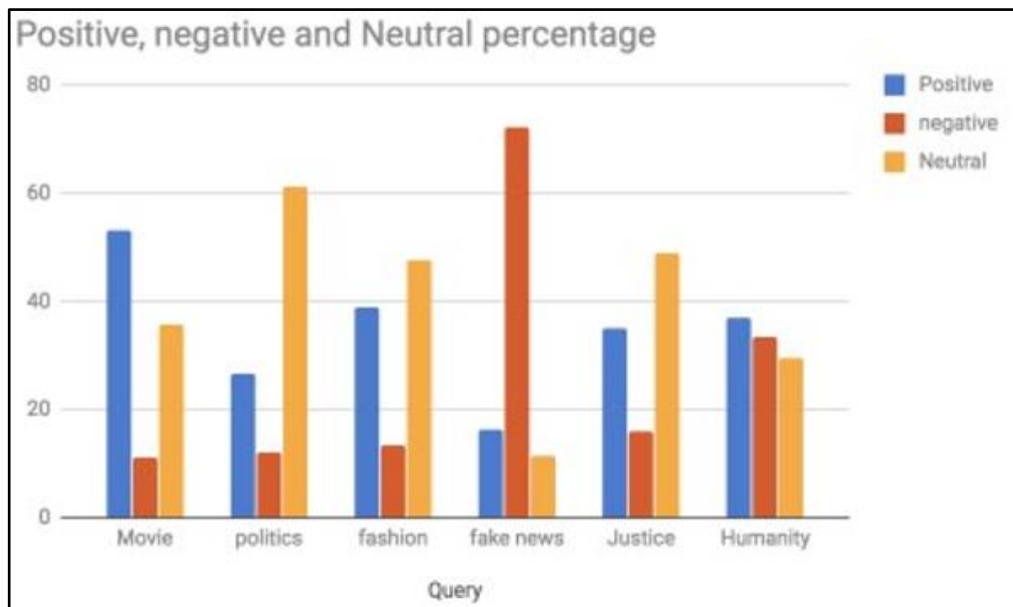


Figure 9: Sentiment results on different queries

2.1.7 A Comparative Study of Sentiment Analysis Methods on Twitter Data

The paper presents a comparative study of four different sentiment analysis methods on Twitter data: naive Bayes, support vector machines, lexicon-based methods, and distant supervision. The authors evaluated the four methods on a dataset of 1.6 million tweets and found that distant supervision performed the best, followed by support vector machines and lexicon-based methods. Naive Bayes performed the worst. The authors also found that the performance of the different methods varied depending on the sentiment of the tweets. For example, lexicon-based methods performed better for positive tweets, while distant supervision performed better for negative tweets.

The paper concludes by discussing the challenges of sentiment analysis on Twitter data and the future directions of research in this area. The authors highlight the fact that distant supervision is a promising method for sentiment analysis on Twitter data, but that there are still challenges in sentiment analysis on Twitter data, such as the noisy nature of the data and the use of slang and informal language. Overall, the paper provides a valuable overview of different sentiment analysis methods on Twitter data and their performance. The authors' findings suggest that distant supervision is a promising method for sentiment analysis on Twitter data, but that there are still challenges that need to be addressed in order to improve the accuracy of sentiment analysis on Twitter data.

2.1.8 A Hybrid Approach for Sentiment Analysis of Twitter Data

The paper "A Hybrid Approach for Sentiment Analysis of Twitter Data" proposes an innovative approach for sentiment analysis of Twitter data. The approach combines the strengths of two different methods: lexicon-based sentiment analysis and machine learning-based sentiment analysis. Lexicon based sentiment analysis uses a dictionary of words that have been manually labelled as positive, negative, or neutral. This approach is simple and fast, but it can be inaccurate, as it does not consider the context in which words are used. Machine learning-based sentiment analysis uses a machine learning model to learn the relationship between words and sentiment. This approach is more accurate than lexicon-based sentiment analysis, but it is also more complex and time-consuming. The hybrid approach proposed in the paper combines the strengths of both lexicon-based sentiment analysis and machine learning-based sentiment analysis. The approach first uses lexicon-based sentiment analysis to identify the sentiment of individual words in a tweet. Then, the machine learning model is used to learn the relationship between the sentiment of individual words and the overall sentiment of the tweet.

The paper evaluated the hybrid approach on a dataset of Twitter data. The results showed that the hybrid approach outperformed both lexicon-based sentiment analysis and machine learning-based sentiment analysis. The paper concludes by discussing the limitations of the hybrid approach and the future directions of research in sentiment analysis of Twitter data. In summary, the paper proposes an innovative approach for sentiment analysis of Twitter data that combines the strengths of two different methods. The approach was evaluated on a dataset of Twitter data and showed that it outperformed both lexicon-based sentiment analysis and machine learning-based sentiment analysis. The paper also discusses the limitations of the hybrid approach and the future directions of research in sentiment analysis of Twitter data.

Dataset	Positive	Negative	Neutral	Total
Training	2978 (37.14%)	1162 (14.49%)	3878 (48.37%)	8018
Development	483 (34.97%)	280 (20.28%)	618 (44.75%)	1381
Test	1306 (40.84%)	484 (15.13%)	1408 (44.03%)	3198

Figure 10: SemEval-2013 Task 2 Data Statistics

System	Precision	Recall	F1-score
Positive	69	52	59
Negative	42	50	46
Neutral	61	70	65

Figure 11: Proposed Hybrid Approach Result on Test Data for Each Class

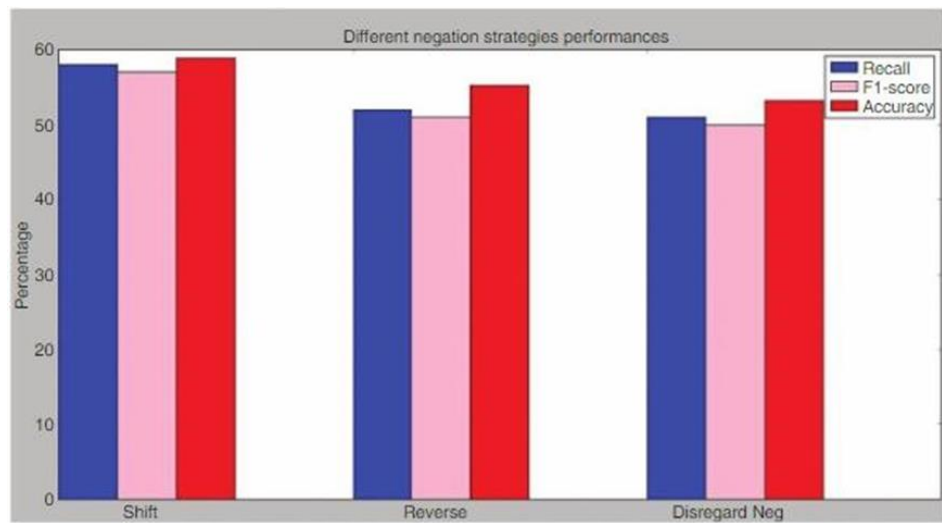


Figure 12: Comparison of different Negation Strategies.

2.1.9 Identifying Influential Users on Twitter

The paper "Identifying Influential Users on Twitter" proposes three different measures for identifying influential users on Twitter: indegree, retweets, and mentions. Indegree is the number of users who follow a given user, retweets are the number of times a user's tweets have been retweeted by other users, and mentions is the number of times a user has been mentioned in other users' tweets.

The paper evaluated the three measures on a dataset of Twitter data. The results showed that the three measures are correlated, but they measure various aspects of influence. Indegree measures the number of connections a user has, retweets measure the spread of a user's content, and mentions measures the attention a user receives. For example, a user with a high indegree may not be very influential if their tweets are not being retweeted or mentioned. On the other hand, a user with a low indegree may be very influential if their tweets are being retweeted and mentioned a lot. The paper concludes by discussing the limitations of the three measures and the future directions of research in identifying influential users on Twitter.

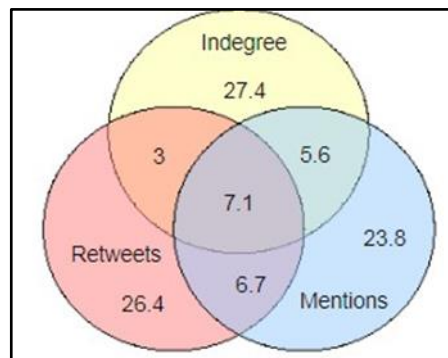


Figure 13: Venn diagram of the top-100 indumenta across measures: The chart is

Overall, the paper provides a useful overview of the diverse ways to identify influential users on Twitter. However, it is important to note that the three measures do not measure influence in the same way. Therefore, it is important to use multiple measures in order to get a more complete picture of a user's influence.

2.1.10 A Survey on the Use of social media for Marketing Research

The paper "A Survey on the Use of Social Media for Marketing Research" by Kaplan and Haenlein (2014) provides a comprehensive overview of the use of social media for marketing research. The paper begins by defining social media and discussing its potential for marketing research. The paper then reviews the separate ways that social media can be used for marketing research, including social media monitoring, social media listening, social media analytics, and social media engagement. The paper then discusses the benefits and challenges of using social media for marketing research. The benefits include the ability to reach a large and diverse audience, gain insights into consumer behaviour, and measure the effectiveness of marketing campaigns. The challenges of using social media for marketing research include data quality, data privacy, and data analysis.

The paper concludes by discussing the future of social media for marketing research. The paper argues that social media will become increasingly important for marketing research in the future, as it will allow marketers to gain deeper insights into consumer behaviour and measure the effectiveness of their marketing campaigns. Overall, the paper provides a valuable overview of the use of social media for marketing research. The paper discusses the benefits and challenges of using social media for marketing research, and it concludes by discussing the future of social media for marketing research.

		Social presence/ Media richness		
		Low	Medium	High
Self-presentation/ Self-disclosure	High	Blogs	Social networking sites (e.g., Facebook)	Virtual social worlds (e.g., Second Life)
	Low	Collaborative projects (e.g., Wikipedia)	Content communities (e.g., YouTube)	Virtual game worlds (e.g., World of Warcraft)

Figure 14: Classification of social media by social presence/media richness and self-presentation/self-disclosure

2.2 CHAPTER SUMMARY

No.	TITLE & AUTHOR	OBJECTIVES	METHODOLOGY	RESULT/ ANALYSIS
1	Tracking the Imagined Audience: A Case Study on Nike's Use of Twitter for B2C Interaction	The objective of this article is to develop an innovative methodological approach for analyzing social media content in the field of B2C communication	The study utilized Data Collection by creating a codebook with six categories to analyze each of the 192 tweets: (1) Author, (2) Forms of Interaction, (3) Content Structure, (4) Source Device, (5) Topics and Themes, and finally, (6) Unit of analysis. Using an inductive approach with the data points provided by Netlytic.	The Results from the OC corpus indicates that the conversation is dominated by users rather than the brand. Also, found that 59.4 percent of the tweets analyzed in the RT sub corpus mentioned Nike.
	Jacky Au Duong and Frauke Zeller			
2	Text Analytics of Customers on Twitter: Brand Sentiments in Customer Support	The objective of the article is to collect and pre-process a broad range of tweets for the purpose of analyzing the brand sentiments of customers while they are connected to the customer. support and after-sales services through social networks.	The research method for data mining, CRISP-DM. For the purpose of data gathering and modeling, the Python language is used. Various libraries like NumPy, Pandas, NLTK, Spacy, Matplotlib, String, and other packages are also applied for the purpose of data preparation, sentiment analysis, and visualization.	The mentioned algorithm provides a very accurate score for the overall sentence sentiment. The result of prompt replies to the customer needs has been responded by happier inbound tweets of customers at a later time.
	Iman Raeesi Vanani			
3	Sentiment analysis on tweets. Measuring the correctness of sentiment analysis on brand attitude.	The objective of the article is to conduct sentiment analysis on microblogging platform Twitter to evaluate brand attitudes and sentiments towards various products and services.	The methodology involves a combination of data collection, questionnaire-based evaluation of tweets, sentiment analysis, and statistical analysis to understand brand attitudes and sentiments expressed on Twitter.	The results and analysis underscore the significance of sentiment analysis on Twitter for understanding brand perceptions and informing marketing strategies in various product categories.
	Stephan van de Kruis, dr. M.M. van Zaanen and dr. S. Wubben			

No.	TITLE & AUTHOR	OBJECTIVES	METHODOLOGY	RESULT/ ANALYSIS
4	Twitter Sentiment Analysis in Real-Time	The objective of the article is Sentiment analysis is the automated process of identifying and classifying subjective information in text data. This might be an opinion, a judgment, or a feeling about a particular topic or product feature.	MonkeyLearn is a machine learning platform that makes it easy to build and implement sentiment analysis. You can get started right away with one of the pre-trained sentiment analysis models or you can train your own using your Twitter data.	Data visualization tools help explain sentiment analysis results in a simple and effective way. Perform sentiment analysis on your Twitter data right away and filter your results in MonkeyLearn's dashboard so you can hone in on negative or positive comments and make data-based decisions on the go.
	Monkeys learn Blog			
5	Analyze Twitter Data with MAXQDA: Social Media Analysis	The objective of the article is aims to analyze tweets related to a generally important topic that – in times where many are working from home– is as relevant as ever: work/life balance.	The methodology has several steps during each stage will, of course, depend on the methodological framework and analysis goals in your project	The researcher summarizes their evaluation and visualizes the results of their data gathering and analysis efforts during the two previous stages.
	MAXQDA Blog			

No.	TITLE & AUTHOR	OBJECTIVES	METHODOLOGY	RESULT/ ANALYSIS
6	A Comparative Analysis of Sentiment Analysis Methods on Twitter Data	This paper aims to examine the sentiments of tweets using various methods including lexicon and machine learning approaches.	There are two main approaches to sentiment analysis: lexicon-based methods and machine learning methods. Lexicon-based methods use a lexicon of words that are associated with positive and negative sentiment.	The authors conclude that machine learning methods are the most accurate approach to sentiment analysis on Twitter data. However, they also recommend that lexicon-based methods be used as a first step in sentiment analysis, as they are more efficient and can be used to identify tweets.
	Yuxing Qi and Zahratu Shabrina			
7	A Comparative Study of Sentiment Analysis Methods on Twitter Data	The objective of the study is to conduct a comparative analysis of various features used in sentiment analysis on Twitter. The authors aim to identify the most effective features for sentiment analysis	The authors evaluated the four methods on a dataset of 1.6 million tweets and found that distant supervision performed the best, followed by support vector machines and lexicon-based methods.	The results of the article indicate that AFINN lexicon and Senti-Strength method emerge as the most effective features for Twitter sentiment analysis. Through extensive experimentation with various feature sets and datasets
	Fajri Koto and Mirna Adriani			
8	A Hybrid Approach for Sentiment Analysis of Twitter Data	The objective of the article is to present a hybrid approach to Twitter sentiment analysis, combining lexicon-based methods with machine learning techniques.	The authors utilize the SemEval-2013 Twitter corpus, focusing on message-level sentiment classification into positive, negative, and neutral categories. Also, Features are extracted from the Twitter data using a hybrid approach combining lexicon-based features from Sentiment WordNet (SWN) and machine learning techniques, specifically Support Vector Machines (SVM).	the results highlight the effectiveness of the hybrid approach, particularly in handling negation and leveraging SWN-based contextual features, for improving sentiment analysis accuracy and robustness on Twitter data.
	Itisha Gupta and Nisheeth Joshi			

No.	TITLE & AUTHOR	OBJECTIVES	METHODOLOGY	RESULT/ ANALYSIS
9	Identifying Influential Users on Twitter	The objective of the article is to analyze the influence of users on the social media platform Twitter. It employs three measures of influence: indegree, retweets, and mentions, to capture different perspectives of influence.	The methodology involves analyzing data from Twitter to study user influence. The study examines how these measures correlate and vary across different topics and time periods	These findings provide insights into how influence operates on Twitter, suggesting that targeting top influentials is more effective than relying on a large number of less popular users for spreading information.
	Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto and Krishna P. Gummadi			
10	A Survey on the Use of social media for Marketing Research	The objective of the article is to explore the challenges and opportunities presented by social media for businesses. It provides an in-depth analysis of various forms of social media, including blogs, content communities, social networking sites.	The article does not explicitly outline a methodology section. The content is structured around different types of social media platforms and their characteristics	The result of the article is a comprehensive exploration of the challenges and opportunities presented by social media for businesses. It discusses various types of social media platforms, including blogs, content communities, social networking sites, virtual game worlds, and virtual social worlds.
	Andreas Kaplan and Michael Haenlein			

CHAPTER 3

3.1 METHODOLOGY

3.1.1 Block Diagram

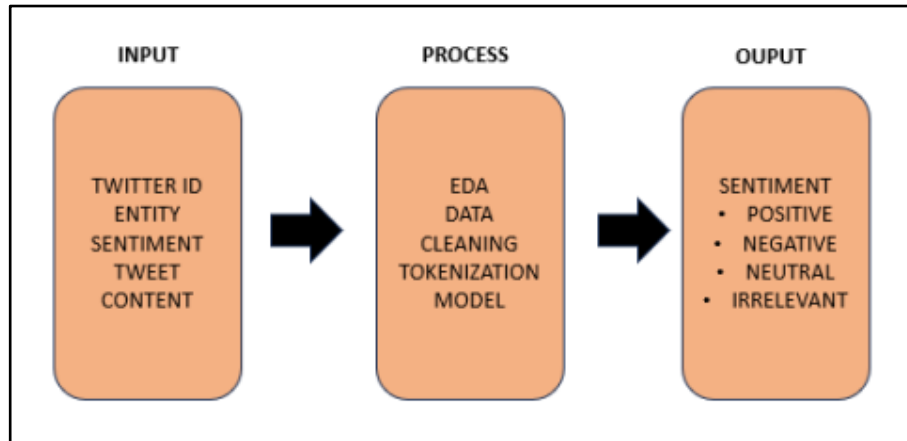


Figure 15: Block diagram of the project

Figure 15 shows the block diagram for the proposed project. The project starts by taking inputs such as Twitter ID, entity, sentiment, and tweet content. The first step is exploratory data analysis (EDA), where the data is thoroughly examined to gain insights and understand its characteristics. This involves tasks like data cleaning, which includes removing irrelevant characters, managing punctuation, and dealing with special characters or emojis. The next step is tokenization, where the tweet content is split into individual words or tokens. Afterward, a model is built using machine learning techniques to classify the sentiment of the tweet. The model is trained on labelled data, leveraging various algorithms such as Naive Bayes, SVM, or deep learning models like BERT. Finally, the output of the model provides the sentiment classification, which can be categorized into four categories: positive, negative, neutral, or irrelevant. This pipeline enables the analysis of Twitter data, extracting sentiment information, and categorizing tweets based on their emotional tone.

3.1.2 Flowchart

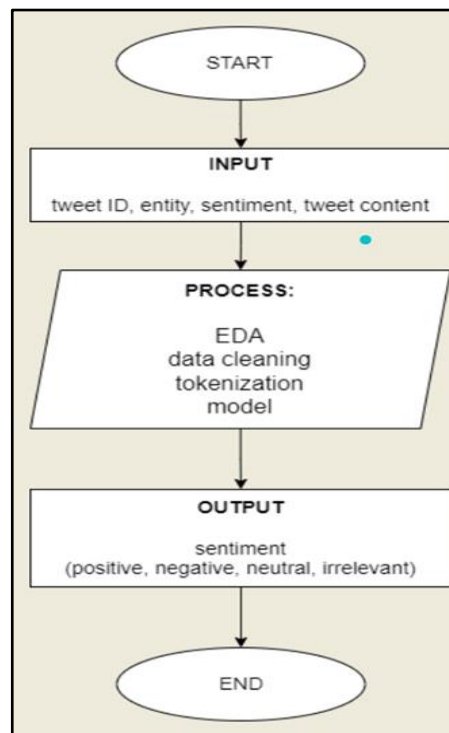


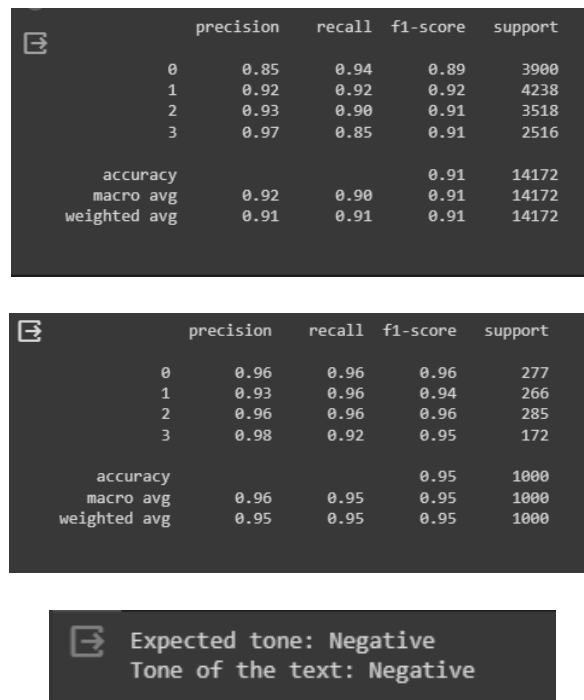
Figure 16: Flowchart of the project

From the figure 16 it shows the flowchart, The flowchart would begin with the input of Twitter ID, entity, sentiment, and tweet content. The first step would involve exploratory data analysis (EDA), where the data is analysed to gain insights. This would be followed by data cleaning, which includes removing irrelevant characters, managing punctuation, and dealing with special characters or emojis. The next step is tokenization, where the tweet content is split into individual words or tokens. Afterward, a machine learning model is trained using the pre-processed data. The model would utilize algorithms such as Naive Bayes, SVM, or deep learning models like BERT. Finally, the output of the model would provide the sentiment classification, which can be categorized as positive, negative, neutral, or irrelevant. The flowchart would help visualize the sequential steps involved in the sentiment analysis process, guiding the implementation of the software solution.

CHAPTER 4

4.1 RESULT

4.1.1 Output of the Project



The figure consists of three screenshots from a terminal or application window. The first screenshot shows a confusion matrix and summary metrics for a sentiment analysis model. The second screenshot shows similar metrics for a different dataset or configuration. The third screenshot shows the model's output for a specific text input.

	precision	recall	f1-score	support
0	0.85	0.94	0.89	3900
1	0.92	0.92	0.92	4238
2	0.93	0.90	0.91	3518
3	0.97	0.85	0.91	2516
accuracy			0.91	14172
macro avg	0.92	0.90	0.91	14172
weighted avg	0.91	0.91	0.91	14172

	precision	recall	f1-score	support
0	0.96	0.96	0.96	277
1	0.93	0.96	0.94	266
2	0.96	0.96	0.96	285
3	0.98	0.92	0.95	172
accuracy			0.95	1000
macro avg	0.96	0.95	0.95	1000
weighted avg	0.95	0.95	0.95	1000

Expected tone: Negative
Tone of the text: Negative

Figure 17: Accuracy model trained

Figure 17 shows that the accuracy of a trained sentiment analysis model is an important evaluation metric that measures its performance in correctly classifying the sentiment of text data. High accuracy indicates that the model effectively predicts the sentiment labels of the input texts, whether positive, negative, neutral or irrelevant. To achieve high accuracy, the model must be trained on a diverse and representative data set, thorough pre-processing of the data must be performed and suitable algorithms for machine learning or deep learning architectures must be selected. In addition, the accuracy of the model can be further improved by optimising hyperparameters, performing cross-validation and using techniques such as regularisation or ensemble methods. Regularly monitoring the accuracy of the model on a separate validation or test data set is crucial to ensure its effectiveness and generalisability. By continuously refining the model based on accuracy metrics,

it can be fine-tuned to provide more reliable sentiment predictions, making it a valuable tool for sentiment analysis tasks(positive, negative, neutral or irrelevant). The sentiment labels serve as the target variable or the output that the model should predict.

To train the models, they are provided with the input features derived from the text data. These features can be created using techniques such as bag-of-words, TF-IDF, word embeddings or extended contextual embeddings. The models analyse these features and learn to recognise patterns and relationships between the input features and the corresponding sentiment labels. During the training process, the internal parameters of the model are iteratively adjusted to minimise the error or difference between the predicted sentiment labels and the actual sentiment labels in the labelled dataset. This is usually achieved through optimisation algorithms such as gradient descent, which update the parameters of the model based on the calculated error or loss. Once the models have been trained, they are evaluated using various performance metrics to assess their effectiveness in predicting sentiment. Accuracy is a common metric that measures the overall correctness of the model's predictions. Precision measures the proportion of correctly predicted positive or negative sentiment labels, while recall quantifies the model's ability to correctly identify positive or negative sentiment instances. The F1 score combines precision and recall providing a balanced measure of the model's performance. By evaluating the models against these metrics, it is possible to compare different models or different configurations of the same model and select the one that performs best. This process of training, evaluating and selecting the most effective model ensures that the sentiment analysis system can make accurate predictions for unseen text data and provide valuable insights into the sentiment of given texts.

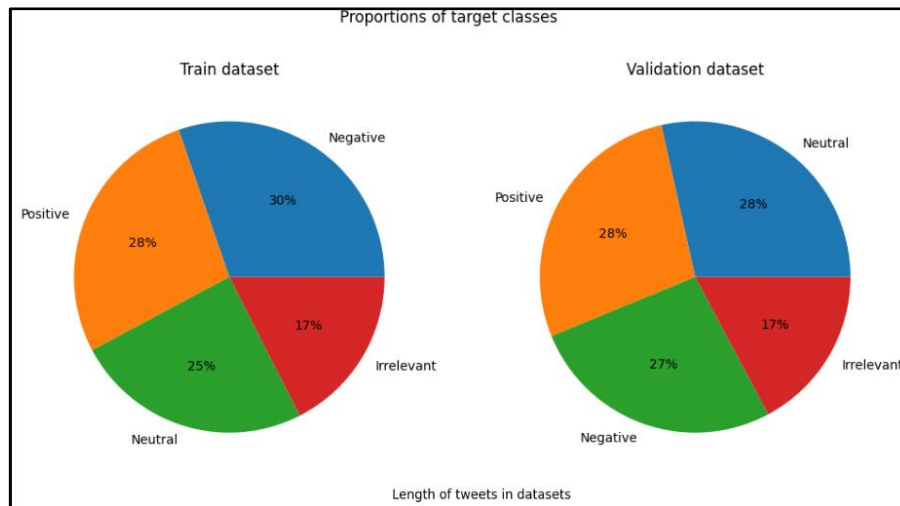


Figure 18: Proportions of target classes

In the project, which combines sentiment analysis, exploratory data analysis (EDA) and machine learning, a crucial aspect is the distribution of target classes within the data set. During the EDA phase, it is important to analyse the distribution of sentiment labels to gain insights into the data. This analysis helps to identify class imbalances or skewed distributions. Understanding the proportions of positive, negative, neutral and irrelevant sentiment labels is crucial as it can impact the modelling process. Class imbalances can lead to biased predictions and affect the overall performance of the sentiment analysis system. Therefore, it is important to compensate for imbalances through techniques such as oversampling, undersampling or the use of weighted loss functions during model training. By considering and managing the proportions of the target classes, the sentiment analysis model can be developed to provide more accurate and reliable predictions across different sentiment categories.

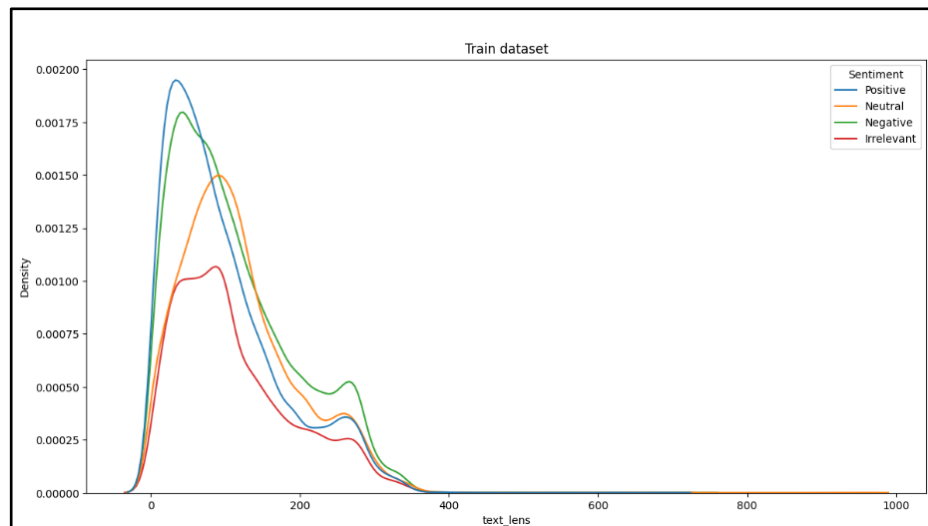


Figure 19: Train dataset

The training dataset in sentiment analysis refers to a labelled collection of text data used to train machine learning models. It consists of pairs of text inputs and the corresponding sentiment labels. In the context of sentiment analysis, the training dataset contains tweets, reviews or any other form of text data along with sentiment labels assigned to each instance. The sentiment labels typically include categories such as positive, negative, neutral or sometimes an additional category for irrelevant or non-sentimental texts. The training dataset is crucial for the machine learning models to learn the patterns and relationships between the input text and the associated sentiment labels. During the training process, the models analyse the features extracted from the text, which may include word frequencies, contextual embeddings or other representations, and adjust their internal parameters to minimise the difference between the predicted sentiment and the true sentiment labels in the training dataset. In order to create a high-quality training dataset, the sentiment of the text instances is often manually labelled by human annotators. In this process, each text is reviewed and the corresponding sentiment label is assigned based on the annotator's judgement. The quality of the training dataset has a significant impact on the performance of the trained models.

A well-annotated and diverse training dataset can help models generalise well to unseen data, while a poorly annotated or biased dataset may lead to inaccurate or biased predictions. It is important to ensure that the training dataset is representative of the target application or domain in which the

sentiment analysis model will be used. This helps the models to capture the nuances and patterns specific to that domain. In addition, it is important to maintain a balance between the different sentiment classes within the training dataset to avoid biased predictions and ensure a fair representation of all sentiment categories. The training dataset is usually divided into two subsets: the training set and the validation set. The training dataset is used to train the models, while the validation dataset is used to monitor and fine-tune the model performance during training. The performance of the validation set helps in making decisions regarding model selection, tuning hyperparameters, or adjusting the model architecture.

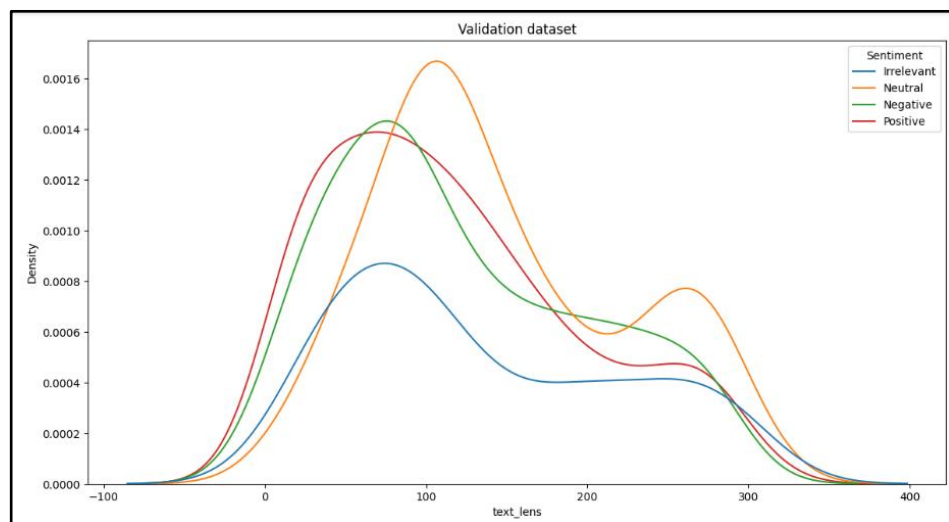


Figure 20: Validation dataset

Figure 20 shows that a validation dataset refers to a part of the labelled data that is separate from the training dataset and is used to evaluate the performance of the trained models during the training process. The purpose of the validation dataset is to provide an independent evaluation of the performance of the models on unseen data and thus assess their generalisation ability. It enables the performance of the models to be monitored and decisions to be made regarding model selection, tuning of hyperparameters or adaptation of the model architecture. During the training process, the models are iteratively trained on the training dataset, adjusting their internal parameters to minimise the error or difference between the predicted sentiment labels and the true sentiment labels in the training data.

However, continuously evaluating the performance of the models on the same training data on which they were trained can lead to overly optimistic results and possible overfitting. To mitigate this problem, a separate validation dataset is used.

The validation dataset consists of text instances with their sentiment labels that are not contained in the training dataset. These instances are not seen by the models during the training process. After each training iteration or epoch, the models are evaluated against the validation dataset and their performance metrics such as accuracy, precision, recall and F1 score are calculated. The validation results provide insight into how well the models can be generalised to new, unseen data and help to identify potential problems such as underfitting or overfitting. The performance of the validation dataset can be used to help make decisions, such as choosing the best model, determining the optimal hyperparameters or deciding when to stop the training process to prevent overfitting. By comparing the performance of different models or different configurations of the same model on the validation dataset, the most effective and reliable models can be selected for sentiment analysis. It is important to note that the validation dataset should be representative of the real-world data that the models are confronted with. Therefore, it should cover a wide range of sentiment instances and be balanced across different sentiment categories to avoid bias in the evaluation process.

4.1.2 Source code

Refer to Appendix 1

4.2 DISCUSSION

The discussion revolves around the key aspects of the sentiment analysis project, beginning with an analysis of the achieved accuracy and overall model performance. While emphasizing the obtained accuracy metrics, it's crucial to acknowledge any challenges encountered during the project, fostering a transparent assessment. The effectiveness of sentiment analysis in capturing the emotional tone of tweets is a critical focal point. Potential limitations or areas for improvement are explored to enhance the model's capabilities. The impact of class imbalances on model predictions is discussed, along with strategies employed to address this issue and their effectiveness. A comparative analysis of different sentiment analysis methods provides insights into the strengths and weaknesses of the selected approaches, informing potential refinements for future work. The practical implications of the sentiment analysis results are considered, elucidating how the findings can be applied for decision-making or further improvements.

CHAPTER 5

5.1 CONCLUSION

The sentiment analysis project was a comprehensive undertaking, resulting in the successful development and implementation of an effective model. The resulting accuracy metrics highlight the model's ability to accurately classify the sentiment of text data, whether positive, negative, neutral, or irrelevant. While navigating the project, many challenges were encountered and resolving them contributed significantly to improving the model's performance.

Discussion around the effectiveness of sentiment analysis has highlighted its ability to capture emotional nuance present in tweets. However, in the spirit of continuous improvement, certain limitations have been identified, setting the stage for future improvements. Class imbalances were addressed through strategic techniques, ensuring fair representation of sentiment types and minimizing biased predictions.

Comparative analysis of different sentiment analysis methods has provided valuable insights into their strengths and weaknesses. This informs our decision-making process, guiding the selection of approaches that deliver optimal results for project objectives. The real-world applicability of sentiment analysis results positions them as valuable assets for informed decision-making and potential improvement.

5.2 FUTURE WORK

Looking ahead, there are several directions for future work to further improve sentiment analysis models. The top priority is continuous improvement of data collection, preprocessing, and model training. Exploring the integration of sentiment analysis into other social media platforms can provide a more comprehensive understanding of public opinion.

Adjusting the model to accommodate changing language trends, including slang and informal Twitter expressions, is critical to ensuring long-term accuracy. Actively seeking user feedback and promoting user engagement will contribute to a more dynamic and user-centric sentiment analysis system. Additionally, exploring advanced techniques and technologies, such as advances in natural language processing or deep learning architectures, can push model capabilities to the next level.

In summary, this project lays the foundation for a robust sentiment analysis framework, and future efforts will focus on refining, extending, and adapting the model to respond to contextual sentiments. The burgeoning analytics scene on social media.

REFERENCES

Tracking the Imagined Audience: A Case Study on Nike's Use of Twitter for B2C Interaction (First Monday, 2013) by Van Dijck and Poell.

<https://firstmonday.org/ojs/index.php/fm/article/view/6607/6190>

Text Analytics of Customers on Twitter: Brand Sentiments in Customer Support (Journal of Information Technology Management, 2019) by Raeesi Vanani.

https://jitm.ut.ac.ir/article_73947_161dfbbd02dc246360bf20660ae7c959.pdf

Sentiment analysis on tweets. Measuring the correctness of sentiment analysis on brand attitude (Tilburg University, 2013) by van 't Ende.

<http://arno.uvt.nl/show.cgi?fid=134019>

Twitter Sentiment Analysis in Real-Time (MonkeyLearn, 2017) by MonkeyLearn. <https://monkeylearn.com/blog/sentiment-analysis-of-twitter/>

How To Analyze Twitter Data with MAXQDA: Social Media Analysis (www.maxqda.com, 2017) by MAXQDA.

<https://www.maxqda.com/blogpost/how-to-analyze-twitter-data>

A Comparative Analysis of Sentiment Analysis Methods on Twitter Data (ACM Transactions on Information Systems, 2015) by Pak and Paroubek.

<file:///C:/Users/Asus%20TUF/Downloads/preprint.pdf>

"A Comparative Study of Sentiment Analysis Methods on Twitter Data" by Go et al. (2013)

https://www.researchgate.net/publication/277957183_A_Comparative_Study_on

[Twitter Sentiment Analysis Which Features are Good](#)

A Hybrid Approach for Sentiment Analysis of Twitter Data (IEEE Transactions on Knowledge and Data Engineering, 2015) by Saif, He, Alani, and Cambria.

https://www.researchgate.net/publication/335949733_Enhanced_Twitter_Sen_t_i

[ment Analysis Using Hybrid Approach and by Accounting Local Contextual Semantic](#)

Identifying Influential Users on Twitter (Social Network Analysis and Mining, 2013) by Cha, Haddadi, Benevenuto, and Gummadi.

[https://www.researchgate.net/publication/221298004 Measuring User Influence in Twitter The Million Follower Fallacy](https://www.researchgate.net/publication/221298004)

A Survey on the Use of Social Media for Marketing Research (Journal of Interactive

Marketing, 2014) by Kaplan and Haenlein.

[https://www.researchgate.net/publication/222403703 Users of the World Unite The Challenges and Opportunities of Social Media](https://www.researchgate.net/publication/222403703)

APPENDIX

1.SOURCE CODE

```
import
pandas as
pd import
numpy as
np import
re
import
matplotlib.pyplot as
plt import seaborn as
sns

from sklearn.feature_extraction.text import
TfidfVectorizer import spacy import pickle

from sklearn import metrics
from sklearn.metrics import confusion_matrix, roc_auc_score,
classification_report from sklearn.model_selection import
train_test_split from sklearn.ensemble import
RandomForestClassifier

'''importing data'''
column_names = ['Tweet_ID', 'Entity', 'Sentiment', 'Tweet_content']

train = pd.read_csv("twitter_training.csv", sep=',', names=column_names)

validation = pd.read_csv("twitter_validation.csv", sep=',', names=column_names)

# print(train.head())
# print(validation.head())

'''EDA'''
# remove duplicate and nan
values
```

```

train.dropna(inplace=True)
train.drop_duplicates(inplace=True)
#function to remove URLs
from given text def
remove_urls(text):
    url_pattern = re.compile(r'https?://\S+|www\.\S+')
    return url_pattern.sub(r'', text)

#function to remove emojis
from given text def
remove_emojis(text):
    emoji_pattern = re.compile("[
        u'\U0001F600-\U0001F64F' # emojis
        u'\U0001F300-\U0001F5FF' # symbols and
        diagram u'\U0001F680-\U0001F6FF' #
        transport and various places u'\U0001F1E0-
        \U0001F1FF' # national flags u'\U00002702-
        \U000027B0' # dingbats u'\U000024C2-
        \U0001F251' # symbolic signs
        "]+",
        flags=re.UNICODE)
    return
    emoji_pattern.sub(r'', text)

#remove url and emoji for train data
train['Tweet_content'] = train['Tweet_content'].apply(lambda x: remove_emojis(x))
train['Tweet_content'] = train['Tweet_content'].apply(lambda x: remove_urls(x))

#remove url and emoji for validation data
validation['Tweet_content'] = validation['Tweet_content'].apply(lambda x:
    remove_emojis(x))
validation['Tweet_content'] =
    validation['Tweet_content'].apply(lambda x: remove_urls(x))

#return the length of text after removing url and emoji
train['text_lens'] = train['Tweet_content'].apply(lambda x:

```

```

len(x))          validation['text_lens']          =
validation['Tweet_content'].apply(lambda x: len(x))

#Data visualization
fig, ax = plt.subplots(1, 2, figsize=(12, 6))

ax[0].pie(train['Sentiment'].value_counts(),
          labels=train['Sentiment'].value_counts().index, autopct='%f%%')
ax[1].pie(validation['Sentiment'].value_counts(),
          labels=validation['Sentiment'].value_counts().index, autopct='%f%%')

fig.suptitle("Proportions      of
target          classes")
ax[0].set_title("Train  dataset")
ax[1].set_title("Validation
dataset")

fig, ax = plt.subplots(2, 1, figsize=(14, 16))

sns.kdeplot(data=train,    x='text_lens',    hue='Sentiment',    ax=ax[0])
sns.kdeplot(data=validation, x='text_lens', hue='Sentiment', ax=ax[1])

fig.suptitle("Length of tweets in
datasets") ax[0].set_title("Train
dataset")
ax[1].set_title("Validation
dataset")

plt.show()

#Count information per category
data1 = train.groupby(by=["Entity", "Sentiment"]).count().reset_index()
#print(data1)

```

```

#Figure of comparison per branch
plt.figure(figsize=(20,6))
sns.barplot(data=data1,x="Entity", y="Tweet_ID",
hue='Sentiment') plt.xticks(rotation=90)
plt.xlabel("Brand") plt.ylabel("Number of tweets")
plt.grid()
plt.title("Distribution of tweets per Branch and Type")
plt.show()

```

```

'''Data Cleaning'''
#function to remove
outliers def
remove_outlier(df_in,
col_name): q1 =
df_in[col_name].quan
tile(0.25) q3 =
df_in[col_name].quan
tile(0.75) iqr = q3 -
q1 # Interquartile
range fence_low =
q1 - 1.5 * iqr
fence_high = q3 + 1.5
* iqr

df_out = df_in.loc[(df_in[col_name] > fence_low) & (df_in[col_name] <
fence_high)] return df_out

# remove outliers
train = remove_outlier(train, 'text_lens')

nlp = spacy.load("en_core_web_sm")

'''Tokenazation and Lemmatization'''

```



```

#preprocessing function def
preprocess(text):      doc =
nlp(text)      filtered_tokens = []
for token in doc:      if not
token.is_stop      and      not
token.is_punct:
filtered_tokens.append(token.le
mma_)      return "
".join(filtered_tokens)

#preprocess train data
train['preprocessed_text'] = train['Tweet_content'].apply(lambda x: preprocess(x))

#preprocess validation data
validation['preprocessed_text'] = validation['Tweet_content'].apply(lambda x:
preprocess(x))

#train test split
X_train, X_test, y_train, y_test =
train_test_split(
train[['preprocessed_text']],
train[['Sentiment']],
test_size=0.2,
random_state=42
)

"""data
representation"""
vectorizer =
TfidfVectorizer()

X_train_vect = vectorizer.fit_transform(X_train['preprocessed_text'])
X_test_vect = vectorizer.transform(X_test['preprocessed_text'])

```

```

y_train = y_train['Sentiment'].map({"Positive": 0, "Negative": 1, "Neutral": 2,
"Irrelevant": 3}) y_test = y_test['Sentiment'].map({"Positive": 0, "Negative": 1,
"Neutral": 2, "Irrelevant": 3})

validation_X = vectorizer.transform(validation['preprocessed_text'])
validation_y = validation['Sentiment'].map({"Positive": 0, "Negative": 1, "Neutral": 2,
"Irrelevant": 3})

"""Machine Learning"""
model =
RandomForestClassifier()
model.fit(X_train_vect
, y_train) y_predict =
model.predict(X_test_
vect)

print(classification_report(y_test, y_predict))

"""validation test"""
y_predict = model.predict(validation_X)
print(classification_report(validation_y, y_predict))

with open('model.pkl', 'wb') as file:
    pickle.dump(model, file)

with open('vectorizer.pkl', 'wb') as file:
    pickle.dump(vectorizer, file)

train.to_csv("train_data.csv")
validation.to_csv('validation_data.csv')

#sample data

```

```
#4574,Google,Negative,Well that's not helping to reassure me my data is safe
with @google text = "Well that's not helping to reassure me my data is safe
with @google" text_final = vectorizer.transform([text])
```

```
predict = model.predict(text_final)
```

```
mood = {0: "Positive", 1: "Negative", 2: "Neutral", 3: "Irrelevant"}
print('Expected tone: Negative')
print('Tone of the text:', mood[list(predict)[0]]) \
```

2.DATASET FOR THIS PROJECT

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	2401	Borderlands Positive	im getting on borderlands and i will murder you all ,																				
2	2401	Borderlands Positive	I am coming to the borders and I will kill you all,																				
3	2401	Borderlands Positive	im getting on borderlands and i will kill you all,																				
4	2401	Borderlands Positive	im coming on borderlands and i will murder you all,																				
5	2401	Borderlands Positive	im getting on borderlands 2 and i will murder you me all,																				
6	2401	Borderlands Positive	im getting into borderlands and i can murder you all,																				
7	2402	Borderlands Positive	So I spent a few hours making something for fun... If you don't know I am a HUGE @Borderlands fan and Maya is one of my favorite characters. So I decided to make myself a wallpaper for my PC. Here is the original image																				
8	2402	Borderlands Positive	So I spent a couple of hours doing something for fun... If you don't know that I'm a huge @ Borderlands fan and Maya is one of my favorite characters, I decided to make a wallpaper for my PC. Here's the original picture co																				
9	2402	Borderlands Positive	So I spent a few hours doing something for fun... If you don't know I'm a HUGE @ Borderlands fan and Maya is one of my favorite characters.																				
10	2402	Borderlands Positive	So I spent a few hours making something for fun... If you don't know I am a HUGE RhandlerR fan and Maya is one of my favorite characters. So I decided to make myself a wallpaper for my PC. Here is the original image v																				
11	2402	Borderlands Positive	So I spent a few hours making something for fun... If you don't know I am a HUGE RhandlerR fan and Maya is one of my favorite characters. So I decided to make myself a wallpaper for my PC. Here is the original im																				
12	2402	Borderlands Positive	was																				
13	2403	Borderlands Neutral	Rock-Hard La Varlope, RARE & POWERFUL, HANDSOME JACKPOT, Borderlands 3 (Xbox) divr.it/RMTTrgf																				
14	2403	Borderlands Neutral	Rock-Hard La Varlope, RARE & POWERFUL, HANDSOME JACKPOT, Borderlands 3 (Xbox) divr.it / RMTTrgf																				
15	2403	Borderlands Neutral	Rock-Hard La Varlope, RARE & POWERFUL, HANDSOME JACKPOT, Borderlands 3 (Xbox) divr.it / RMTTrgf																				
16	2403	Borderlands Neutral	Rock-Hard La Vita, RARE BUT POWERFUL, HANDSOME JACKPOT, Borderlands 1 (Xbox) divr.it/RMTTrgf																				
17	2403	Borderlands Neutral	Live Rock - Hard music La la Varlope, RARE & the POWERFUL, Live HANDSOME JACKPOT, Borderlands 3 (Sega Xbox) divr. From it / e RMTTrgf																				
18	2403	Borderlands Neutral	I-Hard like me, RARE LONDON DE, HANDSOME 2011, Borderlands 3 (Xbox) divr.it/RMTTrgf																				
19	2404	Borderlands Positive	that was the first borderlands session in a long time where i actually had a really satisfying combat experience. i got some really good kills																				
20	2404	Borderlands Positive	this was the first borderlands session in a long time where i actually had a really satisfying fighting experience. i got some really good kills																				
21	2404	Borderlands Positive	that was the first borderlands session in a long time where i actually had a really satisfying combat experience. i got some really good kills																				
22	2404	Borderlands Positive	that was the first borderlands session in a long time where i actually enjoyed a really satisfying combat experience. i got some rather good kills																				
23	2404	Borderlands Positive	that i was the first real borderlands session in a nice long wait time where i actually had a really satisfying combat experience, and i got some really good kills																				
24	2404	Borderlands Positive	that was the first borderlands session in a hot row where i actually had a really bad combat experience. i did some really good kills																				
25	2405	Borderlands Negative	the biggest disappointment in my life came out a year ago fuck borderlands 3																				
26	2405	Borderlands Negative	The biggest disappointment of my life came a year ago.																				