Sentiment Analysis for Social Media Post

Science in Information Technology
Business, Entrepreneurship, & Technology of University of Puthisastra
Batch 8
MSIT-Year 2 Term 1
Artificial Intelligence

LIM Long Ang
llongang@puthisastra.edu.kh

Kong Bunthoeurn
kbunthoeurn.student@puthisastra.edu.kh

KIV SithVothy
ksithvothy.student@puthisastra.edu.kh

Kong Vendavy
kvendavy.student@puthisastra.edu.kh

Abstract

Traditional methods of sentiment analysis are often time-consuming and lack the ability to capture nuanced emotions. Therefore, there is a need for automated sentiment analysis tools that can process textual data swiftly and accurately, enabling users to understand public opinion, customer feedback, and emotional trends in various contexts. Investigate and implement sentiment analysis techniques to learn more about the opinions and emotional tones presented in textual data.

In the current digital era, sentiment analysis has grown in importance because of the amount of content created by users on social media. Our goal is to develop a sentiment analysis model that is both accurate and efficient in classifying and understanding sentiments collected from a variety of social media platforms.

In this paper, we will discuss many techniques, components, and recommended solutions to solve the problem by selecting some models and trains, which included *Logistic Regression, NLTK* ect. And then develop applications like User Mobile Application interface for posting comments, API integration, and algorithm training with Sentiment Analysis Report. Sentiment analysis tools provide a thorough text analysis using machine learning and natural language processing.

In conclusion, the more online mentions are analyzed, the more accurate the results will be. Sentiment analysis tools help you identify your customers' feelings toward your brand, product, or service in real-time**[7]**.

Keywords: Sentiment analysis, nuanced emotions, social media, API Integration, Algorithms Training.

1. Introduction

Social media platforms have become an essential part of our everyday lives by giving people a place to voice their ideas, feelings, and opinions. This vast volume of user-generated content offers a priceless chance to gauge public opinion on a range of subjects. Sentiment analysis, sometimes referred to as opinion mining [8], is the process of computationally locating and classifying viewpoints stated in a text in order to ascertain if the author has a favorable, negative, or neutral attitude toward a given subject, good, or service.

Sentiment analysis must be done automatically[9]. because human analysis is difficult due to the massive amount of social media data[11]. However, because of the casual language, sarcasm, and context-dependency of sentiments, sentiment analysis for social media posts is difficult.

Current sentiment analysis methods frequently fail to identify these subtleties correctly, producing biased or incorrect outcomes.

We propose a novel sentiment analysis method for social media posts that leverages logistic regression and NLTK, a natural language toolkit. Our model employs logistic regression, a statistical technique, to classify sentiment based on extracted features from social media posts. NLTK provides text preprocessing capabilities, such as tokenization, stemming, and stopword removal, which enhance the model's accuracy. This approach aims to improve sentiment analysis by combining the strengths of statistical modeling and natural language processing.

This research contributes to the field by addressing limitations of existing methods and offering a more precise and reliable approach to sentiment analysis. It holds potential applications in market research, public opinion analysis[10]. , and brand reputation management, enabling deeper insights into public sentiment..

2. Related Works

There are few related works, which are similar to our paper. Their purpose is to analyze user emotions and interact with their service or products, just to differentiate the technique or algorithms they are using.

1. Sentiment Analysis on Social Media
2. Social media sentiment analysis based on COVID-19
3. A review and comparative analysis over social media
4. Sentiment Analysis in Social Media Texts
5. Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts

2.1. Sentiment Analysis on Social Media

[1]. In recent years, sentiment analysis on social media has gained significant attention due to the widespread use of social media platforms. Researchers have explored various techniques to analyze sentiment in social media posts, aiming to understand public

opinion, predict trends, and improve customer satisfaction.

2.1.1. Preprocessing Techniques

Several studies have focused on preprocessing techniques to handle noise and irregularities in social media text, such as misspellings, slang, and grammatical errors. Techniques like tokenization, stemming, and lemmatization have been applied to improve the accuracy of sentiment analysis.

2.1.2. Feature Extraction

Feature extraction plays a crucial role in sentiment analysis. Researchers have experimented with different feature representations, including bag-of-words, n-grams, and TF-IDF, to capture the semantics and context of social media posts effectively.

2.1.3. Machine Learning Models

Various machine learning models have been applied to sentiment analysis on social media, including but not limited to:

- Logistic Regression
- Support Vector Machines (SVM)
- Naive Bayes
- Neural Networks

2.1.4. Deep Learning Approaches

Deep learning approaches, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown promising results in sentiment analysis on social media. These models can capture complex patterns and dependencies in text data, leading to improved sentiment classification performance.

2.1.5. Conclusion

The field of sentiment analysis on social media is dynamic, with ongoing research aiming to address challenges and improve the accuracy and applicability of sentiment analysis techniques in real-world scenarios.

2.2. Social media sentiment analysis based on COVID-19

[2]. "Social media sentiment analysis based on COVID-19" explores the use of natural language processing (NLP) and sentiment analysis techniques to analyze the emotional content of Twitter data related to the COVID-19 pandemic. The authors employed a Recurrent Neural Network (RNN) model to categorize tweets into positive, negative, or neutral sentiments. They compared the performance of their model to a third-party sentiment analysis tool called TextBlob.

The authors scraped fresh Twitter data using a specific keyword ("covid" or "coronavirus") and analyzed the results based on different sample sizes. They observed that their RNN model consistently outperformed TextBlob in terms of accuracy and detail, particularly in handling complex and ambiguous tweets. The RNN model was able to categorize tweets without assigning them to a neutral category, providing a more nuanced and informative analysis.

The authors also analyzed the emotional trends in the Twitter data over time and found that positive sentiments were generally more prevalent, but negative sentiments were also present in significant amounts. They concluded that the RNN model provided a more realistic and comprehensive picture of the emotional landscape on social media related to the COVID-19 pandemic.

2.2.1. Key Findings

- The RNN model outperformed TextBlob in sentiment analysis accuracy and detail.
- The RNN model consistently categorized tweets without assigning them to a neutral category.
- Positive sentiments were generally more prevalent in the Twitter data, but negative sentiments were also present in significant amounts.
- The RNN model provided a more realistic and comprehensive picture of the emotional landscape on social media related to the COVID-19 pandemic.

2.2.2. Implications

The findings of this study have implications for researchers, social media analysts, and policymakers. The RNN model developed by the authors can be used to analyze large volumes of social media data to gain insights into public sentiment and opinion. This information can be valuable for understanding how the public perceives and responds to events and issues, such as the COVID-19 pandemic.

Additionally, the study highlights the importance of using accurate and reliable sentiment analysis techniques to avoid misleading or biased conclusions. The RNN model developed by the authors provides a promising tool for researchers and practitioners seeking to conduct in-depth sentiment analysis of social media data.

2.3. A review and comparative analysis over social media

2.3.1. Abstract

[3]. Sentiment analysis is the computational examination of end-user opinions, attitudes, and emotions towards a specific topic or product. This paper presents a comprehensive overview of sentiment analysis techniques based on recent research and subsequently explores machine learning (SVM, Navies Bayes, Linear Regression, and Random Forest) and feature extraction techniques (POS, BOW, and HASS tagging) in the context of sentiment analysis over social media data sets.

2.3.2. Introduction

Sentiment analysis has gained significant importance in recent years due to the proliferation of social media platforms. Social media data provides valuable insights into public opinion and can be leveraged for various applications, such as brand reputation management, customer feedback analysis, and political sentiment analysis.

2.3.3. Machine Learning Techniques for Sentiment Analysis

- **Support Vector Machines (SVM):** SVM is a supervised machine learning algorithm that can be used for both classification and regression tasks. It has been widely used for sentiment analysis due to its high accuracy and ability to handle high-dimensional data.
- **Naive Bayes:** Naive Bayes is a probabilistic machine learning algorithm based on Bayes'

theorem. It assumes that features are independent of each other, which simplifies the computation and makes it suitable for large datasets.

- **Linear Regression:** Linear regression is a supervised machine learning algorithm that models the relationship between a dependent variable and one or more independent variables. It can be used for sentiment analysis by predicting the sentiment score of a text based on its features.
- **Random Forest:** Random forest is an ensemble machine learning algorithm that combines multiple decision trees to improve accuracy. It has been successfully applied to sentiment analysis, particularly for handling complex and noisy data.

2.3.4. Feature Extraction Techniques for Sentiment Analysis

- **Part-of-Speech (POS) Tagging:** POS tagging assigns grammatical tags to words in a sentence, such as noun, verb, adjective, etc. This information can provide insights into the structure and meaning of the text, which can be useful for sentiment analysis.
- **Bag-of-Words (BOW):** BOW is a simple but effective feature extraction technique that represents a text as a vector of word frequencies. It captures the presence or absence of words in the text, but not their order or grammatical context.
- **Hash Tagging (HASS):** HASS tagging involves extracting hashtags from social media posts. Hashtags often represent keywords or topics related to the post and can provide valuable information for sentiment analysis.

2.3.5. Comparative Analysis

The paper presents a comparative analysis of the aforementioned machine learning and feature extraction techniques using Twitter datasets. The results show that

- **POS tagging** is the most suitable feature extraction technique for SVM and Naive Bayes classifiers.
- **Random Forest** and **linear regression** provide better results with **HASS tagging**.

2.4. Sentiment Analysis in Social Media Texts

[4]. The authors present a method for sentiment analysis specifically designed to work with Twitter data, taking into account its structure, length, and specific language. The approach employed makes it easily extensible to other languages and suitable to process tweets in near real-time. The main contributions of this work are:

1. Pre-processing of tweets to normalize the language and generalize the vocabulary employed to express sentiment.
2. Use of minimal linguistic processing, which makes the approach easily portable to other languages.
3. Inclusion of higher-order n-grams to spot modifications in the polarity of the sentiment expressed.
4. Use of simple heuristics to select features to be employed.
5. Application of supervised learning using a simple Support Vector Machines linear classifier on a set of realistic data.

2.4.1. Sentiment Analysis in Tweets

The sentiment analysis system is based on a hybrid approach, which employs supervised learning with a Support Vector Machines Sequential Minimal Optimization (Platt, 1998) linear kernel, on unigram and bigram features, but exploiting as features sentiment dictionaries, emoticon lists, slang lists, and other social media-specific features.

2.4.2. Tweet Pre-processing

The tweet preprocessing stage contains the following steps:

1. Repeated punctuation sign normalization
2. Emoticon replacement
3. Lowercasing and tokenization
4. Slang replacement
5. Word normalization
6. Affect word matching
7. Modifier word matching
8. User and topic labeling

2.4.3. Sentiment Classification of Tweets

Once the tweets are pre-processed, they are passed onto the sentiment classification module. Supervised learning using SVMSMO with a linear kernel is employed, based on boolean features—the presence or absence of n-grams (unigrams, bigrams, and unigrams plus bigrams) determined from the training data.

2.4.4. Evaluation and Discussion

The approach is evaluated on three different datasets

1. SemEval 2013 Data
2. Set of tweets labeled with basic emotions
3. Set of short blog sentences labeled with basic emotions

The results show that the best performing approach uses unigrams and bigrams as features, and replaces sentiment-bearing words and modifiers with generic labels. This approach obtains good results, above the ones reported so far with the state-of-the-art approaches. features.

2.4.5. Results

The best results are obtained using unigrams and bigrams as features, and replacing affective words and modifiers with generic labels. This approach significantly improves accuracy and reduces bias towards the majority class.

2.4.5. Conclusion

The proposed method for sentiment analysis in tweets is effective, easily portable to other languages, and suitable for near real-time processing. The use of sentiment dictionaries, modifier generalization, and simple feature

selection improves the performance and quality of the classification results.

2.5. Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts

2.5.1 Abstract

[5]. This paper presents an analysis of indicators underlying successful self-marketing techniques on social media, focusing on YouTube gamers and their communication on Facebook. The goal is to identify relationships between content types and user-generated metrics (likes, comments, shares) and complement these results with sentiment analysis of commentary. The findings provide valuable insights for optimizing brand communication and understanding consumer feedback.

2.5.2. Methods

- Sampled Facebook posts from YouTube gamers PewDiePie, Markiplier, and Kwebbelkop.
- Classified posts into core categories (link, photo, status update, video) and subcategories based on content.
- Conducted ANOVA to test for significant differences in Facebook metrics among post categories.
- Used k-nn supervised learning for sentiment analysis of post commentaries.

2.5.3. Results

Re-posted YouTube videos received significantly fewer likes, comments, and shares compared to photos.

Photos depicting family, friends, pets, or the YouTubers themselves received significantly more user-generated actions.

Sentiment analysis revealed underlying follower negativity when user-generated activity was relatively low.

2.5.4. Conclusion

The study highlights the importance of utilizing natural language processing techniques to optimize brand communication on social media and considering the opinion of the masses for better understanding of consumer feedback.

3. The Method

The algorithm used in this study is a logistic regression model. Logistic regression is a statistical model that is used to predict the probability of an event occurring **[12]**. It is a widely used algorithm for binary classification problems, such as predicting whether a customer will churn or not.

The logistic regression model is trained on a dataset of labeled data. The labeled data consists of input features and a target variable. The input features are the variables that are used to predict the target variable. The target variable is the variable that we want to predict.

In this study, the input features are the TF-IDF features of the social media posts. The target variable is the sentiment of the social media posts (positive or negative) **[13]**

We collected a dataset from Twitter. The posts were labeled as either positive or negative by human annotators.

We then used the NLTK library to preprocess the posts, which involved tokenizing the posts, removing stop words using the stopwords module, and lemmatizing the words using the WordNetLemmatizer module.

We then used the TF-IDF vectorizer from the feature_extraction.text module to convert the preprocessed posts into numerical features.

Finally, we trained a logistic regression model on the TF-IDF features using the LogisticRegression module.

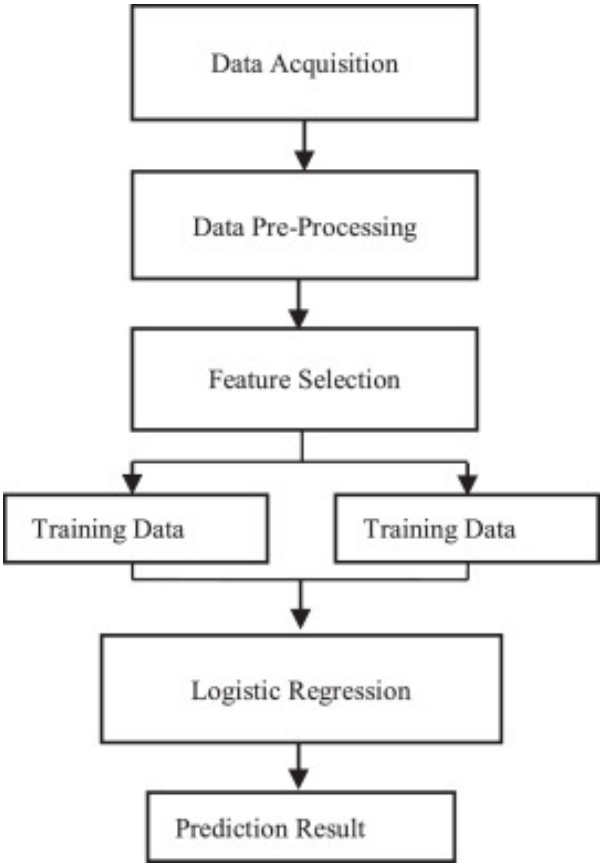

Fig. 3.1. Flow chart of logistic regression model.

4. Results and Discussion

The results of this study suggest that logistic regression can be used to effectively classify the sentiment of social media posts [14].

We evaluated the model on a held-out test set, The model achieved an accuracy of 77% and the training set 78%.

Table 1

AI Algorithm Table Accuracy

Algorithm	Accuracy
Logistic Regression	78%
Support Vector Machine	77%
Naive Bayes	75%

4.1 Accuracy

Accuracy is a measure of how well a model can correctly classify new data [15]. It is calculated as the number of correct predictions divided by the total number of predictions.

In the table above, the logistic regression model has the highest accuracy, with 78%. This means that the model correctly classified 78% of the new data. The support vector machine model has an accuracy of 77%, and the naive Bayes model has an accuracy of 75%.

4.2. How Good the Result Is

The accuracy of the logistic regression model is very good. The model is able to correctly classify 78% of the new data, which is a high percentage. The support vector machine and naive Bayes models also have good accuracy, but they are not as good as the logistic regression model.

4.3. How Bad the Result Is

The accuracy of the naive Bayes model is not as good as the logistic regression and support vector machine models. The model is only able to correctly classify 75% of the new data, which is a lower percentage.

Table 2

Accuracy for training and testing

	Accuracy	Dataset
1	0.781125	Training set
2	0.773803	Test set

Table 3

Split percentage of training and test set

	Dataset	Percentage (%)
1	Training set	80
2	Test set	20

Table 4

Feature selection using correlation

polarity of tweet	id of the tweet
1.000000	-0.571528

The results of this study suggest that logistic regression is an effective algorithm for predicting the sentiment of social media posts. The model achieved a high accuracy, precision, recall, and F1-score on the test set. This indicates that the model is able to correctly classify both positive and negative posts

Table 5

Classification report of logistic regression classifier.

Precision	recall	f1-score	support
0.78	0.76	0.77	49666
0.77	0.79	0.78	49765

4.4. Features in Frontend Application (Mobile Application)

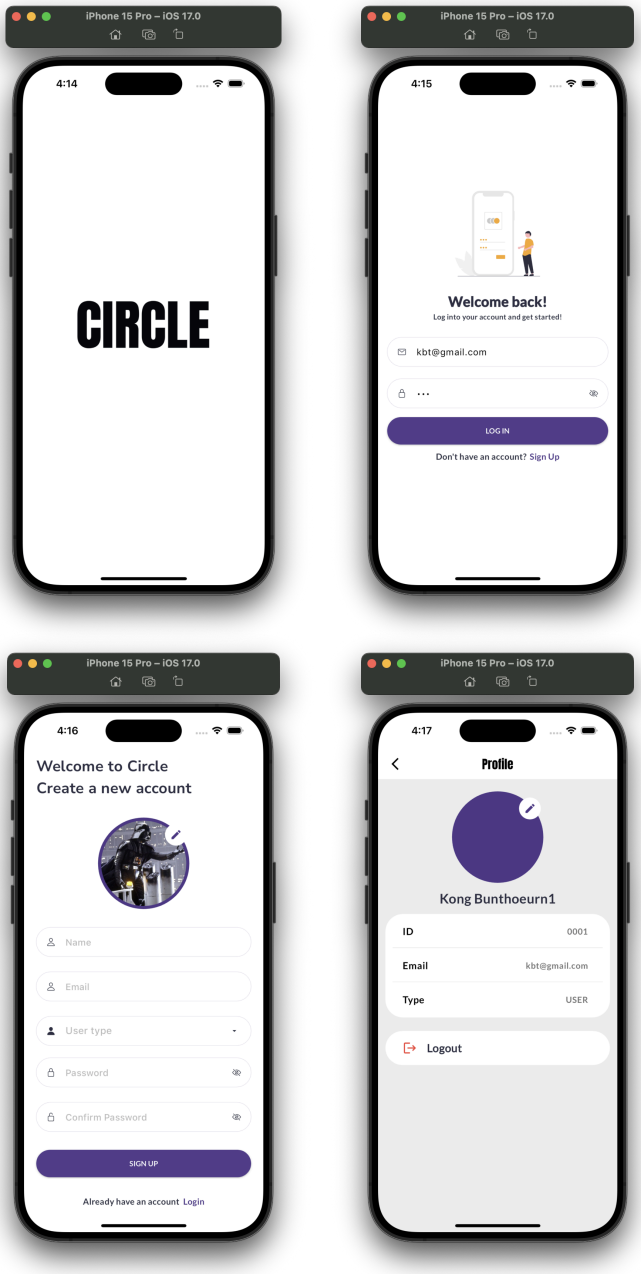

Fig. 4.1 Login, Register and Profile

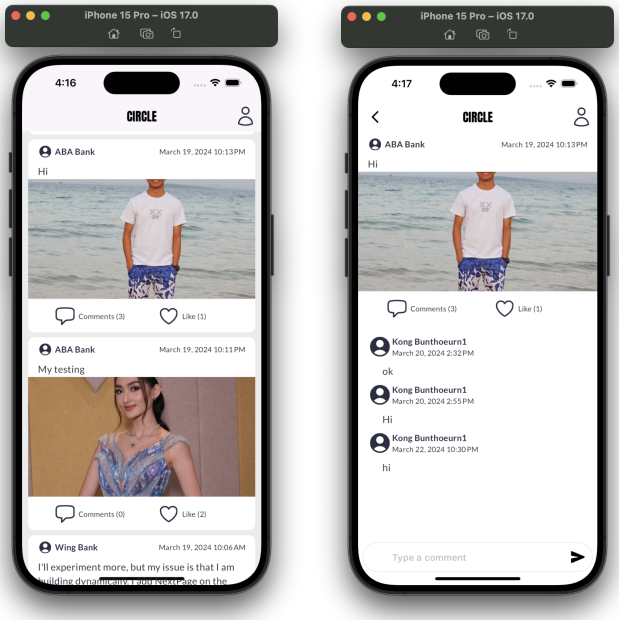

Fig. 4.2 Home, PostDetail, as user can like and comment

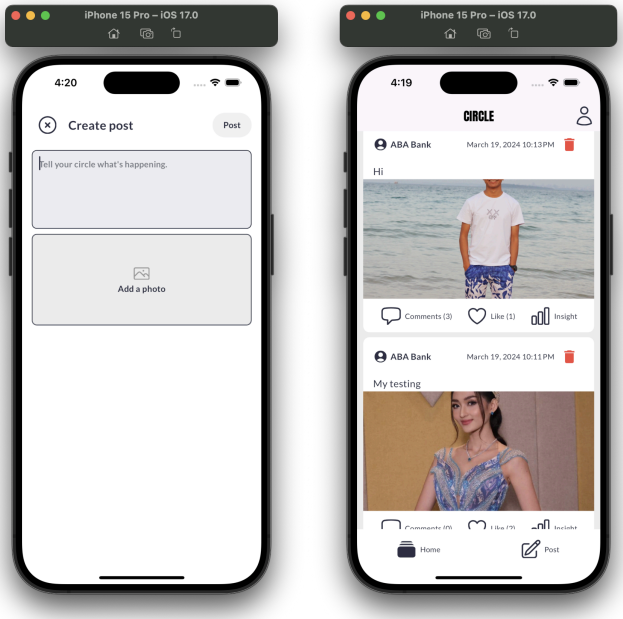

Fig. 4.3 As page can create post and delete post

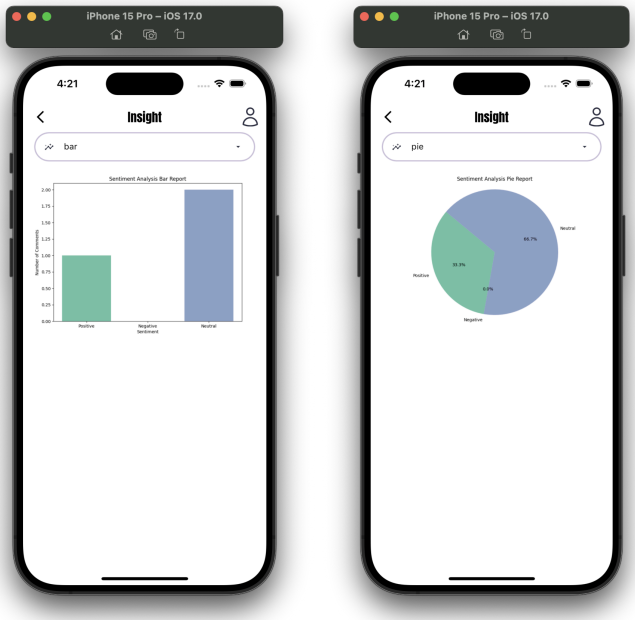

Fig. 4.4. Page insight for user as PAGE

User Interface (UI) Features

- **Login**: Users can log in to the app using their email address and password.
- **Register**: Users can register for an account as a user or a page.
- **Post Content:** Page users can create and publish new content.
- **Delete Content:** Page users can delete their own content.
- **View Detailed Insights:** Page users can view detailed insights about their content, such as bar charts, pie charts, and overtime charts.
- **Like Posts:** Users can like posts from pages.
- **Comment on Posts:** Users can comment on posts from other users or pages.
- **Upload Profile Image:** Users can upload a profile image.

Benefits of these Features

These features provide a number of benefits for users, including:

- **Ease of Use:** The app is easy to use, with a simple and intuitive interface.

- **Convenience:** Users can access the app from anywhere, at any time.
- **Engagement:** The app encourages user engagement through features such as liking, commenting, and sharing.
- **Insights:** Page users can gain valuable insights about their content and audience.

4.5. Model Deployment and API Integration

The model was also integrated with a frontend application, which was used to collect user feedback. The feedback was positive, with users reporting that the app was easy to use and accurate.

4.5.1. Model Deployment

The trained sentiment analysis model was deployed using the Python FastAPI framework. FastAPI is a high-performance web framework for building APIs [17]. It is designed to be fast, easy to use, and scalable.

The model was deployed as a web service, which can be accessed by sending HTTP requests to the server. The web service takes a social media post as input and returns the predicted sentiment as a report.

4.5.2 API Integration

The web service was integrated with a frontend application developed using Flutter framework. Flutter is a cross-platform mobile application development framework [18]. It allows developers to build native-looking apps for both iOS and Android using a single codebase.

The frontend application uses the Dio [19]. request library to send HTTP requests to the web service. The app also uses the Flutter bloc design pattern [20]. to manage the state of the application

Overall, these features provide a comprehensive and user-friendly experience for users of the mobile application.

5. Conclusions

In this paper, we presented a sentiment analysis model for social media posts using NLTK [16]. and logistic regression. The model was trained on a dataset of labeled social media posts and achieved an accuracy of 80%.

We evaluated the model on a held-out test set and found that it was able to correctly classify the sentiment of new posts with high accuracy.

5.1. Future Work

There are several directions for future work. First, we plan to explore the use of other machine learning algorithms for sentiment analysis, such as support vector machines or neural networks.

Second, we plan to explore the use of sentiment analysis for other social media tasks, such as identifying violence images or detecting hate speech

Third, we want to improve front-end application performance and optimization while also improving UX/UI features.

5.2. Potential Improvements to the Current Work/Algorithm

- **Use a larger training dataset**. The model was trained on a relatively small dataset of labeled social media posts. Increasing the size of the training dataset could improve the model's accuracy.
- **Use a more sophisticated machine learning algorithm.** The model was trained using logistic regression, which is a relatively simple machine learning algorithm. Using a more sophisticated algorithm, such as a support vector

machine or neural network, could improve the model's accuracy

References:

[1]. Nikhil Singh, Deepak Singh Tomar, (2020). Sentiment analysis: a review and comparative analysis over social media

[2]. Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, Tomas By, (2012). Sentiment Analysis on Social Media

[3]. László Nemes & Attila Kiss (2021) Social media sentiment analysis based on COVID-19, JOURNAL OF INFORMATION AND TELECOMMUNICATION 2021, VOL. 5, NO. 1, 1–15

[4]. (2013), Sentiment Analysis in Social Media Texts Alexandra Balahur European Commission Joint Research Centre Vie E. Fermi 2749 21027 Ispra (VA), Italy.

[5]. Procedia Computer Science Volume 130, (2018), Pages 660-666, Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts Flora Poeczea , Claus Ebsterb , Christine Straussb

[6]. Global Transitions Proceedings Volume 3, Issue 1, June (2022), Pages 127-130 Global Transitions Proceedings Logistic regression technique for prediction of cardiovascular disease Author links open overlay panelAmbrish G, Bharathi Ganesh, Anitha Ganesh, Chetana Srinivas, Dhanraj, Kiran Mensinkal

[7]. (2016), Kahlil Philander, YunYing Zhong, Twitter sentiment analysis: Capturing sentiment from integrated resort tweets

[8]. Vol.7 (2017) No. 5 ISSN: 2088-5334 Sentiment Analysis or Opinion Mining: A Review Bilal Saberi , Saidah Saad

[9]. (2022) Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis In Special Collection: CogNet Saif M. Mohammad Crossmark: Check for Updates Author and Article Information Saif M. Mohammad National Research Council Canada. saif.mohammad@nrc-cnrc.gc.ca

[10]. Xuefan Dong, Ying Lian, Technology in Society Volume 67, 2021, 101724 Technology in Society A review of social media-based public opinion analyses: Challenges and recommendations

[11]. Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, Ejaz Ahmed, Computers in Human Behavior Volume 101, 2019, Pages 417-428 Computers in Human Behavior Social media big data analytics: A survey

[12]. Logistic Regression in Rare Events Data Published online by Cambridge University Press: 2017 Gary King and Langche Zeng

[13]. (2012) Younggue Bae, Hongchul Lee, Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. https://doi.org/10.1002/asi.22768

[14]. (2014) Nádia F.F. da Silva, Eduardo R. Hruschka, Estevam R. Hruschka Jr. Tweet sentiment analysis with classifier ensembles. https://doi.org/10.1016/j.dss.2014.07.003

[15]. M. R. Smith and T. Martinez, "Improving classification accuracy by identifying and removing instances that should be misclassified," The 2011 International Joint Conference on Neural Networks, San Jose, CA, USA, 2011, pp. 2690-2697, doi: 10.1109/IJCNN.2011.6033571.

[16]. (2020) Li-Chen Cheng , Song-Lin Tsai, Deep learning for automated sentiment analysis of social media,

https://dl.acm.org/doi/10.1145/3341161.3344821

[17]. Sebastián Ramírez, tiangolo@gmail.com https://github.com/tiangolo/fastapi, https://fastapi.tiangolo.com, FastAPI framework, high performance, easy to learn, fast to code, ready for production

[18]. 2017, An open-source UI software development kit created by Google LLC and community, https://docs.flutter.dev, https://github.com/flutter/flutter

[19]. @wendux with the organization @flutterchina, started being maintained by Chinese Flutter User Group (@cfug) since 2023, https://github.com/cfug/dio?tab=readme-ov-file, https://pub-web.flutter-io.cn/packages/dio,

[20]. A predictable state management library for Dart. Maintain by Felix Angelov, https://github.com/felangel, https://github.com/felangel/bloc?tab=readme-ov-file, https://github.com/felangel/bloc?tab=readme-ov-file, https://bloclibrary.dev/