OPINION MINING ON TESLA (2015) USING TWITTER DATA

A Dissertation Submitted
in Partial Fulfillment of the Requirements for the
Degree of

MASTERS OF SCIENCE
in
APPLIED MATHEMATICS

by

SUNIDHI SINGH RAJPUT
(2K22/MSCMAT/47)

Under the Supervision of
Dr. GOONJAN JAIN
Assistant Professor, Department of Applied Mathematics
Delhi Technological University

To the
DEPARTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

June, 2024

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I Sunidhi Singh Rajput hereby certify that the work which is being presented in the thesis entitled "OPINION MINING ON TESLA (2015) USING TWITTER DATA" in partial fulfilment of the requirements for the award of the Degree of Masters of Science, submitted in the Department of Applied Mathematics, Delhi Technological University, is an authentic record of my own work carried out during the period from August, 2023 to May, 2024 under the supervision of Dr.Goonjan Jain.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Sunidhi Singh Rajput

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Dr. Goonjan Jain **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE BY THE SUPERVISOR

Certified that Sunidhi Singh Rajput (2K22/MSCMAT/47) has carried out their search work presented in this thesis entitled ”OPINION MINING ON TESLA (2015) USING TWITTER DATA” for the award of Master of Science from Department of Applied Mathematics, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

DR. GOONJAN JAIN
SUPERVISOR
ASSISTANT PROFESSOR
DEPARTMENT OF APPLIED MATHEMATICS
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

Date:

ABSTRACT

In the current technological landscape, the surge in social media[1] usage has become a predominant platform for the exchange of ideas and beliefs. Opinion mining also referred to as sentiment analysis, has emerged as a pivotal tool for understanding public sentiment. Through the application of Natural Language Processing (NLP), sentiment analysis automates the extraction of attitudes, opinions, and emotions from diverse sources, including text, audio, tweets, and databases. Leveraging data from Kaggle, a renowned hub for data science and machine learning, our research focused on analyzing tweets pertaining to top companies from 2015 to 2020 to gain comprehensive insights into international public sentiment trends. Twitter, with its widespread accessibility, facilitated swift and efficient data collection, offering valuable insights into ongoing topical discussions.

Our study concentrated specifically on examining approximately 30,000 tweets about Tesla from early 2015. Employing methodologies such as CountVectorizer and various classification algorithms including Naive Bayes, Decision Trees, Random Forests (RFC), Logistic Regression, XGBoost, and Support Vector Machines (SVM), we sought to categorize and assess the emotions elicited by these tweets. Our findings underscored the efficacy of Random Forest and SVM in providing the highest classification accuracy, thereby significantly contributing to a nuanced understanding of public sentiment dynamics surrounding Tesla and, by extension, other top companies. This research not only sheds light on the evolving landscape of public opinion but also underscores the potential of sentiment analysis techniques in informing decision-making processes across various domains.

ACKNOWLEDGEMENT

Throughout the process of composing this dissertation, I received considerable assistance and encouragement from numerous individuals. Firstly, I am profoundly grateful to Dr. Goonjan Jain for her substantial contributions to the formulation of the research topics and methodologies. Her insightful feedback compelled me to refine my ideas and elevate the quality of my work. Her guidance and the provision of essential resources were pivotal in enabling me to navigate the correct path and successfully complete this dissertation.

Additionally, I also want to sincerely thank God and my parents for their wise advice and understanding support. Their constant encouragement and support were essential throughout this journey.

Sunidhi Singh Rajput
MSc. Applied Mathematics
DTU, New Delhi

List of Tables

List of Figures

LIST OF ABBREVIATIONS

1. Sentiment Analysis	SA
2. Machine Learning	ML
3. Natural Language Processing	NLP
4. Random Forest Classifier	RFC
5. Extreme Gradient Boosting	XGBoost
6. Decision Tree Classifier	DTC
7. Naïve Bayes Classifier	NBC
8. Support Vector Machine	SVM

CONTENTS

Chapter 1

INTRODUCTION

Opinion mining, often known as Sentiment Analysis (SA)[2], is a powerful tool used to evaluate public sentiment through natural language processing (NLP). This advanced technique automates the process of extracting attitudes, opinions, perspectives, and emotions from a variety of sources, including text documents, audio recordings, tweets, and databases. At its core, SA focuses on detecting subjectivity and polarity within text, while semantic orientation quantifies the text's polarity and the strength of these sentiments. With the immense volume and rapid creation of user-generated content on social media platforms, machine learning models have become essential for accurately determining public opinion. In our digital era, information spreads at an unprecedented speed among users, significantly shaping collective perceptions of events. Consequently, it is imperative to understand and identify popular opinion, which is crucial for professionals and researchers in diverse fields such as human-computer interaction, sociology, marketing, advertising, psychology, economics, and political science.

People living in a society constantly form judgments about the world around them. They develop opinions about people, objects, places, and events, which are recognized as sentimental attitudes. SA involves the study of automated methods to extract these sentiments from written text. The advent of social media platforms has led to an explosion of freely available, user-generated content on the World Wide Web. This abundance of data can be leveraged to provide real-time insights into people's attitudes and opinions. Social media includes blogs, online forums, comment sections on news websites, and social networking sites such as Facebook and Twitter. These platforms have the potential to capture the voices and opinions of millions of individuals.

The ability to share and access real-time opinions from people worldwide has revolutionized computational linguistics and social network analysis. Social media has become an indispensable information source for businesses. Simultaneously, people are more inclined than ever to share details about their lives, knowledge, experiences, and thoughts with the world through social media. They actively engage in societal events by expressing their ideas and making observations about current happenings. This propensity to share knowledge and feelings with society and on social media drives businesses to gather extensive information about their companies, products, and reputations. This data collection enables businesses to make more informed decisions.

Moreover, the impact of social media on SA has been profound, providing a vast and diverse dataset for in-depth analysis. Platforms like Twitter and Facebook empower companies to monitor brand sentiment in real time, allowing them to respond swiftly to customer feedback and market trends. This immediate access to public opinion aids businesses in adapting their strategies quickly, thereby enhancing customer satisfaction and fostering brand loyalty. Additionally, SA helps predict market movements and consumer behavior by identifying trends and patterns in social media discussions.

Beyond business applications, SA on social media is crucial in various other domains, including political campaigns, public health, and disaster response. Politicians and policymakers utilize SA to gauge public reactions to policies and speeches, enabling them to adjust their approaches accordingly. Public health officials can track the spread of diseases or public concerns about health issues by analyzing social media conversations. During disasters, SA assists authorities in understanding public sentiment and needs, improving response efforts and communication strategies.

Overall, the integration of social media with SA offers profound insights and actionable data. This synergy drives advancements across multiple fields, enhancing our understanding of public opinion and behavior. The ability to analyze vast amounts of real-time data from social media platforms provides a comprehensive view of societal trends and public sentiment, facilitating more informed decision-making processes in various sectors.

1.1 Sentiment Analysis

SA, a key area of NLP and Information Extraction, involves analyzing a large corpus of texts to determine the writer's emotions, such as positivity, negativity, or neutrality. It seeks to understand the author's stance on a specific topic or the overall mood of a piece of writing. The surge in internet usage and the proliferation of public opinion sharing have propelled the significance of SA. The web is teeming with structured and unstructured data, making it a formidable challenge to extract and interpret hidden sentiments. SA can be approached at various granularities. At the document level, it evaluates the overall sentiment of an entire text, categorizing it as positive, negative, or neutral. At the sentence level, it identifies and classifies the sentiment expressed in individual sentences. At the phrase level, it determines the polarity of specific phrases within sentences. This process differentiates between objective statements and subjective opinions, identifying which parts of a text contain sentiments. A crucial element of SA is pinpointing the target of the sentiment. In sentences with multiple entities, it's important to discern which entity the sentiment pertains to. SA also assesses the polarity (positive, negative, or neutral) and intensity of the sentiment. Sentiments can be objective (fact-based), positive (expressing happiness or satisfaction), or negative (indicating disappointment or dissatisfaction). Additionally, sentiments can be ranked by their intensity of positivity, negativity, or neutrality.

In summary, SA leverages vast online data to glean insights into public opinion and emotions, proving invaluable across multiple fields, including business, politics, public health, and beyond. Advances in NLP and machine learning continue to refine the accuracy and depth of SA, expanding its utility and impact.

1.2 Classification Levels in Sentiment Analysis

Sentiment Analysis (SA) employs a structured, three-tiered approach for effective sentiment classification:

1. **Document-Level Analysis:** This top layer evaluates the overall sentiment of an entire document. It provides a broad overview, categorizing the whole text as positive, negative, or neutral. This level is ideal for summarizing the general mood of lengthy content such as articles, reviews, or reports.

2. **Sentence-Level Analysis:** The next layer breaks the document into individual sentences, assessing the sentiment within each one. This granular approach captures the nuances in text, recognizing that different sentences can convey varying sentiments, even within a single document. It's particularly useful for documents with mixed sentiments, ensuring that every sentiment is accounted for.

3. **Aspect-Level Analysis:** Also known as word or phrase-level analysis, this deepest layer focuses on specific components within the text. It identifies and evaluates the sentiment associated with particular words or phrases, often related to distinct aspects or features of a subject. This precise analysis is crucial for detailed feedback, such as pinpointing specific strengths or weaknesses in product reviews.

By integrating these three levels, SA can deliver a comprehensive and nuanced understanding of textual sentiments, enhancing accuracy and providing valuable insights across various applications.

1.3 Advantages of Sentiment Analysis

SA is a powerful tool that combines[2] NLP and ML to understand public opinion and emotions from various text sources. This technique is widely used across industries to gain insights into customer feelings, market trends, and overall sentiment towards products, services, or events. By leveraging ML, SA becomes more efficient and accurate, providing actionable insights in real-time. Some key advantages of using SA are as follows:

- **Real-Time Market Feedback:** ML algorithms analyze social media posts and reviews instantly, providing businesses with immediate insights. For example, a company can quickly adjust a marketing campaign based on customer reactions on Twitter, ensuring their message resonates with the audience.

- **Enhanced Customer Support:** By examining feedback with SA, companies can identify recurring issues and address them proactively. For instance, an e-commerce platform can detect and resolve common complaints about delivery delays, improving customer satisfaction and loyalty.
- **Product Improvement:** SA highlights specific features that customers love or dislike, guiding product development. A software company can use this data to enhance popular features and fix problematic ones, ensuring the product meets user expectations and stands out in the market.
- **Crisis Management:** Monitoring sentiment trends can alert companies to potential PR crises early. For example, a brand can detect a spike in negative sentiment about a new policy and respond quickly to mitigate backlash, protecting its reputation and maintaining customer trust.
- **Competitor Benchmarking:** Analyzing sentiments about competitors provides strategic insights. A retail chain can understand why customers prefer a rival's service, then improve its own offerings to gain a competitive advantage, ensuring they remain competitive and appealing to their target market.
- **Targeted Marketing Campaigns:** ML helps segment customer preferences and sentiments, enabling personalized marketing. A fashion brand can create targeted ads based on positive sentiments about certain product lines, increasing engagement and sales, and maximizing the return on marketing investments.

By combining ML with SA, organizations can gain a deeper, more accurate understanding of public opinion, leading to better decision-making and strategic planning across various domains.

1.4 Some Applications

SA has numerous applications across various industries and domains. Here are some notable ones:

- Travel Optimization: By analyzing sentiments in reviews and social media, travelers can plan their trips more effectively, ensuring enjoyable experiences.
- Dining Delight: SA enhances restaurant recommendation systems, suggesting venues aligned with user tastes for a satisfying dining experience.
- Event Engagement: Real-time sentiment monitoring on social media during events allows organizers to gauge audience reactions and tailor strategies for maximum engagement.
- Mental Health Monitoring: Tracking changes in sentiment from online interactions aids in the early detection of mental health issues, enabling timely interventions.
- Environmental Advocacy: Understanding public sentiment towards environmental issues helps organizations tailor campaigns and rally support for conservation efforts.

- Cultural Heritage Appreciation: SA of historical materials assists in preserving cultural legacies, fostering community engagement and appreciation.

These applications demonstrate SA's adaptability, providing insights and driving positive outcomes across diverse fields.

1.5 Issues Faced

- Deciphering sarcasm and irony:

  Deciphering sarcasm and irony in text is akin to navigating through a maze of words, where meanings can sometimes be misleading. Take this scenario: "Oh, wonderful, another meeting!" Despite the word "wonderful" typically indicating positivity, its usage here likely conveys sarcasm, suggesting frustration or annoyance about attending yet another meeting. Similarly, in the statement, "Fantastic, another deadline extension," the term "fantastic" may be used ironically to express exasperation over the ongoing delays. Unraveling such nuanced expressions requires algorithms capable of discerning subtle linguistic cues beyond the literal meanings of words, ensuring accurate interpretation of sentiments encoded within the text.

- Subjectivity Identification:

  Subjectivity identification in text analysis is comparable to sifting through a field of mixed grains, where separating the valuable wheat from the unwanted chaff requires keen discernment. Just as wheat stands out for its value amidst the chaff, subjective opinions shine amidst objective statements in text data. However, distinguishing between the two demands advanced algorithms equipped with the ability to recognize subtle linguistic nuances and contextual cues. For instance, consider the sentence: "The movie was fantastic." Here, "fantastic" indicates a subjective opinion, while a statement like "The movie lasted two hours" is objective, merely stating a fact. To accurately identify subjective opinions amidst objective statements, algorithms must be trained to detect language patterns indicative of personal viewpoints, sentiments, or emotions, ensuring nuanced analysis of textual data.

- Entity-Level Opinion Extraction:

  Entity-level opinion extraction entails isolating opinions tied to particular entities, such as products or services, from a vast sea of textual data. This task is comparable to searching for a needle in a haystack, where the desired opinions represent the needle hidden within the haystack of information. Just as finding a needle requires meticulous search techniques, extracting entity-specific opinions demands precise extraction methods amidst a deluge of text. For example, in a product review dataset, identifying opinions about a specific smartphone model amidst a multitude of reviews requires algorithms capable of pinpointing mentions of the smartphone and extracting associated sentiments accurately. This process necessitates sophisticated natural language processing techniques tailored to recognize entity references and discern corresponding opinions, ensuring precise analysis in sentiment mining tasks.

- Contextual Understanding:

  Contextual understanding in sentiment analysis parallels solving a complex puzzle, where comprehending the complete context of opinions is essential for accurate interpretation. It's akin to piecing together disparate puzzle fragments to reveal the bigger picture, requiring a nuanced approach beyond superficial examination. Just as solving a puzzle demands attention to detail and consideration of how each piece fits into the larger whole, contextual understanding entails comprehensive analysis to grasp the underlying sentiment nuances embedded within the text. For instance, in a customer review, understanding the context of a positive statement about a product may involve considering factors like previous experiences, expectations, or comparison with competitors. This holistic approach ensures that sentiments are interpreted accurately, capturing the subtleties and intricacies of language usage within the broader context.

- Emoji:

  Emojis add complexity to opinion mining as their meanings are subjective and context-dependent. Their limited vocabulary and platform-dependent variability make accurate sentiment analysis challenging. Algorithms must interpret nuances and integrate text and visual data effectively. Emojis can modify adjacent text sentiment or convey additional layers of meaning, demanding thorough contextual understanding. Overcoming these challenges requires advanced natural language processing and machine learning techniques. Ensuring consistent interpretation across datasets is crucial. Emojis enrich communication with emotional context but necessitate careful consideration in opinion mining to extract reliable insights from textual data.

- Cross-Domain Opinion Generalization:

  Cross-domain opinion generalization entails applying sentiment analysis models trained on one domain to other domains, which demands adaptable algorithms capable of discerning sentiment patterns amidst diverse datasets and topics. It's akin to navigating through unfamiliar terrain, requiring algorithms to recognize underlying sentiment nuances across different contexts. This task necessitates robust techniques for transferring knowledge and understanding sentiments beyond the scope of the original training data, ensuring accurate analysis across various domains.

Chapter 2

METHODOLOGY

The full explanation of the methodology used to ascertain public opinion on Tesla company from January to March of 2015 as shown in Figure 2.1.

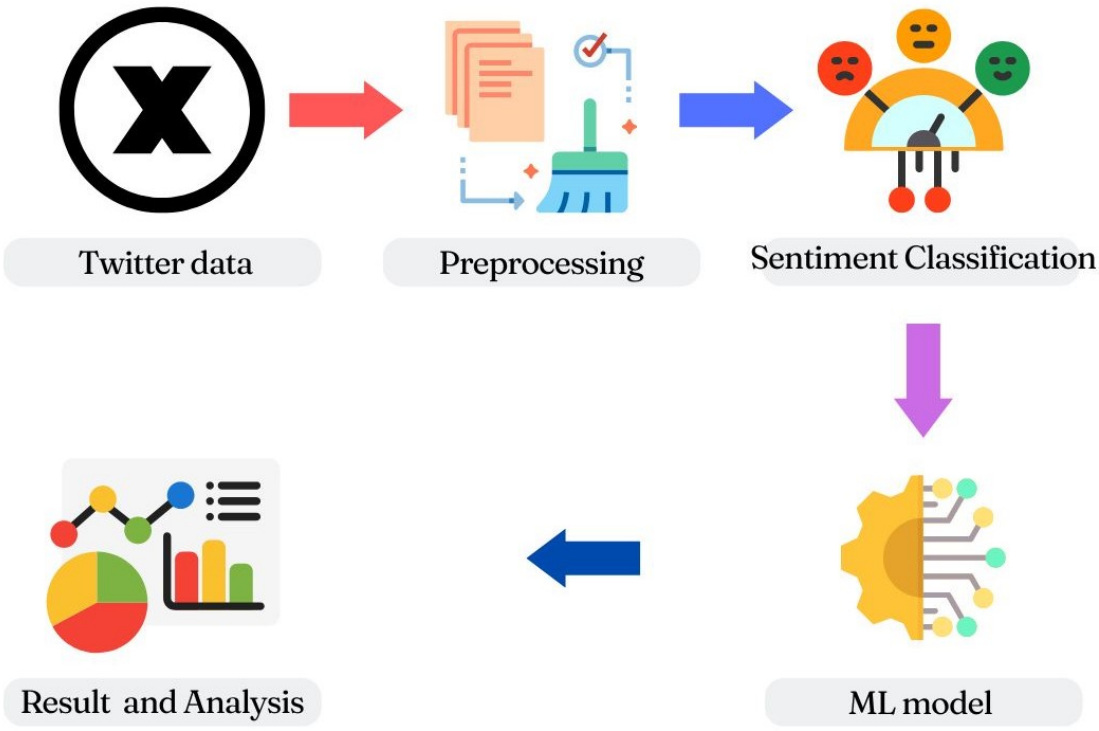

Figure 2.1: Methodology

The method employed to derive sentiment from tweets is outlined as follows:

1. Gathering Data: We collected tweets about Tesla from January to March 2015.
2. Cleaning Text: We removed common, insignificant words from the tweets.
3. Breaking Down Words: Each tweet was split into individual words.

4. Analyzing Sentiment: Using a specialized algorithm, we determined whether each tweet expressed positive or negative sentiment.
5. Data Split: We divided the dataset into two parts: one for training our model and the other for testing its accuracy.
6. Training the Model: Using the training data, we taught a machine learning model to recognize sentiment patterns in tweets.
7. Testing the Model: Finally, we assessed the model's performance by using it to predict sentiment on the test dataset and comparing its predictions to the actual sentiment expressed in the tweets.

This approach allowed us to gauge public opinion on Tesla during the specified timeframe using social media data.

2.1 About the Data

The dataset, featured in the 2020 IEEE International Conference on Big Data within the 6th Special Session on Intelligent Data Mining track[3], was meticulously curated to unveil potential speculators and influencers within the stock market[3]. A vast pool of $43,36,445$ tweets sourced from Kaggle, publicly accessible, served as the foundation. Specifically focusing on Tesla, data extraction was tailored to encompass only tweets from January to March of 2015, leveraging a meticulous approach to handle non-standard date formats. Post conversion to a standardized format (e.g., "$2015-01-0123:18:50$"), the dataset was distilled into a CSV file, with duplicates expunged through Excel functionalities. The resultant dataset boasts $29,135$ unique tweets, featuring tweet IDs and their respective bodies, providing an insightful lens into public sentiment surrounding Tesla during the early months of 2015 -a period marked by Tesla's prominence as one of the world's foremost companies by market capitalization.

2.2 Data Pre-Processing

The final dataset, extracted from the main data, initially contained $29,135$ tweets. After removing duplicates using Excel, we were left with $25,312$ unique tweets, as illustrated in Table 2.1.

The data cleaning process encompassed several crucial steps:

- Transformation: standardizing the tweets by converting all text to lowercase and removing URLs and special characters.
- Stopword Removal: Eliminating common words that do not affect the overall meaning, such as "that" and "the."
- Tokenization: breaking down tweets into individual words using spaces and underscores as delimiters.

These steps were essential for refining the data and preparing it for sentiment analysis.

Table 2.1: Sample Data of the Tweets

tweet_id	body
5.78602E+17	Here's @DanaHull on Tesla software update that seeks to ensure drivers never accidentally run out of juice- http://bayareane.ws/1B5RHSp$TSLA
5.60502E+17	Boston-Power Aims to Rival Tesla With Gigawatt Lithium Battery Factories. http://bit.ly/1CzhQxj# $TSLA $LIT $ILHMF
5.53289E+17	@FoxBusiness @MonicaCrowley @Varneyco Penalizing struggling Americans? put a luxury tax on $TSLA luxury cars @larry_kudlow @SpeakerBoehner
5.8204E+17	Testing calling out day trades next week on TWTR - $AAPL $FB $TSLA $IWM $SPY $GOOGL $AMZN $NFLX $TWTR - follow for trades.
5.58311E+17	Posted Ystrday: $TSLA wow, 6$ move had this post for members ystrday big 400% $$$ http://stks.co/p1KIe"
5.54161E+17	What #sector will you #invest in today? http://optionmillionaires.com $TSLA $SINA $GRMN $SSYS
5.82595E+17	$TSLA - I sold out at 190.50 -
5.55373E+17	#WorldBank Wednesday â€“ Global #GDP Outlook Cut By 10% $TSLA $SPY $SPX $INDU $COMP $TASR http://stks.co/e1Ud0
5.74827E+17	$TSLA - Tesla to slash Chinese workforce http://uk.advfn.com/news/SEEK/2015/article/65768570?xref
5.72363E+17	"Free course on \""Beta\"" of #stocks http://bit.ly/BetaStra $CMG $TSLA $BIDU "

2.3 Classification of Sentiment

Tweets are given a polarity score ranging from −1 to +1, with −1 being the most negative and +1 being the most positive[4]. We can fine-tune the classification by choosing an interval value. For instance, if the polarity score exceeds 0.01, the tweet is labeled positive; if it falls below −0.01, it's labeled negative, and any score in between −0.01 and 0.01 is deemed neutral. An example of this classification is depicted in Figure 2.2[5], which illustrates the distribution of sentiments in a sample dataset. Interestingly, an early 2015 analysis of Tesla-related tweets revealed that approximately 57.6 % of individuals expressed neutral sentiments, indicating a lack of strong positive or negative feelings towards the company as represented in Figure 3.4.

Table 2.2: Sample Tweets for Polarity and Sentiment

body	polarity	Sentiment
news cars getting hacked press days company planet id trust protect tsla	0	neutral
sector invest today tsla sina grmn ssys	0	neutral
well nothing said let tsla whipsaw yesterday completely sync market	0.1	Positive
get day free trial esf live trading room visit spy fb nqf spx ymf tsla nflx twt qqq dji	0.268182	Positive
exchanges circuit break hft play regular traders got stuck queue lol esf spy aapl tsla nflx	0.4	Positive
seriously get free email trade alerts stocks like tsla nflx amzn aapl fb twtr baba pcln	0.033333	Positive
tesla big layoffs chinawhat predictable mess tsla made far	-0.06875	negative
initial tsla reaction	0	neutral
greatest stocktrading partnership pairtrade hedge stocks spy twtr tsla iwm qqq	1	Positive
tsla increased vxx position	0	neutral

2.4 Model Building

This section delves into the process of classifying and predicting sentiment from tweets using six powerful ML algorithms: Random Forest, Gradient Boosting, Decision Tree, Naive Bayes[6], Logistic Regression, and Support Vector Machine. To effectively feed textual data into these models, it's essential to convert the words or tokens into numerical or vector representations. In this case, we utilized the Count Vectorizer technique to transform the text strings into numerical vectors, which can then be used as input for our chosen ML models.

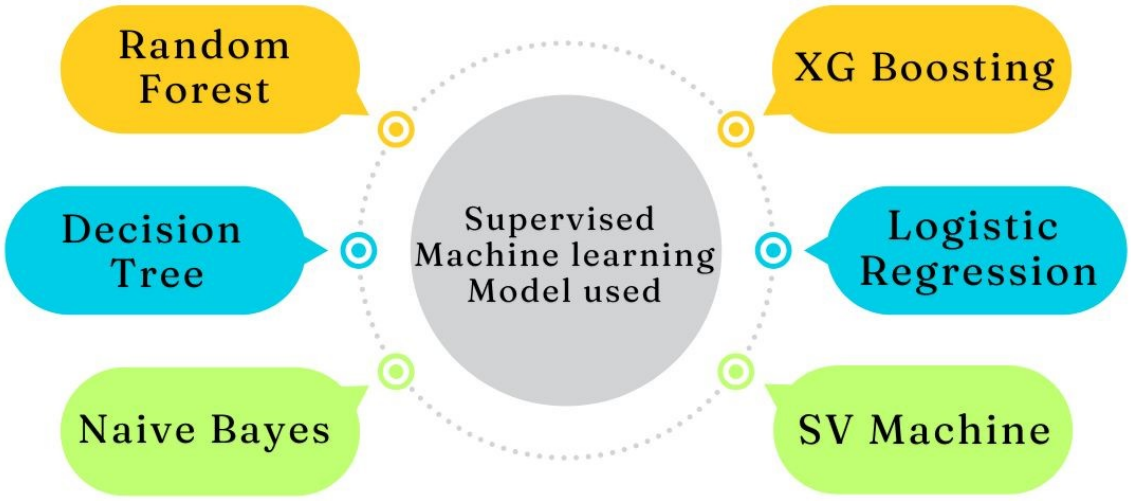

Figure 2.2: Used ML Algorithms

CountVectorizer

CountVectorizer is a technique used to transform a collection of text strings into a frequency-based representation. It essentially counts the occurrences of each word or token within a text corpus. The sklearn library provides the CountVectorizer function, which enables this conversion. However, one limitation of CountVectorizer is its inability to distinguish between more or less important words. It simply treats words with higher frequencies as statistically significant without considering their actual importance or context. Additionally, CountVectorizer does not capture relationships between words, such as linguistic similarities or semantic connections.

2.4.1 Random Forest

The Random Forest Classifier (RFC) is a machine learning algorithm that combines multiple randomized decision trees and aggregates their predictions through averaging. It consists of two main stages: random forest construction and prediction using the generated random forest classifier. In random forest construction, a subset of features is randomly selected, and the best split point for these features is determined to calculate nodes. The network is divided into daughter nodes based on the optimal split point, and this process is repeated until a specified number of nodes is reached. Finally, a forest is built by repeating these steps multiple times, resulting in a collection of decision trees.

Random Forest is known for its adaptability to various learning tasks and its ability to provide measures of variable importance, offering insights into the significance of different features in the classification process.

2.4.2 Gradient Boosting

XGBoost, short for Extreme Gradient Boosting[7], has become a cornerstone of machine learning due to its exceptional performance and versatility. By sequentially combining weak learners, predominantly decision trees, XGBoost constructs a potent ensemble model. Its iterative approach, employing gradient boosting, continuously minimizes loss functions, enhancing predictive accuracy. Additionally, XGBoost integrates regularization techniques to curb overfitting and bolster generalization capabilities, ensuring robust performance across diverse datasets and applications. Renowned for its scalability, efficiency, and reliability, XGBoost remains a top choice for addressing a myriad of machine learning challenges encountered in various domains. Its ability to handle complex datasets and deliver superior results, coupled with its adaptability and efficiency, makes XGBoost indispensable for practitioners seeking reliable and accurate solutions in the ever-evolving landscape of machine learning.

2.4.3 Decision Tree

The fundamental concept of the Decision Tree Classifier (DTC) lies in simplifying complex decisions into a series of more straightforward ones, aiming to achieve the desired solution iteratively. This approach involves segmenting a population into branch-like segments, forming an inverted tree structure comprising root, internal, and leaf nodes. Each decision leads to a leaf node, which signifies whether the sentiment is positive, negative, or neutral without further branching. However, potential limitations arise as errors may accumulate across levels in large trees, particularly when dealing with numerous classes. This scenario can lead to an excessive number of terminals compared to actual classes, subsequently increasing search time and memory space requirements. Hence, while the DTC offers an intuitive approach to decision-making, its scalability and efficiency may be compromised in certain scenarios due to these inherent limitations.

2.4.4 Naïve Bayes

The Naïve Bayes Classifier (NBC)[8] is a text mining technique utilized for opinion mining tasks, effectively categorizing opinions as positive or negative. Leveraging Bayes' Theorem, NBC quantifies the likelihood of a hypothesis given prior knowledge, making it a robust method for text classification. Specifically, Multinomial Naïve Bayes is commonly employed in text mining, operating as a supervised learning algorithm[9] to classify text based on the probability of a class within a document. Research has shown that the accuracy of NBC is not solely contingent on the degree of feature dependencies, as measured by class-conditional mutual information between features[10]. Instead, a more reliable predictor of accuracy is the loss of information regarding class features under the naive Bayes model[10] assumptions. This highlights the nuanced nature of NBC's performance assessment and underscores the importance of understanding its underlying assumptions for effective application.

2.4.5 Logistic Regression

At the heart of logistic regression[8] lies the concept of the logit, which represents the natural logarithm of an odds ratio. This mathematical framework enables logistic regression to model the probability of an outcome based on individual characteristics. Functioning akin to linear regression, logistic regression is tailored for scenarios with a binomial response variable. Widely utilized in epidemiological studies, logistic regression serves as a robust tool, facilitating simultaneous analysis of multiple explanatory variables while mitigating the influence of confounding factors. Its versatility and ability to account for complex relationships make it indispensable for researchers seeking to uncover insights into various phenomena while controlling for potential confounders

2.4.6 Support Vector Machine

The Support Vector Machine (SVM)[11] is a mathematical framework and algorithm designed to maximize a specific mathematical function relative to a set of data points. Renowned for its robust theoretical foundation and impressive empirical performance, SVM has garnered widespread adoption across various domains. Particularly in text classification tasks, SVM excels at distinguishing between positive and negative texts. Its effectiveness in this domain is attributed to several advantages, including its capability to effectively handle large feature sets. This ability to manage extensive feature spaces makes SVM a preferred choice for tasks involving high-dimensional data, such as natural language processing and text analysis.

2.5 Predictions of Model

In our project, we utilized machine learning algorithms with the help of a Python-based toolkit called scikit-learn to analyze our dataset. This involved forecasting, visualizing, and analyzing the gathered data, as well as evaluating the performance of machine learning algorithms[12] based on the following parameters:

Accuracy

Measures how often the classifier makes the correct prediction.

$$\text{accuracy} = \frac{\text{number of correct predictions}}{\text{total of predictions}}$$

Precision

Indicates the proportion of tweets we classified as positive, negative, or neutral that genuinely were positive, negative, or neutral.

$$\text{Precision} = \frac{\text{true positive}}{(\text{true positive} + \text{false positive})}$$

Recall

Shows the proportion of tweets that were actually positive, negative, or neutral, which our classification correctly identified.

$$\text{Recall} = \frac{\text{true positive}}{(\text{true positive} + \text{false negative})}$$

F1 score

A common method for measuring classifier performance in sentiment analysis, as it calculates the harmonic mean between precision and recall.

$$\text{f1 score} = 2 * \frac{\text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})}$$

[13]

Chapter 3

RESULTS AND DISCUSSION

3.1 Results

In the realm of machine learning, a comprehensive evaluation of various algorithms unfolds in Table 3.1. This table encapsulates the performance metrics of six distinct methods, each with its own strategic approach. The contenders include the classic Naive Bayes, the decision-making Decision Tree, the ensemble might of Random Forest, the versatile Logistic Regression, the boosting powerhouse Gradient Boosting, and the versatile SVC with CountVectorizer.

Table 3.1: Performance Comparison for ML models

MODEL	POSITIVE			NEGATIVE			NEUTRAL		
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
Random Forest	0.93	0.86	0.90	0.87	0.65	0.74	0.89	0.98	0.93
XG Boost	0.95	0.80	0.87	0.87	0.62	0.72	0.85	0.98	0.91
Decision Tree	0.81	0.89	0.85	0.69	0.74	0.72	0.93	0.86	0.89
Multinomial Naive Bayes	0.71	0.86	0.78	0.67	0.51	0.58	0.88	0.82	0.85
Logistic regression	0.95	0.80	0.87	0.87	0.62	0.72	0.85	0.98	0.91
Support Vector Machine	0.93	0.86	0.90	0.87	0.65	0.74	0.89	0.98	0.93

In this table 3.2, each row presents the accuracy achieved by a specific machine learning model. The accuracy metric assesses the overall correctness of the model's predictions[6], providing insight into its effectiveness in distinguishing between different classes or categories.

Table 3.2: Accuracy of the Models

ML Model	Accuracy
Random Forest	0.89
XG Boost	0.87
Decision Tree	0.82
Multinomial Naive Bayes	0.74
Logistic regression	0.85
Support Vector Machine	0.89

3.2 Analysis

Figures 1 and 2 present the positive and negative word clouds derived from the dataset. These word clouds visually represent the most frequently used words, highlighting both positive and negative sentiments. In a word cloud, words that appear more often are displayed in larger fonts, making it easy to identify key themes. This method is increasingly popular for quickly grasping the central ideas of textual content. The word clouds help illuminate the sentiments surrounding Tesla in 2015. Prominent words in the positive word cloud include ”best,” ”important,” ”autonomous,” ”nicely,” ”perfect,” ”happy,” ”popular,” ”hot,” ”detailed,” ”advanced,” and ”excited.” Conversely, the negative word cloud features words such as ”stupid,” ”poor,” ”moron,” ”crap,” ”sick,” ”crude,” ”idiot,” and ”impossible.”

Figure 3.1: Positive word cloud

Figure 3.2: Negative word cloud

Figure 3.3 illustrates the polarity scale for each tweet in the dataset, showcasing Polarity vs. Tweet Serial Number. Most of the polarity scores are concentrated within the [0, 0.50] range, suggesting an overall positive sentiment towards the company. Despite this, individual sentiments vary from negative to positive. This highlights the significant influence of media and politicians in shaping public opinion, whether favorable or unfavorable. In early 2015, with approximately 35,000 tweets analyzed, the sentiment distribution was nearly evenly split, with a majority of tweets classified as neutral as shown in Figure 3.4.

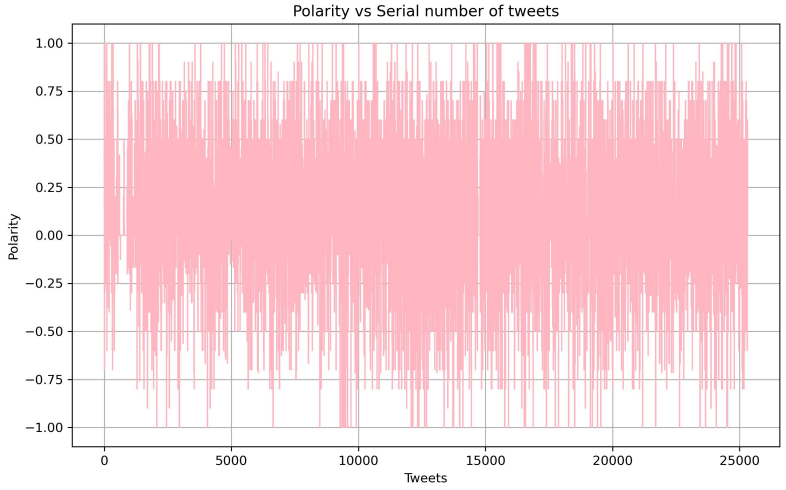

Figure 3.3: Polarity vs. Tweet Serial Number

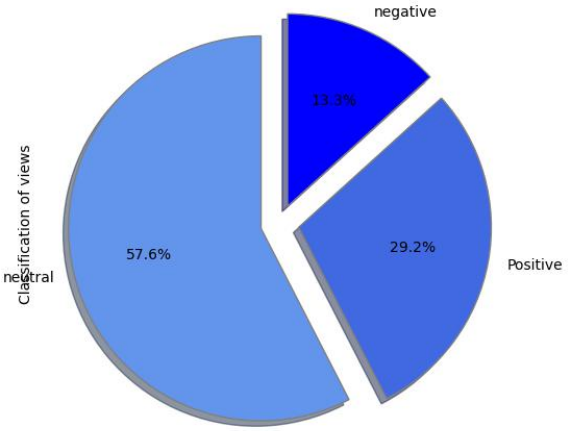

Figure 3.4: Pie Distribution of Sentiments

Figure 3.5 displays the confusion matrix for the Random Forest model using CountVectorizer. The true label represents the actual sentiment of the tweets, while the predicted label shows the sentiment as predicted by the model. A confusion matrix is a specialized table that presents data on actual and predicted classifications, allowing for the assessment of a classification algorithm's performance. According to this matrix, the Random Forest model correctly predicted the sentiment of 89.1 percent of tweets, with 10.9 percent incorrectly classified.

Similarly, Figure 3.6 shows the confusion matrix[14] for the Support Vector Machine (SVM) model, also using CountVectorizer. The SVM model correctly predicted the sentiment of 88.9 percent of tweets, with an error rate of 11.1 percent.

Figure 3.7 provides a comparative analysis of the accuracy of different machine learning algorithms employed for sentiment analysis. Random Forest and SVC demonstrate the highest accuracy levels, achieving 89.1 percent and 88.9 percent respectively. In contrast, Naive Bayes shows the least accuracy, with a rate of 74.1 percent.

3.3 Discussion

This study focused on analyzing tweets related to Tesla Company during early 2015 to understand user perceptions. Notable advancements made by Tesla in 2015 include the introduction of battery packs, such as the Tesla Powerwall and Powerpack, which received significant orders within a short period. Additionally, the launch of the Model X luxury SUV and the announcement of solar power products marked key milestones for the company.

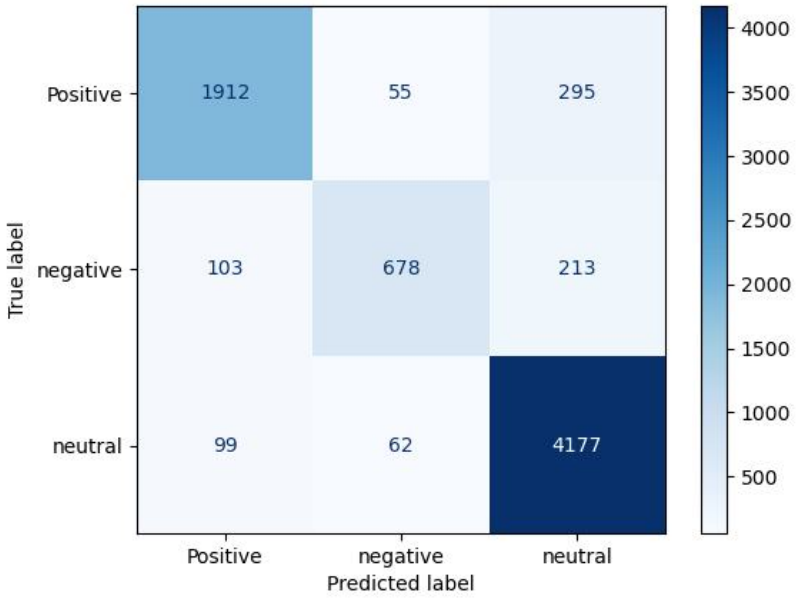

Figure 3.5: Confusion Matrix of Random Forest Model

With the proliferation of data generated on social media platforms like Twitter, traditional methods of data processing prove inadequate due to the sheer volume of information. Hence, there's a pressing need for innovative technical resources and methodologies to efficiently handle such data. Tesla's rapid growth parallels the expansion of the electric vehicle market. Notably, Tesla's acquisition of the NUMMI plant in Fremont, California, and its subsequent IPO on NASDAQ in 2010, signify significant milestones in its journey, making it the first American car company to do so since Ford Motor Company in 1956.

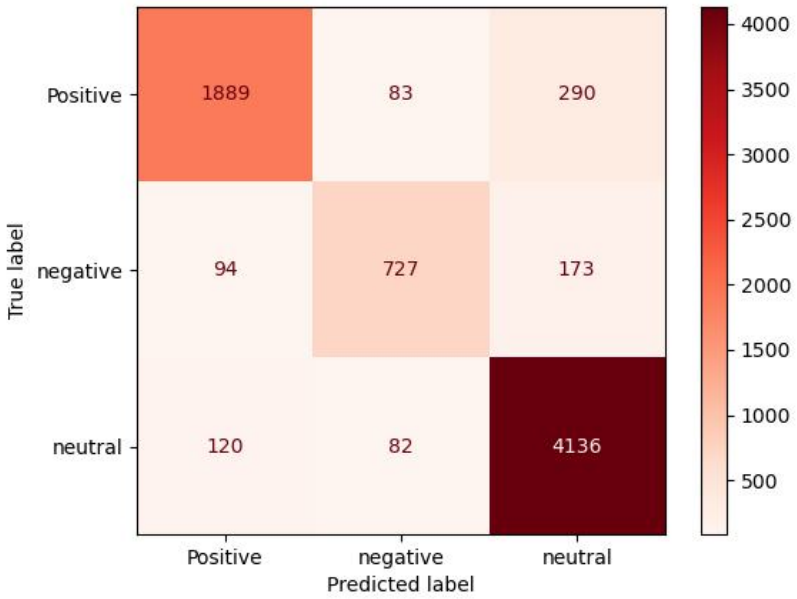

Figure 3.6: Confusion Matrix of Support Vector Machine

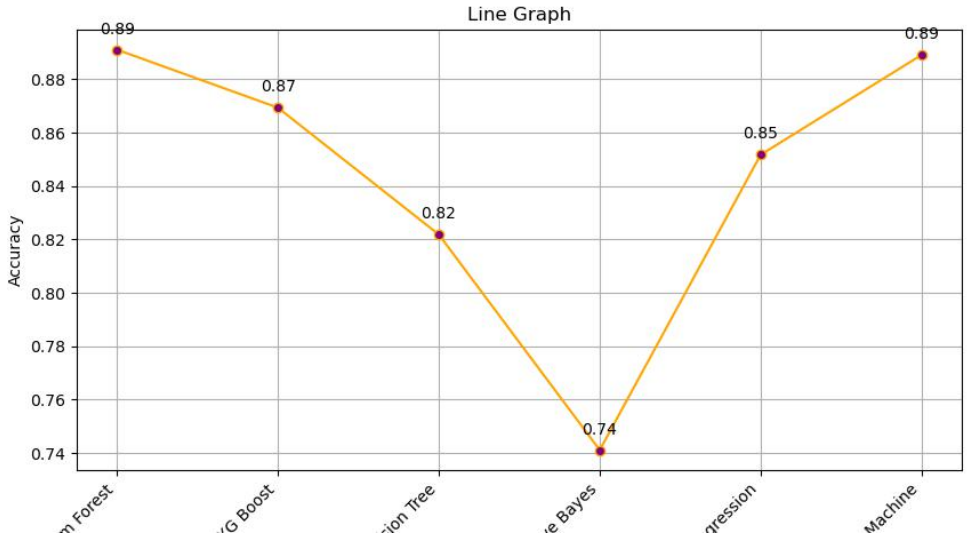

Figure 3.7: Accuracy of ML models

Chapter 4

CONCLUSION

Social media platforms have democratized communication, offering individuals from all walks of life a platform to express their opinions and reactions in real-time, potentially reaching millions of users. Twitter, in particular, stands out for its ability to facilitate rapid information gathering on current topics, allowing for the swift collection of public sentiments.

Our study employed six distinct machine learning models for sorting and prediction tasks, with Random Forest and SVM emerging as the top performers, boasting accuracy rates of 89.1 percent and 88.9 percent respectively. However, the absence of certain keywords in tweets resulted in the omission of some admissible content, potentially limiting the comprehensiveness of our dataset. Despite this, the sentiment analysis of early 2015 provided valuable insights, given the substantial number of opinions expressed about Tesla.

Looking ahead, our future research endeavors will explore additional text processing techniques, particularly in local languages, to capture a more diverse range of public opinions. By extending our analysis to encompass various information exchange platforms beyond Twitter, including Facebook, YouTube comments, and reputable newspapers, we aim to gain a more comprehensive understanding[6] of public sentiment.

Bibliography

[1] Yuan Xu. Evaluation & analysis of movie aspects: Based on sentiment analysis. *Frontiers in Management Science*, 2(3):64–116, 2023.

[2] María-Teresa Martín-Valdivia, Eugenio Martínez-Cámara, Jose-M Perea-Ortega, and L Alfonso Ureña-López. Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40(10):3934–3942, 2013.

[3] Andrzej Janusz, Guohua Hao, Daniel Ka luża, Tony Li, Robert Wojciechowski, and Dominik Ślezak. Predicting escalations in customer support: Analysis of data mining challenge results. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 5519–5526. IEEE, 2020.

[4] Nikmatul Husna Binti Suhendra, Pantea Keikhosrokiani, Moussa Pourya Asl, and Xian Zhao. Opinion mining and text analytics of literary reader responses: A case study of reader responses to kl noir volumes in goodreads using sentiment analysis and topic. In *Handbook of research on opinion mining and text analytics on literary works and social media*, pages 191–239. IGI Global, 2022.

[5] Steven Robert Hazen et al. *The impact of wolves on elk hunting in Montana*. PhD thesis, Montana State University-Bozeman, College of Agriculture, 2012.

[6] Sebastián Moreno Araya and Jorge Pereira Gude. Analysis of first-year university student dropout through machine learning models: A comparison between universities. 2023.

[7] Yiliu Paul Tu. E-business. new challenges and opportunities for digital-enabled intelligent future: 23rd wuhan international conference, whiceb 2024, wuhan, china, may 24–26, 2024, proceedings, part iii. 2024.

[8] Joao Paulo Goulart Pedrosa. Exploiting machine learning techniques to predict alzheimer's disease progression. 2021.

[9] Abdel Karim Kassem. *Intelligent system using machine learning techniques for security assessment and cyber intrusion detection*. PhD thesis, Université d'Angers, 2021.

[10] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. Citeseer, 2001.

[11] Mohamed Farouk. *Principal component pyramids using image blurring for non-linearity reduction in hand shape recognition*. PhD thesis, Dublin City University, 2015.

[12] Konstantinas Korovkinas. *Hybrid method for textual data sentiment analysis*. PhD thesis, Vilniaus universitetas, 2020.

[13] Min Dai. A hybrid machine learning-based model for predicting flight delay through aviation big data. *Scientific Reports*, 14(1):4603, 2024.

[14] Shawni Dutta, Samir Kumar Bandyopadhyay, and S Kumar Bandyopadhyay. Employee attrition prediction using neural network cross validation method. *International Journal of Commerce and Management Research*, 6(3):80–85, 2020.