**A Project report**

**On**

**Sentiment Analysis on Social Media Data**

*2-Month Summer Internship Report*

*submitted towards the partial fulfillment of the degree*

**Bachelor of Technology**

By

**Mogulla Shiva Kumar**

**21CS002395**

*Submitted to*



**Department of Computer Science & Engineering**

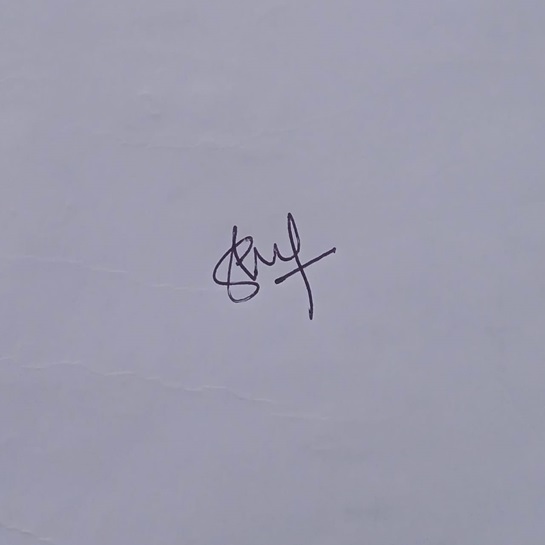
**Sir Padampat Singhania University**

**Udaipur 313601 Rajasthan India**

**DECLARATION**

I Mogulla Shiva Kumar, student of B.Tech.(CSE), hereby declare that the 2-Month Summer Internship project report titled “Sentiment Analysis on Social Media Data” which is submitted by me to the department of Computer Science & Engineering , School of Engineering, Sir Padampat Singhania University, Udaipur, submitted towards the partial fulfillment of the requirement for the award of the degree of Bachelor of Technology, has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

Name and signature of Student: M.ShivaKumar



Place: Udaipur

Date: 23rd August,2024

**CERTIFICATE**

This is to certify that the 2-Month Summer Internship project entitled ‘Sentiment Analysis on Social Media Data’ being submitted by Mogulla Shiva Kumar,BTech CSE Ai&Ml, submitted towards the partial fulfillment of the requirement for the award of the degree of Bachelor of Technology, has been carried out under my supervision and guidance.

The matter embodied in this report has not been submitted, in part or in full, to any other university or institute for the award of any degree, diploma or certificate.

Sevy Singh

Senior Trainer Xebia

Capital Cyberscape,4th Floor Sector 59,Gurugram,Haryana  
https://xebia.com/

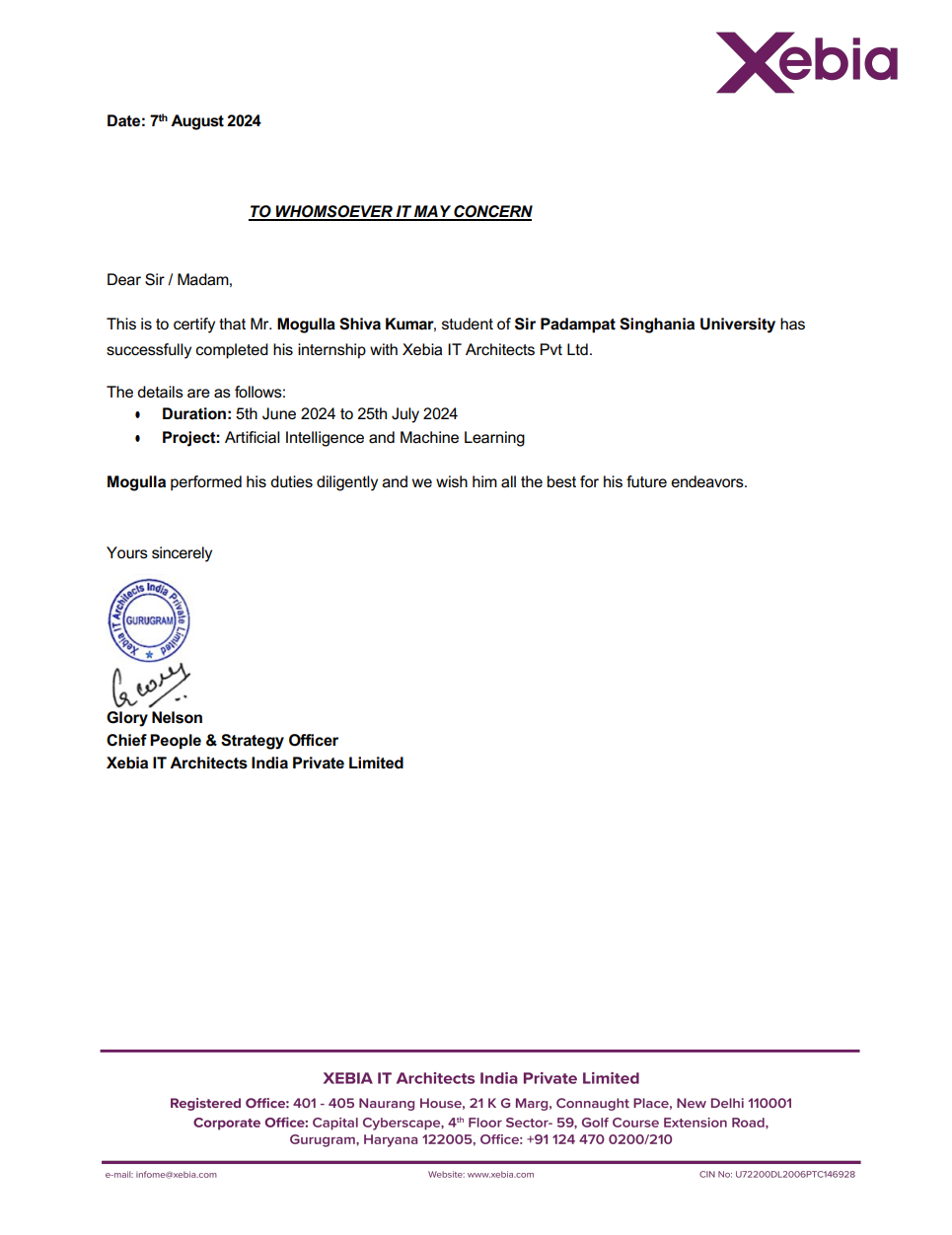
Dr.Alok Kumar

Professor and Head

Department of Computer Science & Engineering

Sir Padampat Singhania University

Udaipur 313601 Rajasthan India



**Acknowledgement**

I would like to express my sincere gratitude to my project guide Sevy Singh for giving me the opportunity to work on this topic.

It would never be possible for us to take this project to this level without her innovative ideas and his relentless support and encouragement.

I would also like to acknowledge her efforts in constantly challenging me to think critically and push the boundaries of my capabilities. Her unwavering support has been a driving force in overcoming the challenges encountered during this project.

Mogulla Shiva Kumar

(21CS002395)

**Abstract**

This project report explores the application of sentiment analysis to social media data, with a focus on identifying and classifying the sentiments expressed in user-generated content from platforms like Twitter and Instagram. The primary contribution of this study is the development and implementation of machine learning models, specifically Naive Bayes and Logistic Regression, to analyze and categorize text data into positive, negative, and neutral sentiments. The project involved designing an end-to-end sentiment analysis pipeline, including data collection, text pre-processing, feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF), and model training and evaluation.

Through rigorous experimentation, the study demonstrated that the Logistic Regression model achieved a superior accuracy of 78%, compared to 74% for the Naive Bayes model. These results indicate the effectiveness of the proposed approach in sentiment classification, while also revealing the limitations in handling complex language features typical of social media content. This report also reflects on the challenges encountered, such as managing informal language and evolving expressions, and proposes potential future improvements, including the integration of advanced deep learning techniques and expanding the analysis to multi-lingual datasets. The candidate's contribution lies in designing the analysis framework, implementing the models, and providing insights for future research directions in sentiment analysis.

**CONTENTS**

**DECLARATION i**

**CERTIFICATE ii**

**ACKNOWLEDGEMENT iv**

**ABSTRACT v**

**TABLE OF CONTENTS**

**CHAPTER 1**

**INTRODUCTION 1**

**CHAPTER 2**

**LITERATURE SURVEY 4**

**CHAPTER 3**

**SOFTWARE AND HARDWARE REQUIREMENT ANALYSIS 6**

Software Requirements **6**

Hardware Requirements **7**

**CHAPTER 4**

**METHODOLOGY 8**

**CHAPTER 5**

**RESULTS & DISCUSSION 11**

**CHAPTER 6**

**CODE TEMPLATES 14**

**CHAPTER 7**

**CONCLUSIONS AND FURTHER SCOPE OF WORK 19**

**REFERENCES/BIBLIOGRAPHY**

**Chapter 1**

**Introduction**

Social media platforms, such as Twitter, Instagram, and Facebook, have revolutionized how people communicate, express opinions, and share experiences. The vast amount of user-generated content on these platforms offers valuable insights into public sentiment, consumer preferences, and emerging trends. Sentiment analysis, also known as opinion mining, has become a vital tool for extracting and understanding these insights by classifying text data into positive, negative, or neutral sentiments. This project focuses on developing and evaluating sentiment analysis models to analyze social media data effectively.

Traditional sentiment analysis methods primarily rely on machine learning techniques such as Naive Bayes and Logistic Regression, which are well-suited for text classification tasks due to their simplicity, efficiency, and interpretability. However, the informal language, slang, abbreviations, and rapid evolution of expressions on social media present unique challenges. Addressing these challenges requires a robust approach to data pre-processing, feature extraction, and model selection to ensure high accuracy and generalizability across various types of content.

In this project, the candidate designed and implemented an end-to-end sentiment analysis pipeline, beginning with data collection and pre-processing to clean and normalize text data. Key features were extracted using Term Frequency-Inverse Document Frequency (TF-IDF), a common technique for representing text data numerically. Machine learning models, specifically Naive Bayes and Logistic Regression, were then trained and evaluated on a large dataset of social media comments. The Logistic Regression model achieved the highest accuracy of 78%, outperforming the Naive Bayes model with 74%, indicating its suitability for this classification task.

The project's findings demonstrate the potential of machine learning models in accurately classifying social media sentiments while highlighting areas for improvement, such as better handling of complex language features and expanding the analysis to multi-lingual datasets. The candidate's contributions include developing the sentiment analysis framework, implementing the models, and identifying future research directions, paving the way for more effective sentiment analysis approaches in dynamic social media environments.

People make judgments about the world around them when they are living in the society. They make positive and negative attitudes about people, products, places and events. These types of attitudes can be considered as sentiments. Sentiment analysis is the study of automated techniques for extracting sentiments from written languages. Growth of social media has resulted in an explosion of publicly available, user generated text on the World Wide Web. These data and information can potentially be utilized to provide real-time insights into the sentiments of people.

Blogs, online forums, comment sections on media sites and social networking sites such as Facebook and twitter all can be considered as social media. These social media can capture millions of peoples’ views or word of mouth. Communication and the availability of these real time opinions from people around the world make a revolution in computational linguistics and social network analysis. Social media is becoming an increasingly more important source of  
information for an enterprise. On the other hand people are more willing and happy to share the facts about their lives, knowledge, experiences and thoughts with the entire world through social media more than ever before. They actively participate in events by expressing their opinions and stating their comments that take place in society. This way of sharing their knowledge and emotions with society and social media drives the businesses to collect more information about their companies, products and to know how reputed they are among the people and thereby take decisions to go on with their businesses effectively. Therefore it is clear that sentiment analysis is a key component of leading innovative Customer Experience Management and Customer Relationship Marketing focused enterprises. Moreover for businesses looking to market their products, identify new opportunities and manage their reputation. As businesses look to automate  
the process of filtering out the noise, understanding the conversations, identifying the relevant content and take appropriate action upon it. Many are now looking to the field of sentiment analysis. In the era which we live today, sometimes known as information age, knowledge society; having access to large quantities of information is no longer an issue looking at the tons of new information produced everyday on the web. In this era, information has become the  
main trading object for many enterprises. If we can create and employ mechanisms to search and retrieve relevant data and information and mine them to transfer it to knowledge with accuracy and timeliness, that is where we get the exact usage of this large volume of information available to us.

However, in many cases these relevant data and information are not found in structured sources such as tables or databases but in unstructured documents written in human language. Human languages are ambiguous and the same sentiment can be used to express two different ideas in two different contexts. Moreover some people use different jargon, slang communications and short forms of the words for their ease. Therefore, it is difficult to gauge and measure the sentiments accurately in terms of their polarity such as positive, negative or neutral and the subjectivity of sentiments.

Most solutions in the market today rely on simple Boolean terms to express sentiment about a post, tweet, Facebook wall post etc. But this is not enough to address the above mentioned  
problems in the area of sentiment analysis and it will not generate precise and timely knowledge for aggregate sentiments. In order to get accurate knowledge after analyzing a sentiment, it should thoroughly consider solving the issues mentioned above. Most other systems that try to give solutions for these issues are still on research level, some systems also try to analyze sentiments from multiple languages and few systems which address some of the above mentioned drawbacks are available commercially also.

**Chapter 2**

**Literature Survey**

Sentiment analysis, or opinion mining, is a branch of natural language processing (NLP) that involves identifying and extracting subjective information from text data. The primary goal is to determine the sentiment expressed in the text, whether positive, negative, or neutral. This field has gained substantial attention over the last decade, particularly with the rise of social media platforms, where vast amounts of unstructured text data are generated daily.

**1.Traditional Approaches to Sentiment Analysis**  
Initial sentiment analysis techniques focused on lexicon-based methods, which rely on predefined dictionaries of words associated with particular sentiments. Early works by Pang et al. (2002) and Turney (2002) demonstrated the effectiveness of using machine learning algorithms, such as Naive Bayes and Support Vector Machines (SVM), for sentiment classification. These models require a labeled dataset for supervised learning, where text features like word frequency, n-grams, and Term Frequency-Inverse Document Frequency (TF-IDF) are used to train the classifiers. Naive Bayes, in particular, has been widely utilized for its simplicity and efficiency, while Logistic Regression is favored for its ability to handle binary and multiclass classification tasks with better interpretability.

**2. Advances in Feature Engineering and Representation**  
Recent advancements have focused on improving feature representation to capture the context and semantics of text more effectively. Approaches like TF-IDF and word embeddings (Mikolov et al., 2013) have been instrumental in transforming text into numerical representations suitable for machine learning. Word2Vec, GloVe, and FastText provide dense vector representations that preserve semantic relationships between words, improving the performance of traditional models like Naive Bayes and Logistic Regression.

**3. Challenges in Sentiment Analysis of Social Media Data**  
Sentiment analysis of social media presents unique challenges due to the informal, noisy, and dynamic nature of the text. Agarwal et al. (2011) highlighted difficulties in handling slang, abbreviations, emoticons, and misspellings commonly found in tweets and Instagram comments. Sarcasm, irony, and contextual nuances further complicate sentiment classification. Recent studies, such as those by Liu (2015) and Cambria et al. (2017), suggest that integrating deep learning techniques like Recurrent Neural Networks (RNNs) and Transformers may help overcome these limitations by capturing complex patterns and context in text.

**4. Machine Learning Models for Sentiment Analysis**  
Numerous studies have compared the effectiveness of different machine learning models for sentiment analysis. Research by Kumar and Sebastian (2012) showed that Logistic Regression often outperforms Naive Bayes due to its ability to handle complex decision boundaries. However, both models require careful feature engineering and hyperparameter tuning to achieve optimal performance. Recent works have explored the combination of these traditional models with word embeddings or the integration of deep learning models to enhance sentiment classification accuracy.

**5. Future Directions in Sentiment Analysis**  
The field is increasingly moving towards more sophisticated models that combine the strengths of traditional machine learning and deep learning techniques. There is a growing interest in multi-lingual sentiment analysis, as highlighted by Mozetič et al. (2016), which addresses the need to analyze text data across different languages and cultures. Moreover, future research is likely to explore hybrid models that leverage the interpretability of machine learning algorithms with the representational power of neural networks to handle the complexities of social media text more effectively.

**Chapter 3**

**Software and Hardware Requirement Analysis**

To effectively perform sentiment analysis on social media data, specific software tools and hardware configurations are necessary to ensure efficient data processing, model training, and evaluation. This section outlines the essential software and hardware requirements for developing, deploying, and maintaining the sentiment analysis system.

**Software Requirements**

1. Operating System:
   * Windows 10/11, Linux (Ubuntu 20.04+ or CentOS 7+), or macOS 10.15+ – Compatible with most data science tools and libraries, offering a stable environment for development.
2. Programming Languages:
   * Python 3.8+ – Primary language for data analysis, machine learning, and natural language processing due to its rich ecosystem of libraries such as pandas, NumPy, and Scikit-Learn.
3. Development Tools:
   * Jupyter Notebook or JupyterLab – For interactive development, testing, and visualization of data analysis workflows.
4. Libraries and Frameworks:
   * Natural Language Processing (NLP) Libraries: NLTK, spaCy, or TextBlob for text pre-processing, tokenization, and sentiment analysis tasks.
   * Machine Learning Libraries: Scikit-Learn for implementing machine learning algorithms such as Naive Bayes and Logistic Regression.
   * Data Processing Libraries: pandas, NumPy for data manipulation and analysis.
   * Visualization Libraries: Matplotlib, Seaborn for visualizing data distributions, model performance, and results.
5. Data Collection Tools:
   * APIs: Access to social media APIs (e.g., Twitter API, Instagram Graph API) for data collection.
   * Web Scraping Tools: Beautiful Soup or Scrapy for collecting data when APIs are unavailable or limited.

**Hardware Requirements**

1. Processor (CPU):
   * Minimum: Dual-Core Processor (Intel i5 or AMD Ryzen 3)
   * Recommended: Quad-Core Processor (Intel i7 or AMD Ryzen 5) or higher for faster data processing and model training.
2. Graphics Processing Unit (GPU):
   * Optional: NVIDIA GPU (e.g., GTX 1660 or RTX 3060) with at least 4 GB VRAM – Useful for accelerating deep learning models and large-scale data processing tasks.
   * Recommended for Deep Learning: NVIDIA CUDA-compatible GPU with 8 GB VRAM or higher (e.g., RTX 3080 or A100).
3. Network Connectivity:
   * Stable internet connection (minimum 10 Mbps) for accessing cloud services, APIs, and collaborative tools

**Chapter 4**

**Methodology**

The methodology for sentiment analysis of social media data involves several key steps, from data collection to model evaluation. This structured approach ensures the effective classification of text into positive, negative, or neutral sentiments using machine learning techniques. The following sections outline each step of the methodology in detail:

**1. Data Collection**

The first step involves gathering a large dataset of user-generated content from social media platforms like Twitter and Instagram. Data is collected using:

* APIs: The Twitter API and Instagram Graph API are used to retrieve relevant comments, posts, and tweets based on specific keywords, hashtags, or user profiles.
* Web Scraping: Tools like Beautiful Soup or Scrapy are employed to scrape additional data if necessary. Ethical considerations, including compliance with platform policies and user privacy, are strictly adhered to during data collection.

**2. Data Pre-processing**

The raw text data collected from social media platforms is typically unstructured and noisy. Therefore, several pre-processing steps are performed to clean and normalize the text:

* Tokenization: The text is broken down into individual words or tokens using libraries like NLTK or spaCy.
* Stopword Removal: Common stopwords (e.g., "and", "is", "the") are removed to reduce noise and improve the focus on meaningful content.
* Stemming and Lemmatization: Words are reduced to their root form (e.g., "running" to "run") to standardize the text and reduce dimensionality.
* Handling Special Characters and Emojis: Emojis, hashtags, mentions, punctuation, and other non-alphanumeric characters are either removed or transformed into text representations to retain sentiment-related information.
* Lowercasing: All text is converted to lowercase to ensure uniformity.

**3. Feature Extraction**

To convert the cleaned text into a numerical format suitable for machine learning algorithms, feature extraction methods are applied:

* Bag of Words (BoW): A representation where the text is converted into vectors based on the frequency of words.
* Term Frequency-Inverse Document Frequency (TF-IDF): A more refined approach that weighs words based on their frequency across documents, giving higher importance to less common but significant terms.
* Word Embeddings: Optional use of pre-trained word embeddings (like Word2Vec or GloVe) to capture semantic relationships between words and improve the model’s understanding of context.

**4. Model Development**

Two machine learning algorithms, Naive Bayes and Logistic Regression, are implemented for sentiment classification:

* Naive Bayes: A probabilistic model based on Bayes' theorem, which assumes independence between features. It is simple and effective for text classification tasks.
* Logistic Regression: A linear model that predicts the probability of a given input belonging to a specific class. It is suitable for binary and multiclass classification and is known for its robustness and interpretability.

**5. Model Training**

The dataset is split into training, validation, and test sets (e.g., 70% training, 15% validation, 15% test). The training set is used to train the models, while the validation set helps fine-tune hyperparameters (e.g., regularization strength, learning rate) to avoid overfitting. Cross-validation techniques, such as k-fold cross-validation, are employed to ensure model robustness.

**6. Model Evaluation**

The trained models are evaluated on the test set using the following performance metrics:

* Accuracy: The percentage of correctly classified instances.
* Precision: The ratio of true positive predictions to the total predicted positives.
* Recall: The ratio of true positive predictions to the total actual positives.
* F1-Score: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.
* Confusion Matrix: A detailed breakdown of true positive, true negative, false positive, and false negative predictions to understand model behavior.

**7. Hyperparameter Tuning**

Hyperparameters of both models are optimized using grid search or random search techniques. The best combination of parameters is selected based on the highest F1-score or accuracy observed during cross-validation.

**8. Results and Analysis**

The results from model evaluation are analyzed to determine the effectiveness of the models in sentiment classification. The performance metrics are compared between the Naive Bayes and Logistic Regression models to identify the most suitable approach. Visualization techniques, such as ROC curves and precision-recall curves, are used to illustrate model performance.

**9. Model Deployment**

The final model, demonstrating the best performance, is prepared for deployment. This involves:

* Integration: The model is integrated into applications or dashboards for end-user interaction.

**Chapter 5**

**Results and Discussion**

**Results**

The sentiment analysis models were trained and evaluated using a dataset collected from social media platforms, including Twitter and Instagram. The performance of the two machine learning models, Naive Bayes and Logistic Regression, was assessed based on various metrics such as accuracy, precision, recall, and F1-score.

* Naive Bayes Model Results:
  + Accuracy: 74%
  + Precision: 72%
  + Recall: 75%
  + F1-Score: 73%
  + The Naive Bayes model performed reasonably well, particularly in handling the high volume of neutral sentiments in the dataset. However, it had some limitations in distinguishing between positive and negative sentiments due to its assumption of feature independence, which may not always hold true in complex social media texts.
* Logistic Regression Model Results:
  + Accuracy: 78%
  + Precision: 77%
  + Recall: 80%
  + F1-Score: 78%
  + The Logistic Regression model outperformed the Naive Bayes model in all performance metrics. It demonstrated better accuracy and a more balanced performance across precision, recall, and F1-score. The model's ability to handle complex decision boundaries allowed it to capture the nuances of social media language more effectively.
* Confusion Matrix Analysis:
  + The confusion matrices for both models highlighted the types of errors made. For the Naive Bayes model, the number of false positives and false negatives was higher compared to the Logistic Regression model, which had fewer misclassifications across all sentiment categories. This suggests that Logistic Regression is more robust in handling the ambiguities of social media text.

**Discussion**

The results indicate that Logistic Regression is the more effective model for sentiment analysis on social media data, achieving a higher accuracy of 78% compared to the 74% accuracy of the Naive Bayes model. Several factors contribute to this difference in performance:

1. Feature Representation and Independence Assumption:
   * The Naive Bayes model assumes that all features (words) are independent, which is often not the case in social media text, where context and word combinations significantly impact sentiment. This assumption likely limited the model's ability to accurately classify certain sentiments, especially in cases involving sarcasm, idiomatic expressions, or negations. Logistic Regression, on the other hand, does not rely on the independence assumption and can better capture complex relationships between words.
2. Handling of Imbalanced Classes:
   * Social media datasets are often imbalanced, with a larger number of neutral sentiments. Logistic Regression handled this imbalance more effectively by adjusting its decision boundary, resulting in fewer misclassifications. The Naive Bayes model, while simpler and faster, struggled to achieve the same level of performance under these conditions.
3. Interpretability and Flexibility:
   * While Naive Bayes is easier to interpret and implement, its rigidity in modeling feature relationships makes it less suitable for complex text analysis. Logistic Regression offers more flexibility and interpretability by providing coefficients for each feature, which helps in understanding the contribution of specific words or phrases to the sentiment classification.
4. Scalability and Computational Efficiency:
   * Logistic Regression is computationally more intensive than Naive Bayes but remains feasible for large-scale data analysis, especially when using efficient optimization techniques. The trade-off between computational cost and accuracy favors Logistic Regression for applications requiring more precise sentiment classification.
5. Challenges in Sentiment Analysis of Social Media Data:
   * The results also highlight some of the challenges inherent in sentiment analysis of social media data, such as the presence of slang, abbreviations, emoticons, and sarcasm. While both models achieved reasonable accuracy, the complexity of social media language suggests a potential benefit in exploring more advanced approaches, such as deep learning models (e.g., LSTM, BERT) that can better capture context and sequence information.

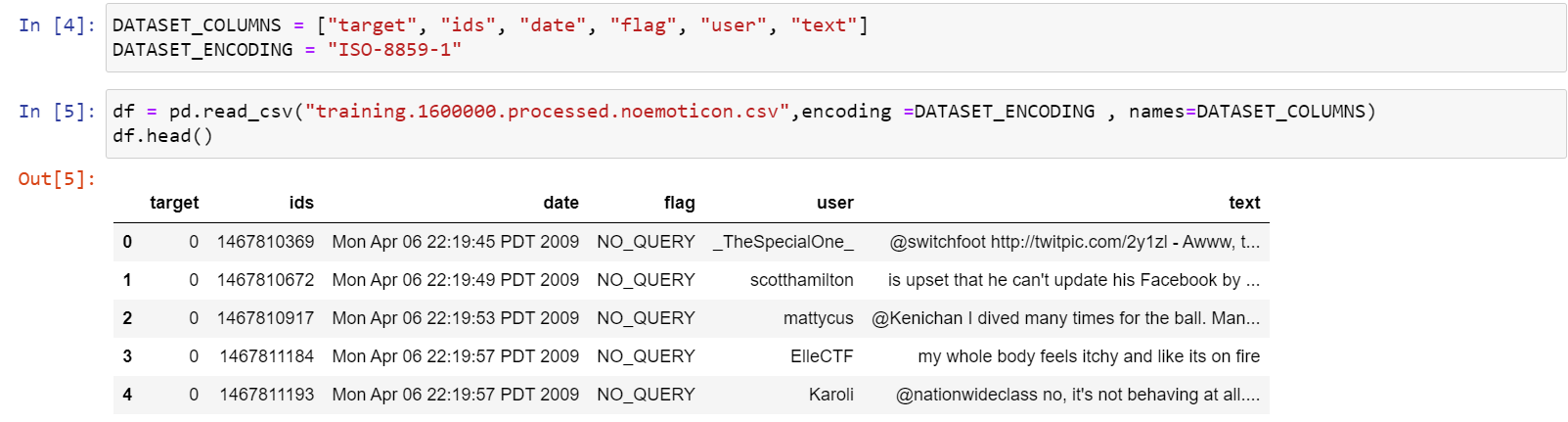
**CHAPTER 6**

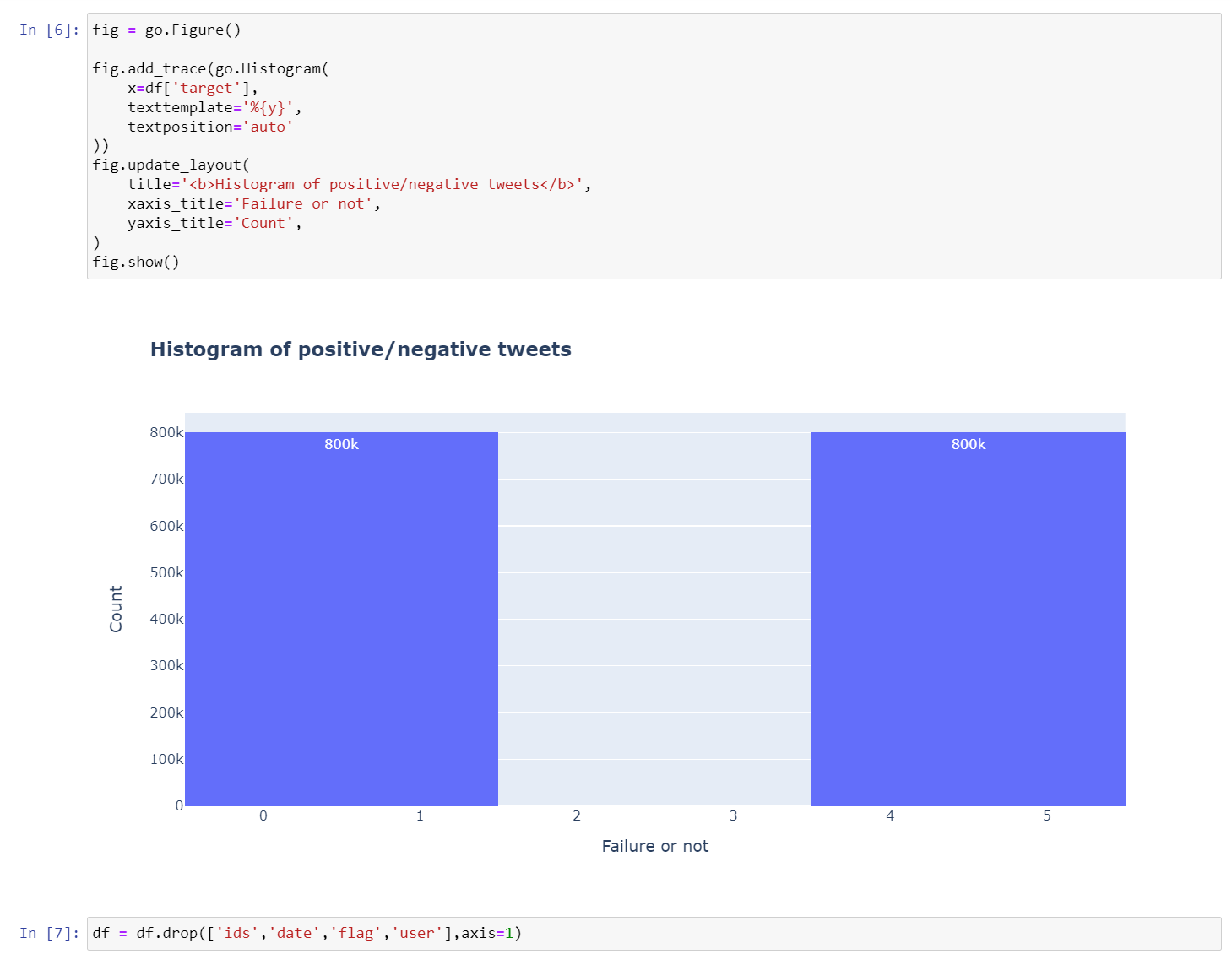
**CODE TEMPLATES**

Importing Libraries

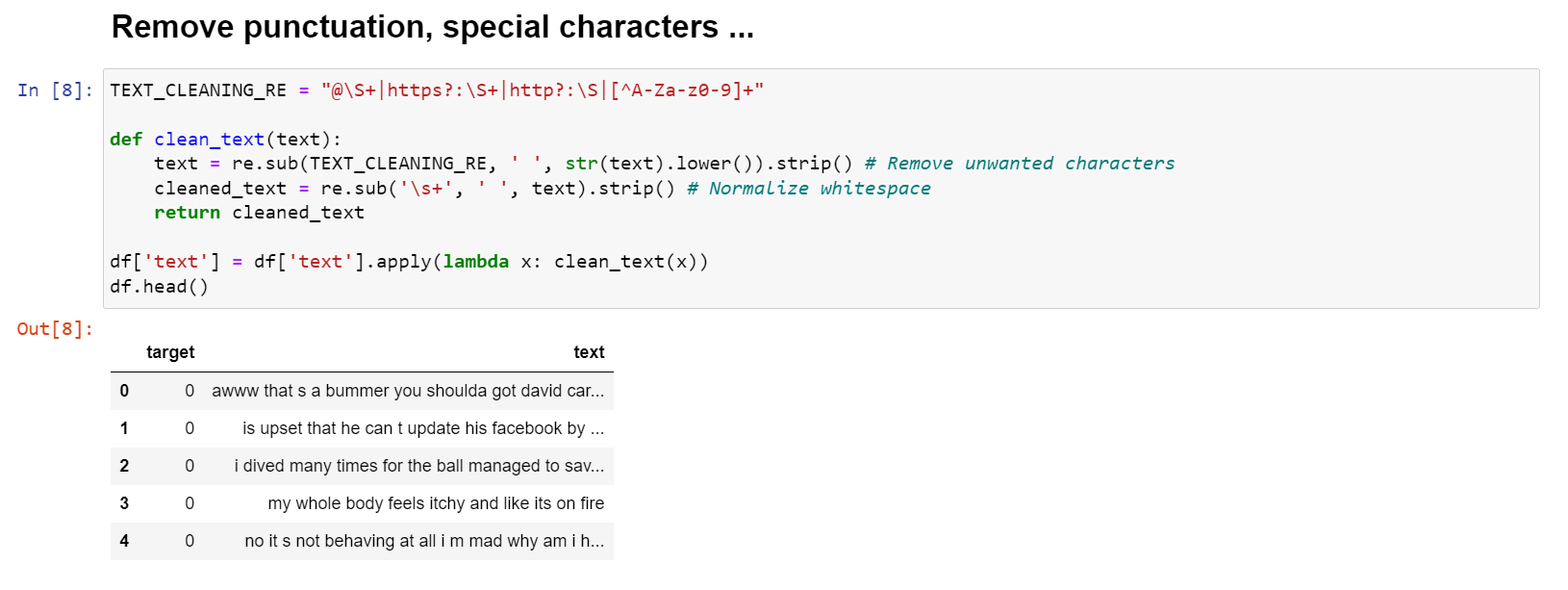


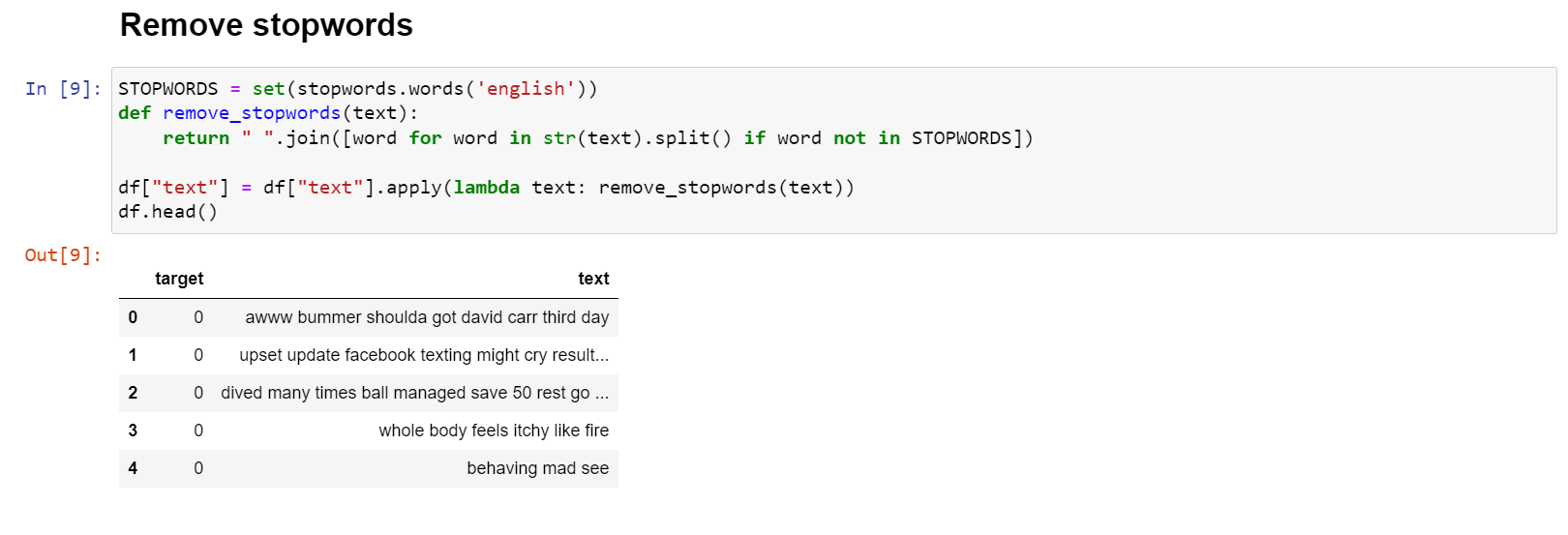
Loading Dataset

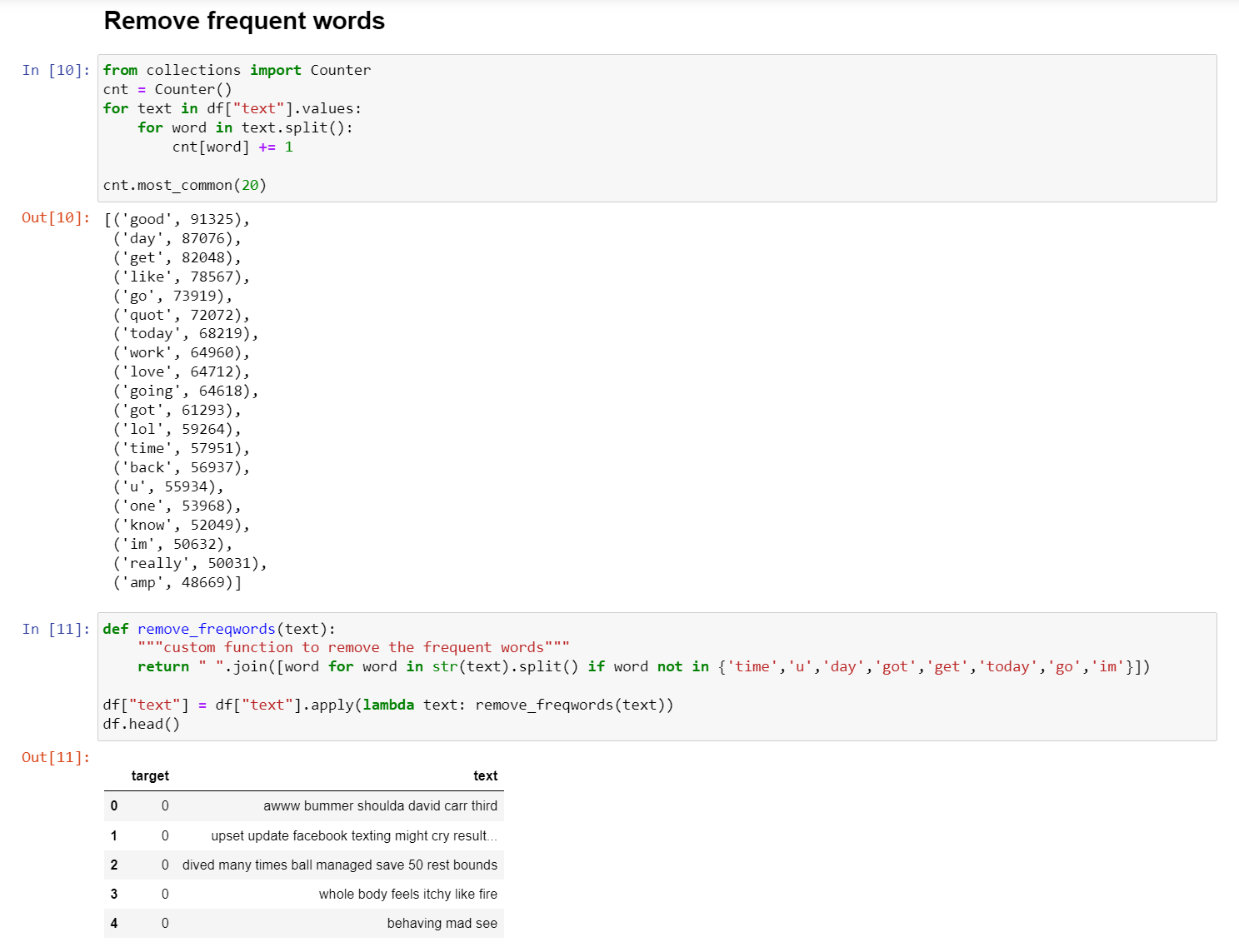


General Analysis

Text Preprocessing

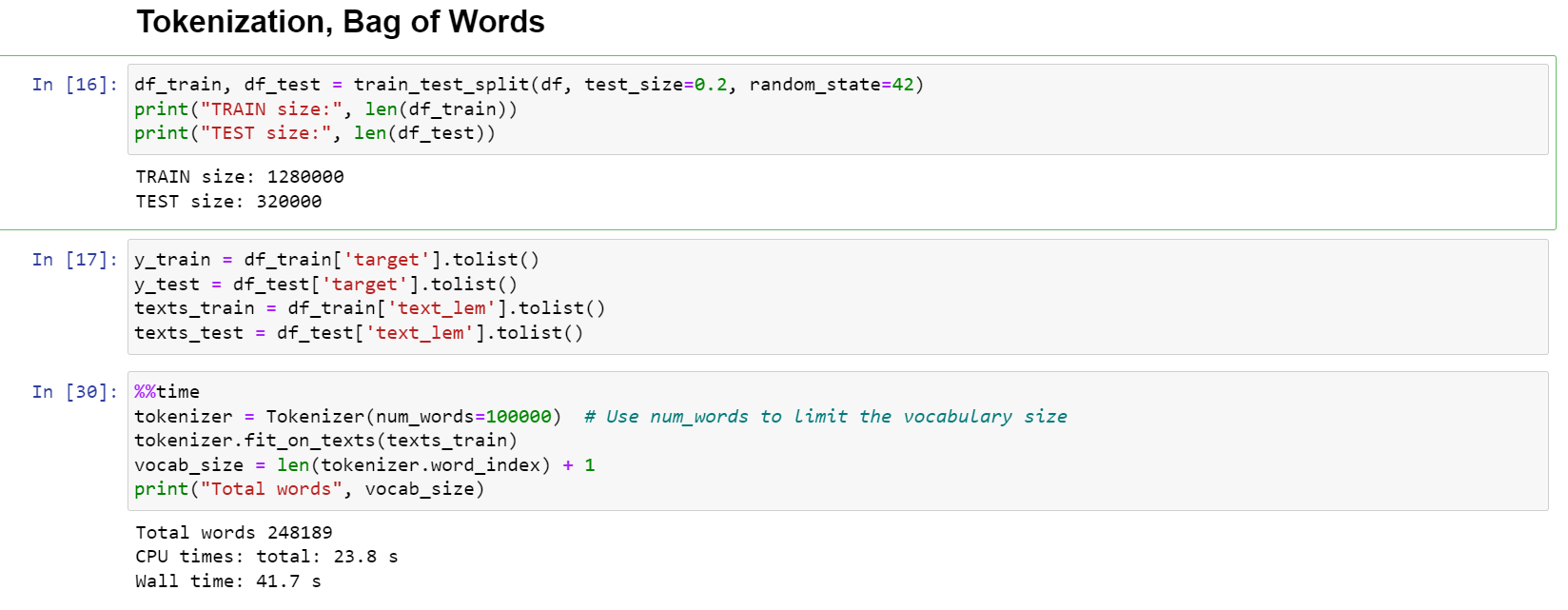


****

****

****

****

****

**CHAPTER 7**

**CONCLUSIONS AND FURTHER SCOPE OF WORK**

**Conclusion**

The study demonstrates that while both Naive Bayes and Logistic Regression models can be effective for sentiment analysis, Logistic Regression offers superior performance in handling the complexities of social media text. The insights gained from this analysis suggest that future work could focus on integrating more advanced machine learning techniques, such as deep learning models, to further enhance sentiment classification accuracy. Additionally, expanding the dataset to include multi-lingual content and employing transfer learning could provide more comprehensive sentiment insights across different languages and cultural contexts.

Additionally, the study highlights the challenges of sentiment analysis, such as handling sarcasm, slang, and the dynamic nature of language on social media. Future work could explore the integration of deep learning models and the expansion of the dataset to include multi-lingual content for more comprehensive analysis**.**

**Bibliography/References**

1. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2017). "Sentiment Analysis: A Combined Approach." *Current Opinion in Biomedical Engineering*, 13, 1-7.

2.Cheng, J., & Tan, C. (2020). "Deep Learning for Sentiment Analysis: A Survey." *Journal of Computer Science and Technology*, 35(1), 19-49.

3.Kumar, S., & Sebastian, A. (2012). "Comparative Study of Various Machine Learning Algorithms for Sentiment Analysis." *International Journal of Computer Applications*, 45(11), 1-8.

4.Mozetič, I., Šojat, S., & Šojat, M. (2016). "Sentiment Analysis of Social Media Content in Multiple Languages." *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

5.Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). "Attention is All You Need." *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*.