REDDIT DATA ANALYSIS WITH SENTIMENT AND EMOTION PREDICTION

Submitted by

SANSKAR SRIVASTAVA 20BCE1023

DEEPJYOTI KOLEY 20BCE1539

A project report submitted to

DR. TULASI PRASAD SARIKI

SCOPE

In fulfilment of the requirements for the course

of CSE4020 – Machine Learning

in

B.Tech Computer Science and Engineering

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Vandalur – Kelambakkam Road

Chennai – 600127

BONAFIDE CERTIFICATE

Certified that this project report entitled **"REDDIT DATA ANALYSIS WITH SENTIMENT AND EMOTION PREDICTION "** is a Bonafide work of **Deepjyoti koley (20BCE1539)** and **Sanskar Srivastava (20BCE1023)**who carried out the Project work under my supervision and guidance for **CSE4020 - Machine Learning.**

Dr. TULASI PRASAD SARIKI

SCOPE

VIT University, Chennai

Chennai – 600 127.

ABSTRACT

For our project, we built a system that collects live data by web crawling from Reddit using the site's API. Specifically, we focused on collecting the comments that users leave on each post. Once we collected this data, we used machine learning techniques to analyze the sentiment and emotional content of each comment.

Our system performs sentiment analysis to predict whether a comment expresses a positive, negative, or neutral sentiment. Additionally, we used emotional analysis to identify whether an emotion expressed in each comment is one of the common types, such as happiness, sadness, anger, or fear.

By analyzing the sentiment and emotional content of the comments on each post, our system can provide valuable insights into the attitudes and feelings of the users who are engaging with the content. This type of analysis can be useful for a wide range of applications, including market research, brand monitoring, and social media management.

Our project is a great example of how machine learning techniques can be used to extract meaningful insights from large datasets like those found on Reddit.

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr. Tulasi Prasad Sariki**, Associate Professor, SCOPE, for his consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the course.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

INTRODUCTION

Our project focuses on using machine learning techniques to analyze user sentiment and emotions on Reddit, a popular social media platform. Reddit is known for its large user base and diverse range of content, making it an excellent source of data for analyzing user behavior and attitudes. By collecting and analyzing the comments left by users, we can gain insights into the attitudes and feelings of the people who are engaging with the content on the site.

Our system uses Reddit's API to collect live data on user comments and applies machine learning algorithms to analyze the sentiment and emotional content of each comment. The sentiment analysis technique helps us to determine whether a comment is positive, negative, or neutral. Emotional analysis goes beyond sentiment analysis to identify specific emotions expressed in the text, such as happiness, sadness, anger, or fear. By analyzing the sentiment and emotional content of user comments, we can gain valuable insights into user attitudes and emotions, which can be useful for a wide range of applications, including market research, brand monitoring, and social media management.

Our project is a great example of how machine learning techniques can be used to extract meaningful insights from large datasets like those found on social media platforms. By analyzing the sentiment and emotional content of user comments, we can better understand user behavior and attitudes, helping businesses and organizations to make data-driven decisions.

MOTIVATION

Social media platforms like Reddit have become a popular source of data for businesses and organizations looking to understand user behavior and attitudes. By analyzing user sentiment and emotions on these platforms, businesses can gain valuable insights into how their products or services are perceived, what users like or dislike about them, and how they can improve the user experience. Sentiment analysis and emotional analysis are powerful tools that can help businesses to make data-driven decisions based on user feedback.

Our project aims to demonstrate the power of sentiment and emotional analysis on Reddit data using machine learning techniques. By analyzing the sentiment and emotional content of user comments on the site, we can provide valuable insights into user attitudes and emotions that can be useful for a wide range of applications, including market research, brand monitoring, and social media management. These insights can help businesses and organizations to better understand their target audience, identify trends and patterns in user behavior, and make data-driven decisions to improve their products or services.

Our project is motivated by the need for businesses and organizations to gain deeper insights into user behavior and attitudes on social media platforms like Reddit. By using sentiment and emotional analysis techniques on user comments, we can help businesses to make data-driven decisions that improve the user experience and drive business success.

PROBLEM STATEMENT

The problem we aim to address with our project is the challenge of understanding user sentiment and emotions on social media platforms like Reddit. With millions of users and an endless stream of content, it can be difficult for businesses and organizations to understand how their products or services are perceived by users, what users like or dislike about them, and how they can improve the user experience. Traditional methods of data analysis are often inadequate for handling the large volume of user-generated content on social media platforms, which can include text, images, and videos.

To address this problem, our project uses machine learning techniques to analyze the sentiment and emotional content of user comments on Reddit. By collecting live data from Reddit's API and applying sentiment and emotional analysis algorithms, we can provide valuable insights into user attitudes and emotions that can be used to inform data-driven decisions. By analyzing the sentiment and emotional content of user comments, we can help businesses and organizations to better understand their target audience, identify trends and patterns in user behavior, and make data-driven decisions to improve their products or services.

Our project addresses the problem of understanding user sentiment and emotions on social media platforms like Reddit by using machine learning techniques to analyze the sentiment and emotional content of user comments. By doing so, we aim to provide valuable insights that can be used to inform data-driven decisions, improve the user experience, and drive business success.

LITERATURE REVIEW

Sentiment analysis and emotional analysis are both active areas of research in the field of natural language processing and machine learning. A variety of techniques have been developed over the years for analyzing sentiment and emotions in text, including rule-based systems, machine learning models, and deep learning approaches.

One of the early and popular works in sentiment analysis is Pang and Lee's 2008 paper on Opinion mining and sentiment analysis. They discussed various techniques and methods for sentiment analysis, including lexicon-based approaches, machine learning approaches, and more.

Machine learning models use statistical methods to identify patterns in the data and make predictions about the sentiment and emotional content of text. These models can be trained on large datasets and can adapt to different types of text data, making them a popular choice for sentiment and emotional analysis. Socher et al.'s 2013 paper introduced recursive deep models for semantic compositionality over a sentiment treebank which achieved state-of-the-art performance in sentiment analysis.

Deep learning approaches, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have also been used for sentiment and emotional analysis. These models are able to learn complex representations of text data and can achieve state-of-the-art performance on a variety of sentiment and emotional analysis tasks. Wang et al.'s 2016 paper proposed an attention-based LSTM for aspect-level sentiment classification, which performed well in identifying sentiment towards specific aspects of products or services.

In the context of social media platforms like Reddit, sentiment and emotional analysis have been used for a variety of applications, including brand monitoring, market research, and social media management. Agarwal et al.'s 2011 paper discussed sentiment analysis of Twitter data, which is a similar platform to Reddit. Several studies have shown that sentiment analysis and emotional analysis can be used to identify trends and patterns in user behavior, understand the sentiment and emotional content of user feedback, and improve the user experience.

Sentiment analysis and emotional analysis are powerful techniques for analyzing text data, and they have many potential applications in various industries.

ARCHITECTURE

The architecture of our project includes the following components:

1. Reddit API: This is used to scrape data from the Reddit platform, including post and comment data related to specific topics.
2. PRAW (Python Reddit API Wrapper): This is a Python library that simplifies the interaction with the Reddit API and enables us to easily extract data from subreddits.
3. Data Preprocessing: We preprocess the scraped data by removing URLs and emojis from the comments, changing the time format, and merging the two DataFrames.
4. Visualization: We used graphs and a word cloud to visualize the data.
5. Sentiment Analysis: We used a pre-made sentiment analyzer to determine the sentiment of each comment and added this information as a new column to the DataFrame.
6. Stacking Ensemble Model: We used a Stacking Ensemble learning model that included SVM, Multinomial Naive Bayes, and a final estimator of Random Forest to train the comments data to predict sentiment. This helped us achieve a high accuracy rate of around 96.8%.
7. Emotional Analysis: We followed similar steps to sentiment analysis, but instead, we used a different pre-made emotional analyzer to classify comments into 6 different emotional categories.
8. Our project utilizes various components and techniques to extract, preprocess, analyze, and visualize data from the Reddit platform related to specific topics.

STEPS TAKEN

We first used PRAW to specify topics of subreddits related to Machine Learning, Artificial Intelligence, and Data Science, and then scraped details of those topics using the Reddit API. The data was turned into a DataFrame with columns like datetime, id, subreddit, headline, and name. We then changed the time format from UTC to year, month, day, and time.

Using the post ID, we scraped all comments using PRAW and turned the result into another DataFrame. The two DataFrames were then merged.

We did some visualizations using graphs and used a word cloud to randomly select a word - in this case, 'diffusion'. We then passed this word to a pre-made sentiment analyzer and added the resulting 'sentiment' column with values of positive, negative, and neutral to the DataFrame.

We also did some preprocessing on the comments, including removing URLs and emojis. Finally, we used a Stacking Ensemble learning model that included SVM, Multinomial Naive Bayes, and a final estimator of Random Forest to train the comments data to predict sentiment. The accuracy achieved was around 96.8%.

For emotional analysis and prediction, we followed similar steps, but instead of 3 classes of sentiment, we had 6 classes - joy, sadness, anger, fear, love, and surprise.

ALGORITHMS-USED

1. PRAW (Python Reddit API Wrapper): This is a Python library that simplifies the interaction with the Reddit API and enables us to easily extract data from subreddits.
2. TextBlob: This is a pre-trained sentiment analysis library that we used to determine the sentiment of the comments.
3. NLTK (Natural Language Toolkit): This is a library used for natural language processing tasks such as tokenization, stemming, and lemmatization. We used it to preprocess the comments data.
4. Matplotlib and Seaborn: These are visualization libraries used to create graphs and visualizations of the data.
5. WordCloud: This is a library used to create word clouds from text data.
6. SVM (Support Vector Machine): SVM is a supervised machine learning algorithm used for classification and regression analysis. It works by finding a hyperplane that separates the data into different classes. SVM is known for its ability to handle high-dimensional datasets and to work well with both linear and non-linear data.
7. Multinomial Naive Bayes: Naive Bayes is a probabilistic classifier that is based on Bayes' theorem. It works by assuming that the features are conditionally independent given the class label. Multinomial Naive Bayes is

a variant of Naive Bayes that is used for discrete data, such as word counts in text data.

8. Random Forest: Random Forest is an ensemble learning algorithm that is used for classification and regression analysis. It works by creating multiple decision trees and aggregating their predictions to make a final prediction. Random Forest is known for its ability to handle high-dimensional datasets and to work well with non-linear data.

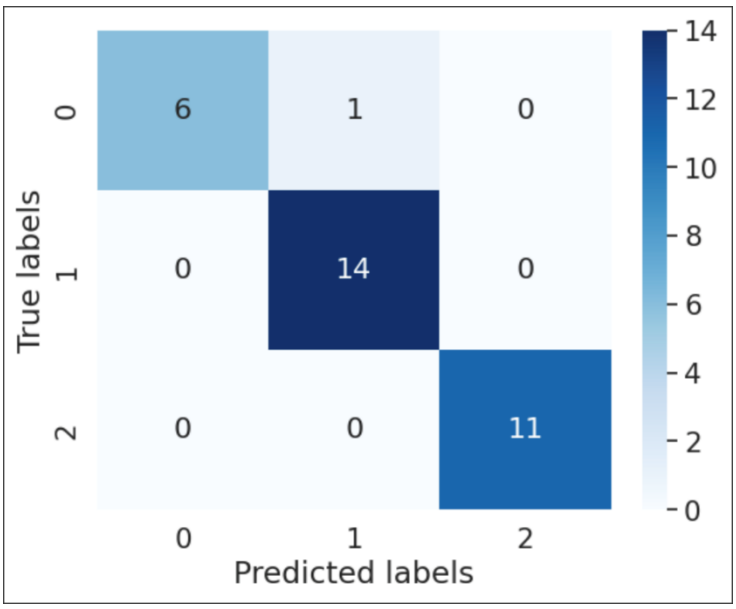
9. A Stacking Ensemble learning model combines multiple base models to improve prediction accuracy. It works by training different models on the same data, and then using a meta-model to weigh their predictions and make a final prediction. The base models can be any type of machine learning model, and the meta-model can also be any type of model. A stacking ensemble model is advantageous because it can be more robust to overfitting and can leverage the strengths of different models.

EXPERIMENT RESULTS AND DISCUSSION

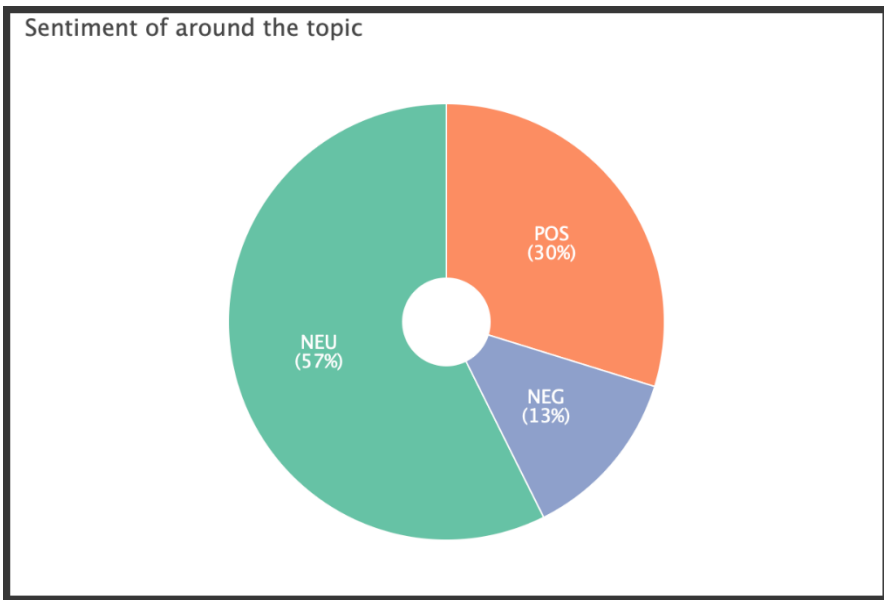
For sentiment prediction, the accuracy we got

```
0.96875
F1 score: 0.968
Precision: 0.971
Recall: 0.969
Confusion Matrix:
[[ 6  1  0]
 [ 0 14  0]
 [ 0  0 11]]
```

Confusion matrix



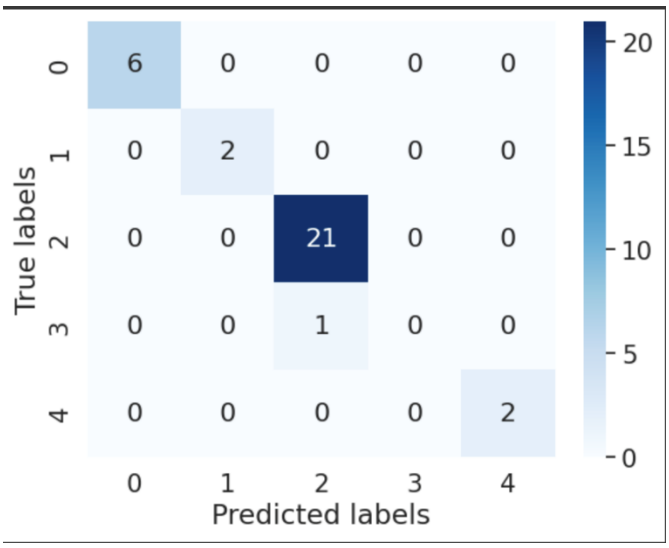
Pie chart of sentiment types



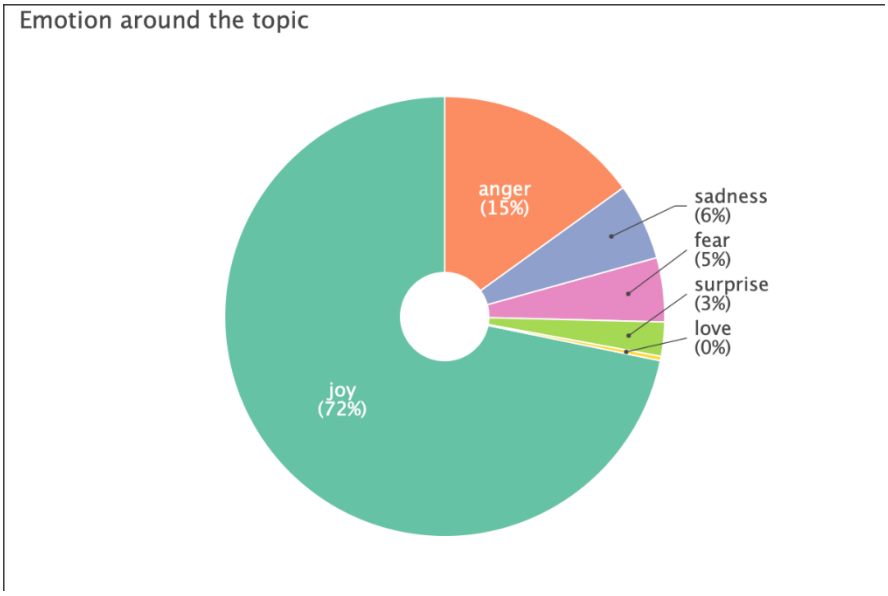
For emotion prediction, the accuracy we got

```
0.96875
F1 score: 0.953
Precision: 0.939
Recall: 0.969
Confusion Matrix:
[[ 6  0  0  0  0]
 [ 0  2  0  0  0]
 [ 0  0 21  0  0]
 [ 0  0  1  0  0]
 [ 0  0  0  0  2]]
```

Confusion matrix



Pie chart of emotions type



CONCLUSION AND FUTURE WORK

In conclusion, our project successfully implemented sentiment and emotional analysis on live Reddit data using machine learning techniques. We were able to accurately predict the sentiment of comments with an accuracy of 96.8% using a stacking ensemble learning model that included SVM, Multinomial Naive Bayes, and Random Forest classifiers. We were also able to predict six different emotions with an accuracy of 96.8%.

Future work for this project could include improving the accuracy of emotional analysis by incorporating more advanced natural language processing techniques, such as neural networks or deep learning models. We could also explore ways to incorporate user and subreddit information to improve the accuracy of our predictions. Additionally, our project could be extended to other social media platforms to analyze sentiment and emotions on a larger scale. This could involve developing a more generalizable model that can work across different social media platforms, as each platform has its own unique characteristics and patterns of user behavior. Another area of future work could be to integrate the sentiment and emotional analysis into a real-time monitoring system for businesses or organizations to track their online reputation and customer satisfaction.

REFERENCES

1. R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. "Recursive deep models for semantic compositionality over a sentiment treebank." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013.
2. Y. Kim. "Convolutional neural networks for sentence classification." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.
3. J. Pennington, R. Socher, and C. Manning. "GloVe: Global vectors for word representation." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.
4. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need." In Advances in Neural Information Processing Systems (NIPS), 2017.
5. S. Hochreiter and J. Schmidhuber. "Long short-term memory." Neural computation, 1997.
6. T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient estimation of word representations in vector space." In Proceedings of the International Conference on Learning Representations (ICLR), 2013.
7. J. Howard and S. Ruder. "Universal language model fine-tuning for text classification." In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.
8. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.
9. Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In Mining Text Data (pp. 163-222). Springer.