IT - 490 : Thesis

Comparative Sentiment Analysis of Public Perception: Monkeypox vs. COVID-19 Behavioral Insights

Supervisor:
Barsha Roy
Lecturer,
Department of IT, UITS

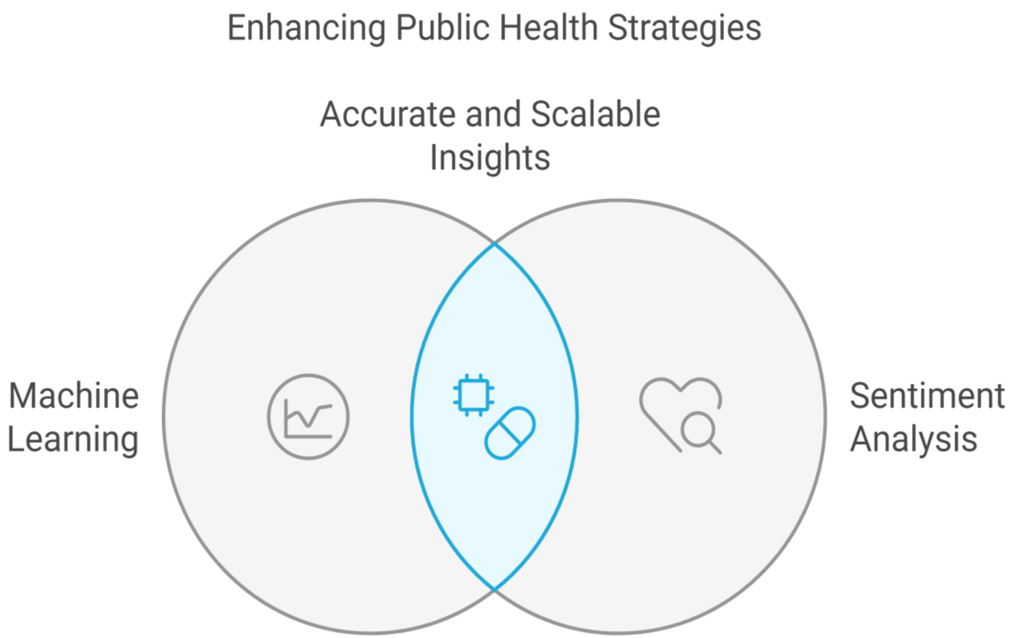
Team Members:
1. Mostafa Mohaimen Akand Faisal (2014755010)
2. Rabeya Amin Jhuma (2014755028)

Outline

- Introduction
- Motivation
- Problem-Statement
- Objectives
- Related-Works
- Dataset
- Methodology
- Applied-Algorithms
- Result-Analysis
- Comparison
- Future-Directives
- Conclusion
- Reference

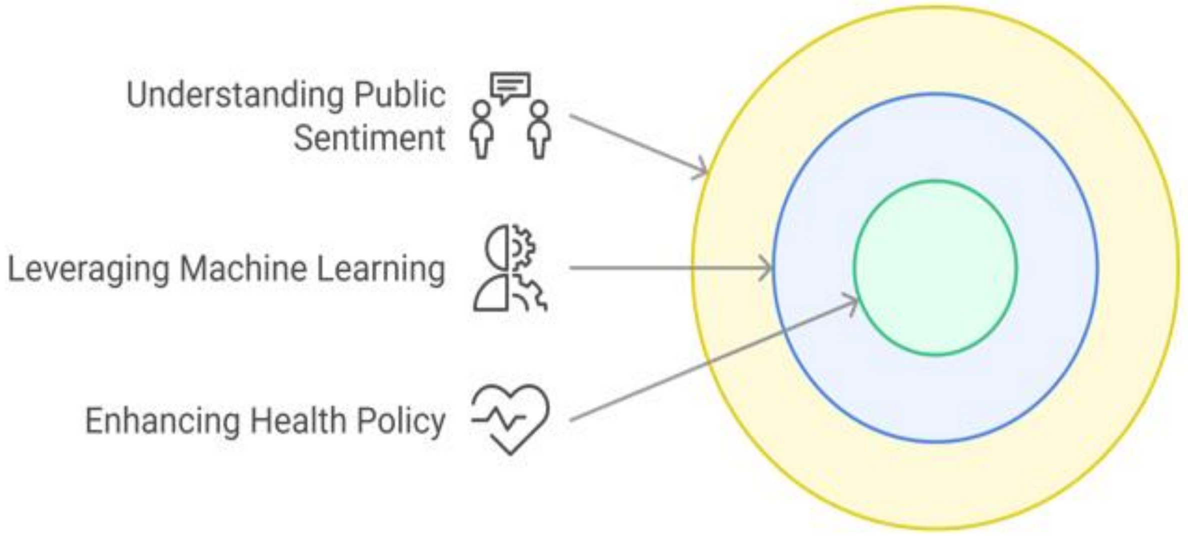
Introduction

- **Emergence of Infectious Diseases**: Monkeypox and COVID-19 pose major challenges to public health management.
- **Role of Machine Learning**: ML algorithms enable accurate, scalable sentiment analysis, offering actionable insights.
- **Impact**: Supports better-informed health strategies and effective communication.



Motivation

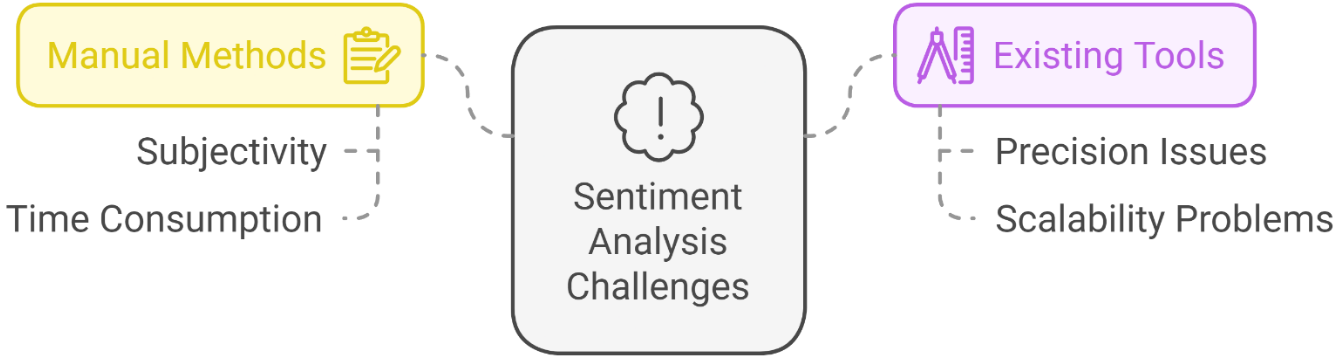
- **Understanding Public Sentiment**: Identify concerns and misconceptions to guide communication.
- **Leveraging Machine Learning**: Enable accurate and scalable sentiment analysis.
- **Enhancing Health Policy**: Provide actionable insights for targeted interventions.



Problem Statement

Current Sentiment Analysis Challenges:

- Manual sentiment analysis methods are often subjective and time-consuming.
- Existing tools lack the precision and scalability needed to compare public perceptions of diseases like Monkeypox and COVID-19.



Objectives

- Develop scalable sentiment analysis tools.
- Analyze public sentiment for Monkeypox and COVID-19.
- Guide health strategies through insights.
- Enhance sentiment classification accuracy.
- Facilitate comparative behavioral analysis.

Related Works

Papers	Summary	Models Used	Best Performanc	Dataset Used
1. Fine-tuned Sentiment Analysis of COVID-19 Vaccine–Related Social Media Data: Comparative Study	The study uses two major social media platforms: Twitter and Reddit, used a fine-tuned DistilRoBERTa model aiming to understand how sentiments changed over time	Fine-tuned **Distil Roberta** (a robust variant of Roberta optimized for efficiency)	Accuracy: **95.92%,** F1-Score: **95.92%,** Precision: **95.61%**	9.5 million tweets from January 1, 2020, to March 1, 2022, filtered for "COVID Vaccine."
2. A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets	The research analyzes public sentiments using VADER and TextBlob. It develops machine learning models, with the SVM model achieving the highest accuracy. The findings aim to assist health authorities in understanding public perceptions and shaping effective health policies	K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, Logistic Regression, Multilayer Perceptron (MLP), Naïve BayesXGBoost	Accuracy: **93.48%** (using TextBlob annotation + Lemmatization + CountVectorizer + SVM)	107,000 tweets collected between July 2022 and September 2022

Related Works

Papers	Summary	Models Used	Best Performanc	Dataset Used
3. Sentiment Analysis and Text Analysis of the Public Discourse on Twitter about COVID-19 and MPox	This paper first study to compare discourse on both COVID-19 and MPox but lacks filtering of bot-generated tweets, impacting sentiment accuracy.	Random Forest, LR, Multilayer Perceptron, SVM, Naïve Bayes, KNN, XGB	RF = 90.05 LR = 89.51 Multilayer Perceptron = 91.86 SVM = 92.95	The study analyzed a total of 61,862 Tweets collected over a specific time frame, from May 7, 2022, to March 3, 2023.
4. Improving Public Health Policy by Comparing the Public Response during the Start of COVID-19 and Monkeypox on Twitter in Germany: A Mixed Methods Study	Highlights platform-specific sentiment trends but limited by user demographics and lack of real-time analysis.	SVM, LR, Naïve Bayes, LDA	SVM = 88.3 LR = 85.6 Naïve Bayes = 81.2	For COVID-19, tweets were collected between January 1, 2020, and March 31, 2020. A total of 8,532 tweets. For monkeypox, tweets were collected from May 1, 2022, to July 31, 2022. Total 7,404 tweets,

Dataset Details

Name: Covid-19 Twitter Dataset, Monkeypox Tweets, Tweets on monkeypox, Monkeypox tweets data

COVID-19 Data: 147,475

Monkeypox Data: 106,638

Dataset Attributes: ID, date, tweet sources ,hashtags, mentions, liked tweets, sentiment, tagging.

Methodology

Experimental framework

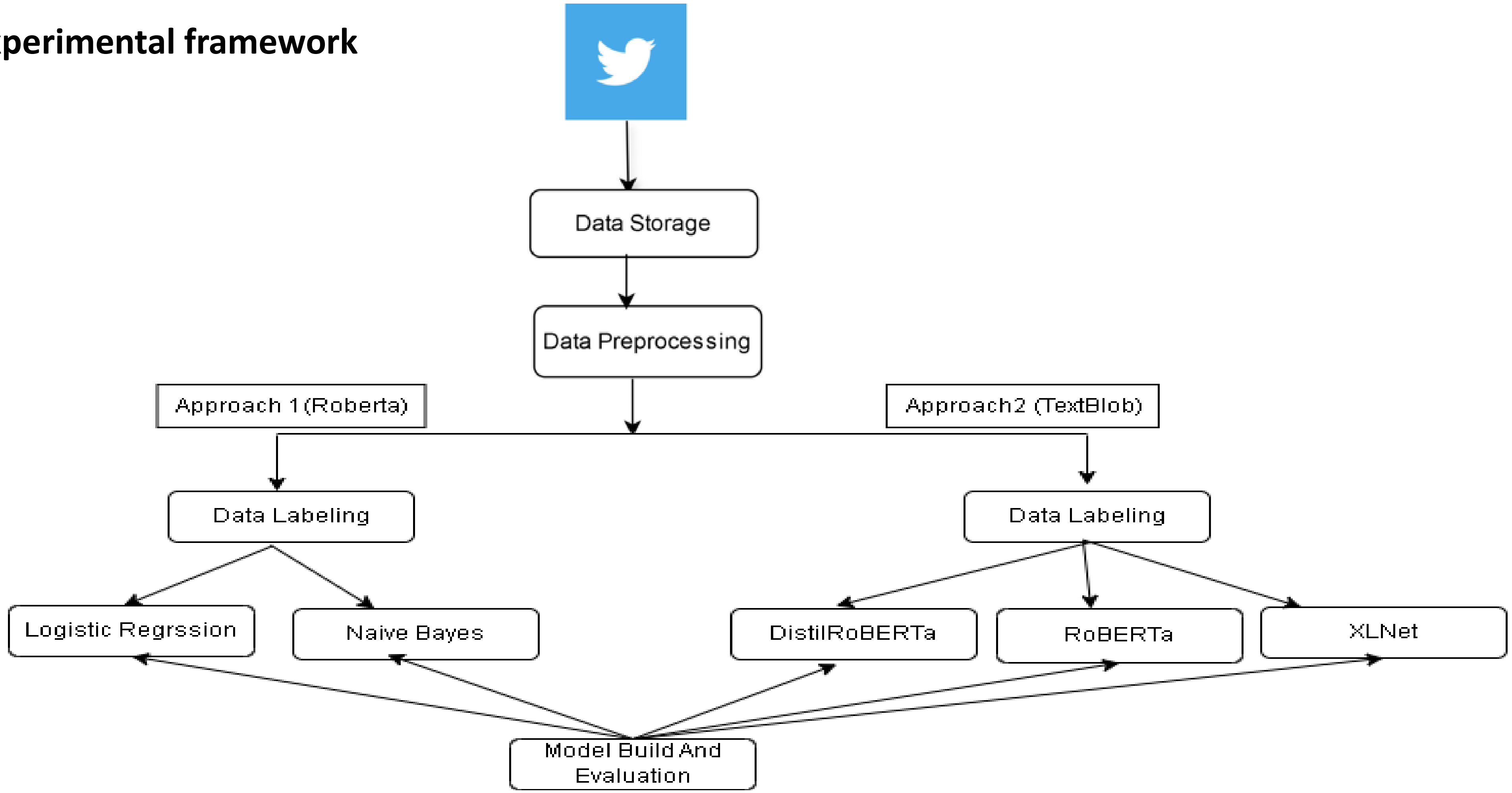


Fig 1: Flow of Work

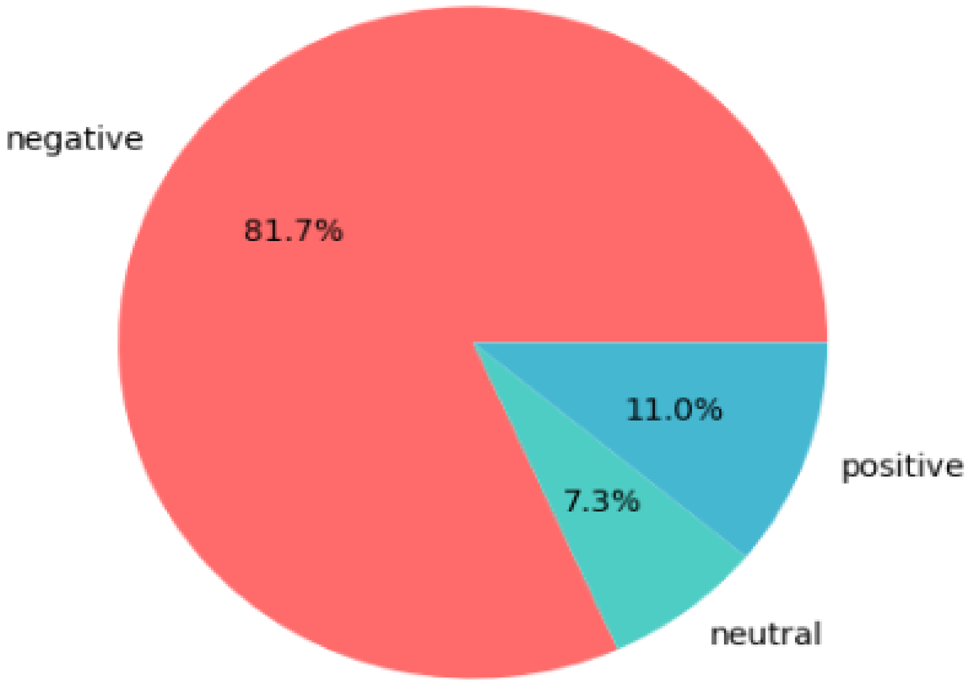
Methodology Cont’d

Data Preprocessing

- **Text Cleaning:** Removing URLs, special characters, and irrelevant words.
- **Stopword Removal:** Custom stopwords for COVID-19 and Monkeypox keywords.
- **Tokenization:** Breaking down tweets into words.
- **TF-IDF Vectorizer:** Captures the importance of words in the text data.
- **RoBERTa:** Enables capturing rich contextual representations of text.
- **TextBlob :** Makes text data analysis straightforward.
- **Sentiment Assignment:** Categorizing tweets into positive, negative, and neutral based on polarity.

Data Visualization(approach 1)

COVID-19



MONKEYPOX

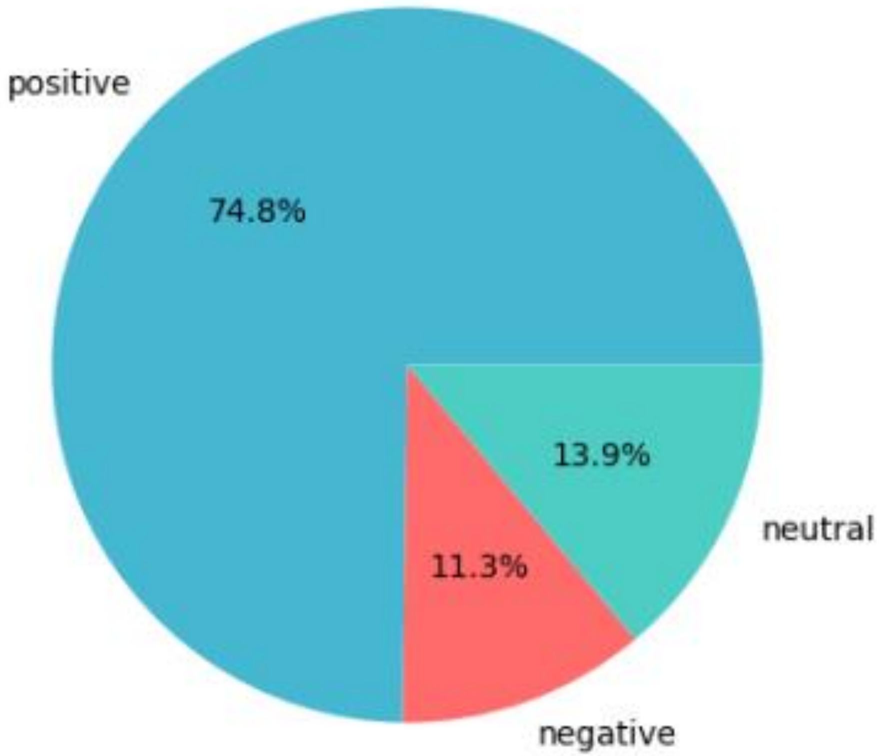
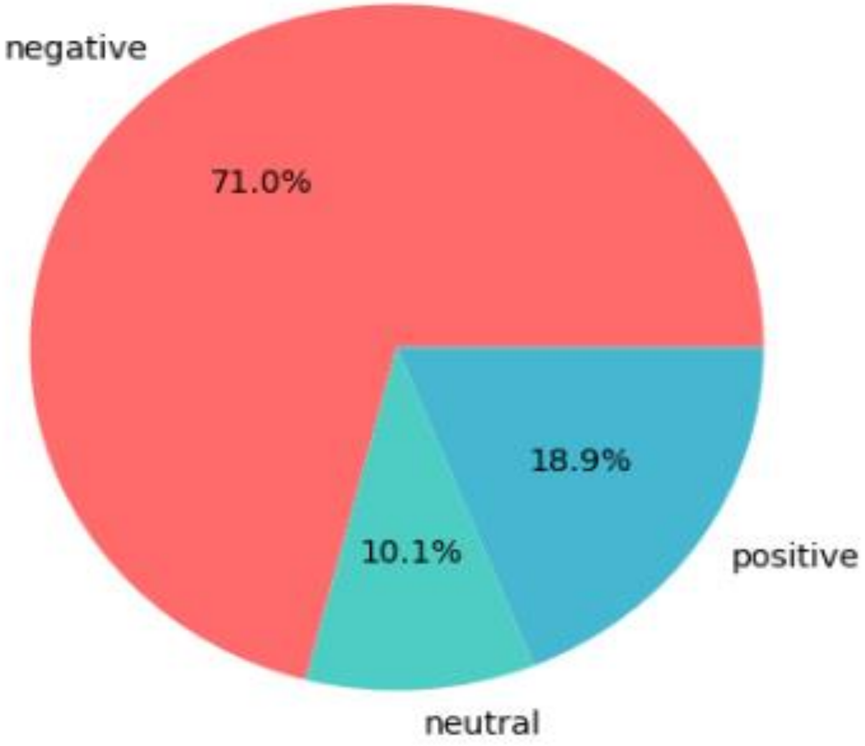


Fig 2: Sentiment Distribution (RoBERTa Sentiment Analysis)

Methodology Cont'd

Data Visualization(approach 2)

COVID-19



MONKEYPOX

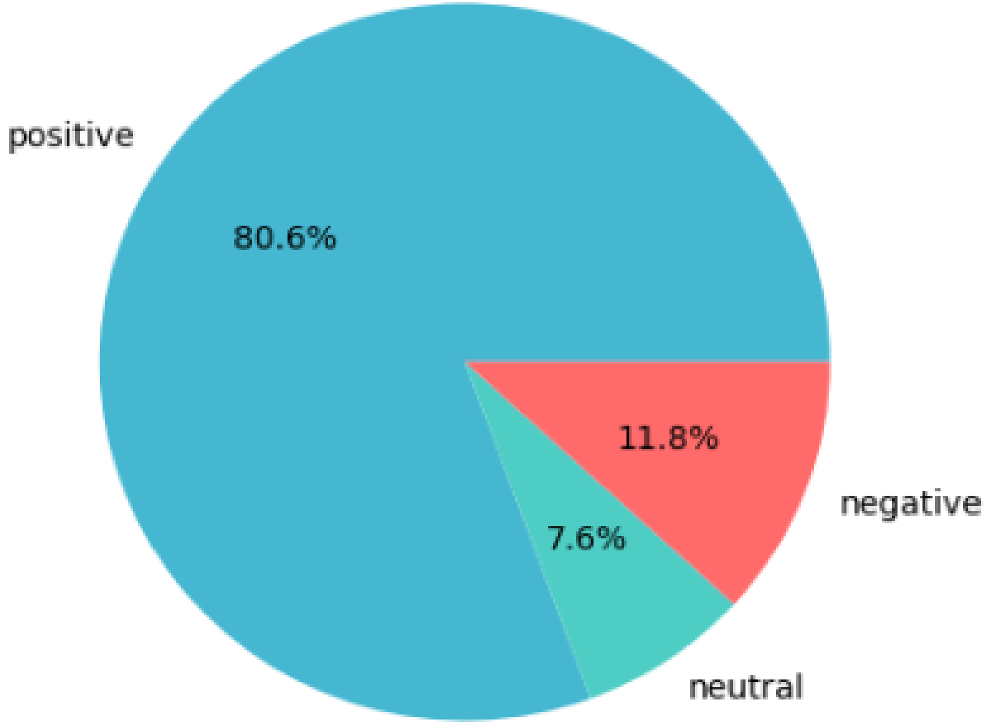


Fig 3: Sentiment Distribution (TextBlob Sentiment Analysis)

Methodology Cont’d

Applied Methods

- **Traditional Machine Learning Models**
  - **Logistic Regression**:
    Trained a multinomial Logistic Regression model as a baseline for sentiment classification.
  - **Naive Bayes**:
    Trained a Naive Bayes classifier (MultinomialNB) for sentiment analysis, utilizing a TF-IDF vectorizer for feature extraction.
- **Transformer-based Models**
  - **ROBERTA:**
    Fine-tuned (Pre-trained Transformer) the Roberta-base model for sentiment classification.
  - **XLNet:**
    fine-tuned (Pre-trained Transformer) architecture capable of leveraging bidirectional context for better understanding of input text.
  - **DistilRoBERTa:**
    Trained a transformer-based DistilRoBERTa model for sentiment classification, being a distilled version of RoBERTa

Environment Setup

- **Platform**: Google Colab
- **Hardware Configuration:**
  - **GPU Enabled:** Configured through Runtime > Change runtime type > GPU.
  - **RAM:** High-RAM mode (12GB) was utilized via Colab to support the computational demands of the experiment.

Model Comparison

Parameter/Attribute	Logistic Regression	Naive Bayes
Loss Function	Not applicable (Logistic regression optimizes log-loss internally)	Multinomial log-likelihood loss
Optimizer	lbfgs (default for Scikit-learn multinomial Logistic Regression)	Uses Maximum Likelihood Estimation (MLE) for fitting the model
Activation Function	Softmax	No activation function (direct calculation based on probabilities)
Learning Rate	Not directly adjustable	Not applicable (model parameters are directly computed)
Vectorization/Embedding	TF-IDF (max_features = 5000)	TF-IDF vectorization (max_features=5000)
Input Text Preprocessing	Lowercasing, tokenization, and TF-IDF	Lowercasing, tokenization, and TF-IDF
Evaluation Metric	COVID(87.69%), Monkeypox(82.96%)	COVID(70.00%), Monkeypox(94.00%)
Dataset Split	80% Training, 20% Testing	80% Training, 20% Testing

Model Comparison Cont'd

Parameter/Attribute	ROBERTA	XLNet	DistilRoberta
Loss Function	Sparse Categorical Crossentropy	SparseCategorical Crossentropy	SparseCategorical Crossentropy
Optimizer	AdamW	Adam	Adam
Activation Function	GeLU	Linear (logits for multi-class classification)	Linear (logits for multi-class classification)
Learning Rate	2e-5	2e-5	3e-5
Vectorization/Embedding	ROBERTA Tokenizer	XLNet tokenizer	DistilRoBERTa tokenizer
Input Text Preprocessing	Tokenization, padding, truncation (max length = 50)	Tokenization (max_length=64)	Tokenization (max_length=128)
Evaluation Metric	COVID(97.99%), Monkeypox (95.36%)	COVID(97.05%), Monkeypox (96.82%)	COVID(97.51%), Monkeypox (95.19%)
Dataset Split	80% Training, 20% Testing	80% Training, 20% Testing	80% Training, 20% Testing

Correlation

CFR (Extremely Case Fatality Rate Comparison)

COVID-19 CFR : 77.08%

Monkey pox CFR42: 42.38%

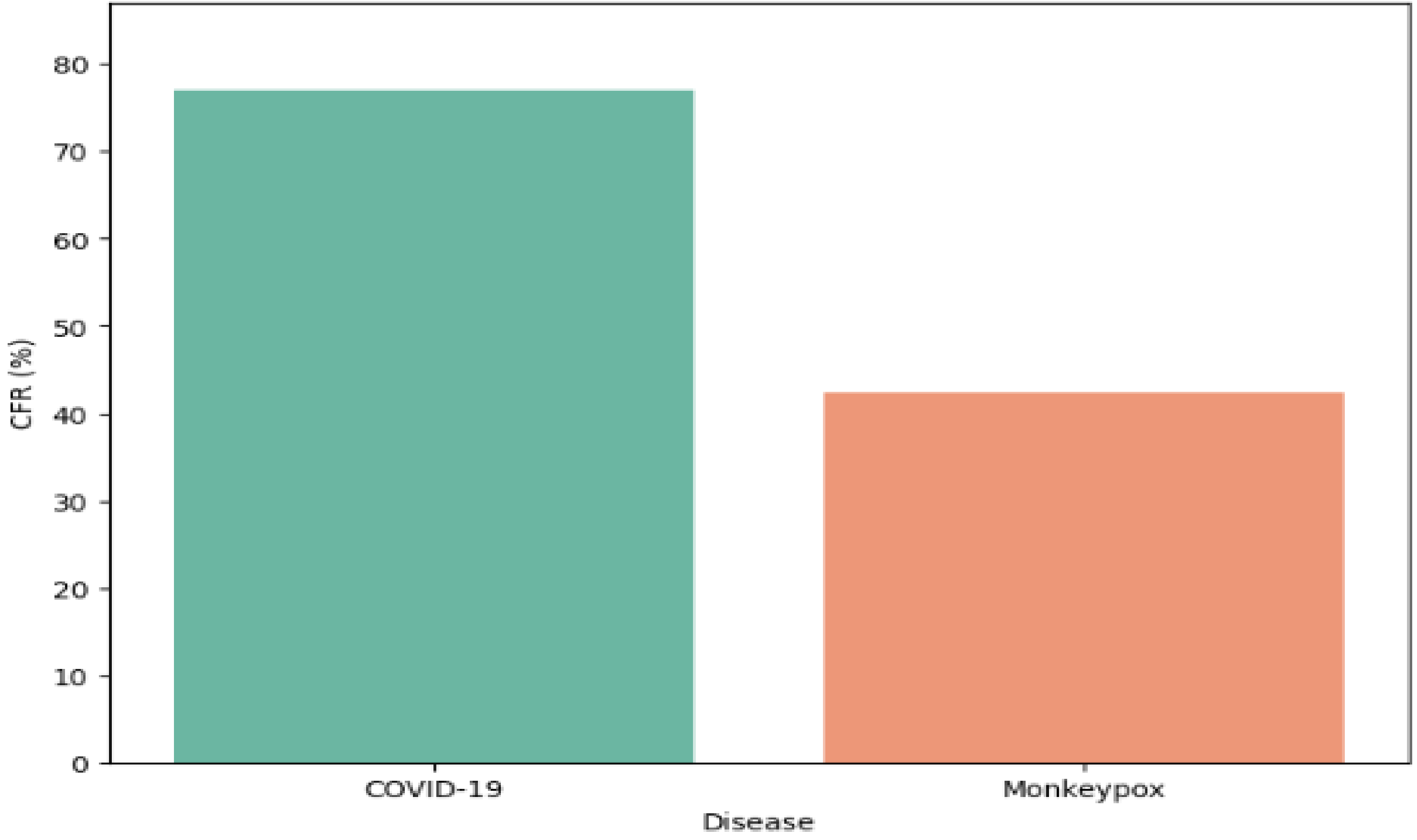


Fig 2: CFR Comparison

CFR Calculation:

CFR (%) = (Deaths/Cases)×100

Deaths: Refers to the total number of deaths

Cases: Refers to the total number of confirmed cases

Percentage Conversion: The result of the division is multiplied by 100

Correlation Cont'd

Age Distribution of Cases (Based on Death Rate)

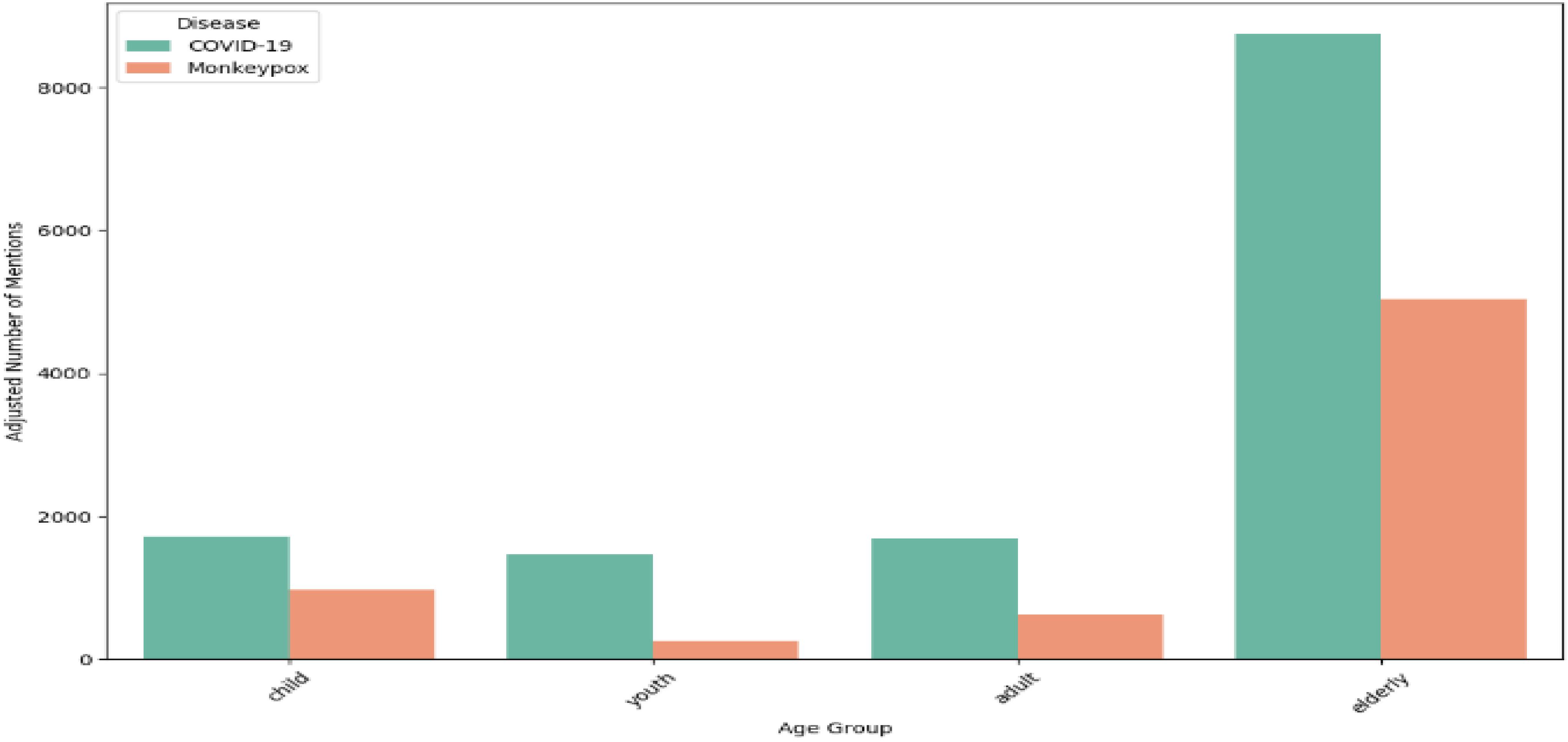


Fig 2: Age Distribution of Cases COVID vs Monkey pox

Correlation Cont'd

Gender Distribution of Cases

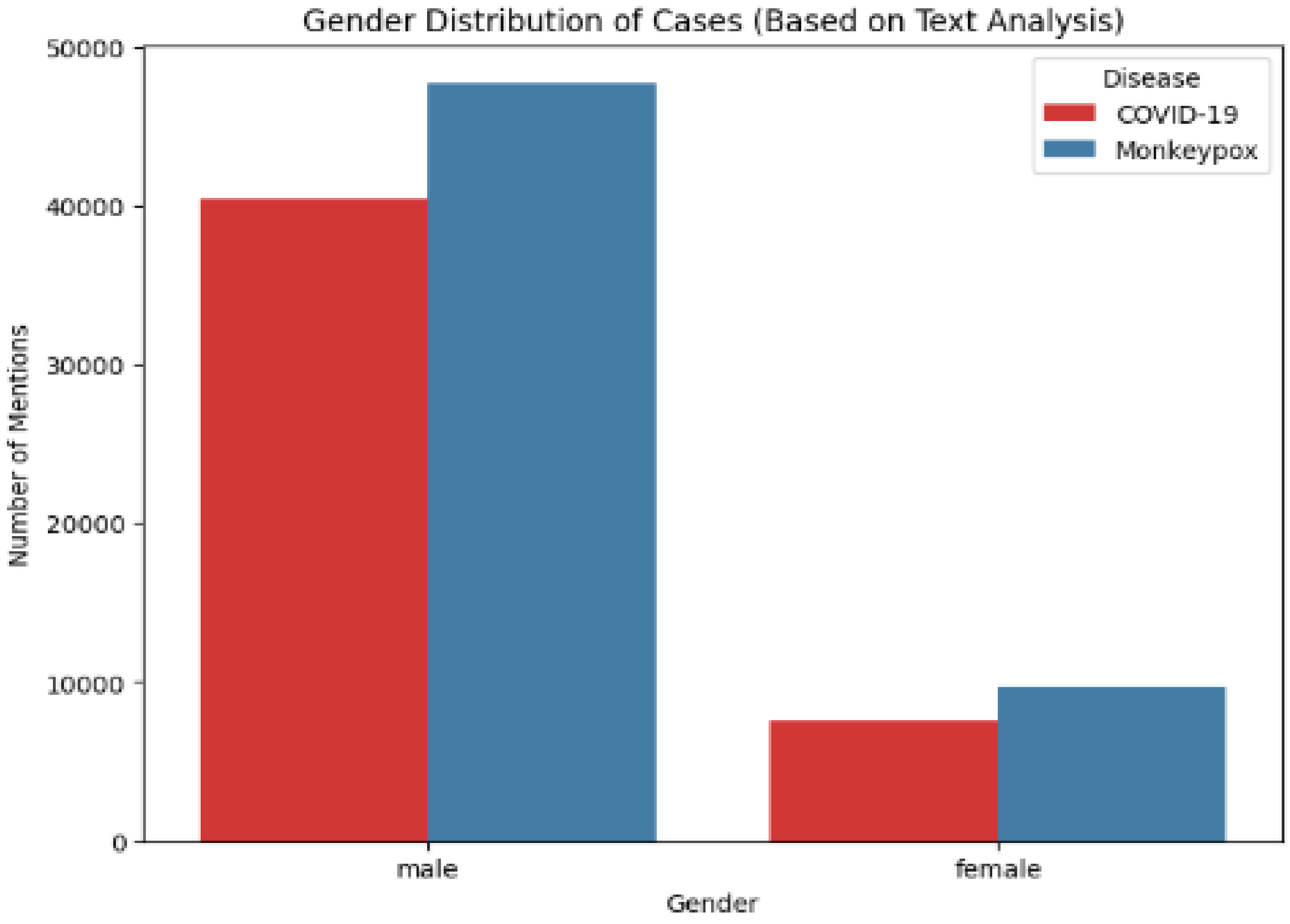
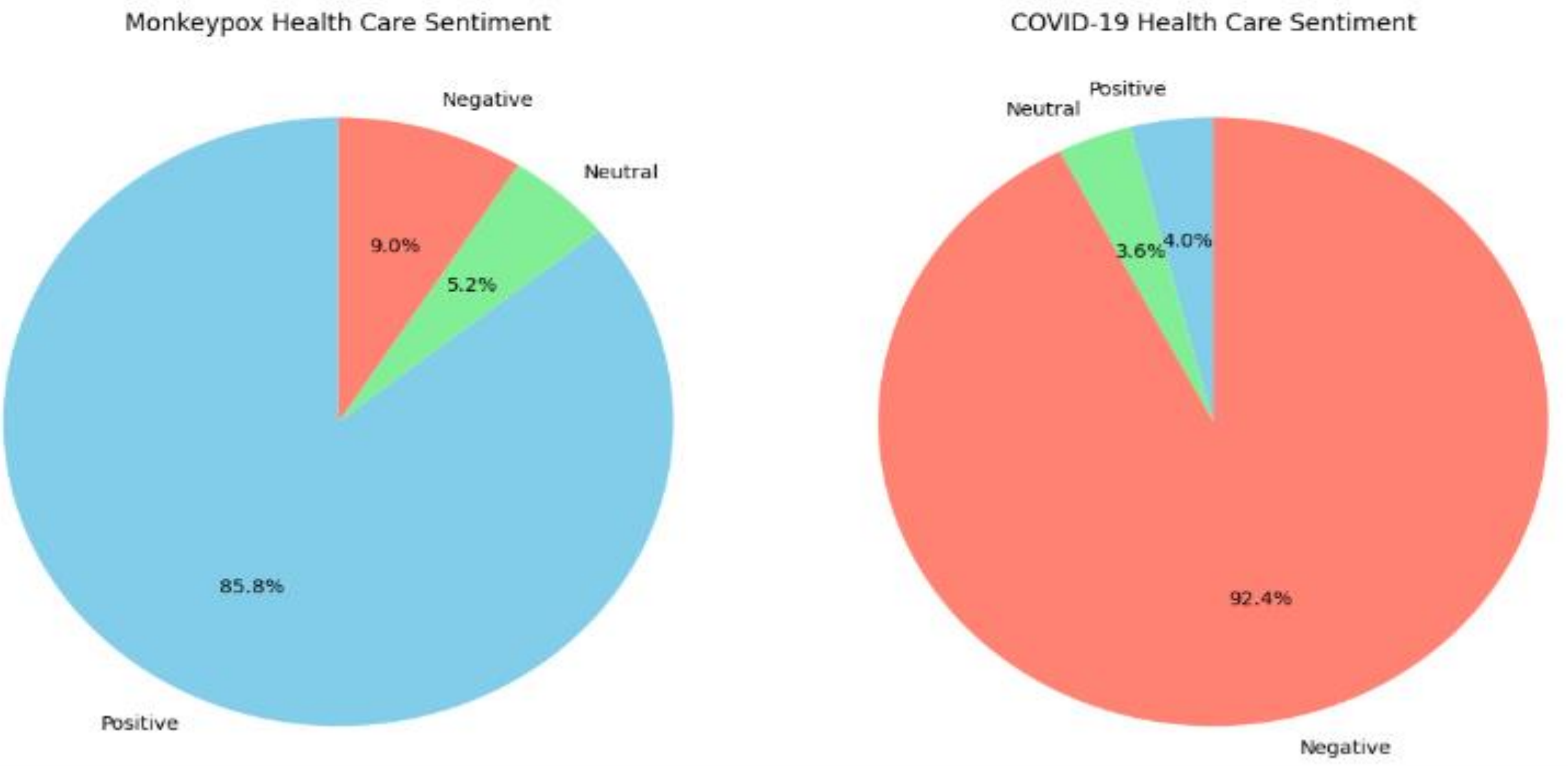


Fig 2: Gender Distribution of Cases COVID vs Monkey pox

Correlation Cont'd

Healthcare System Trust Analysis



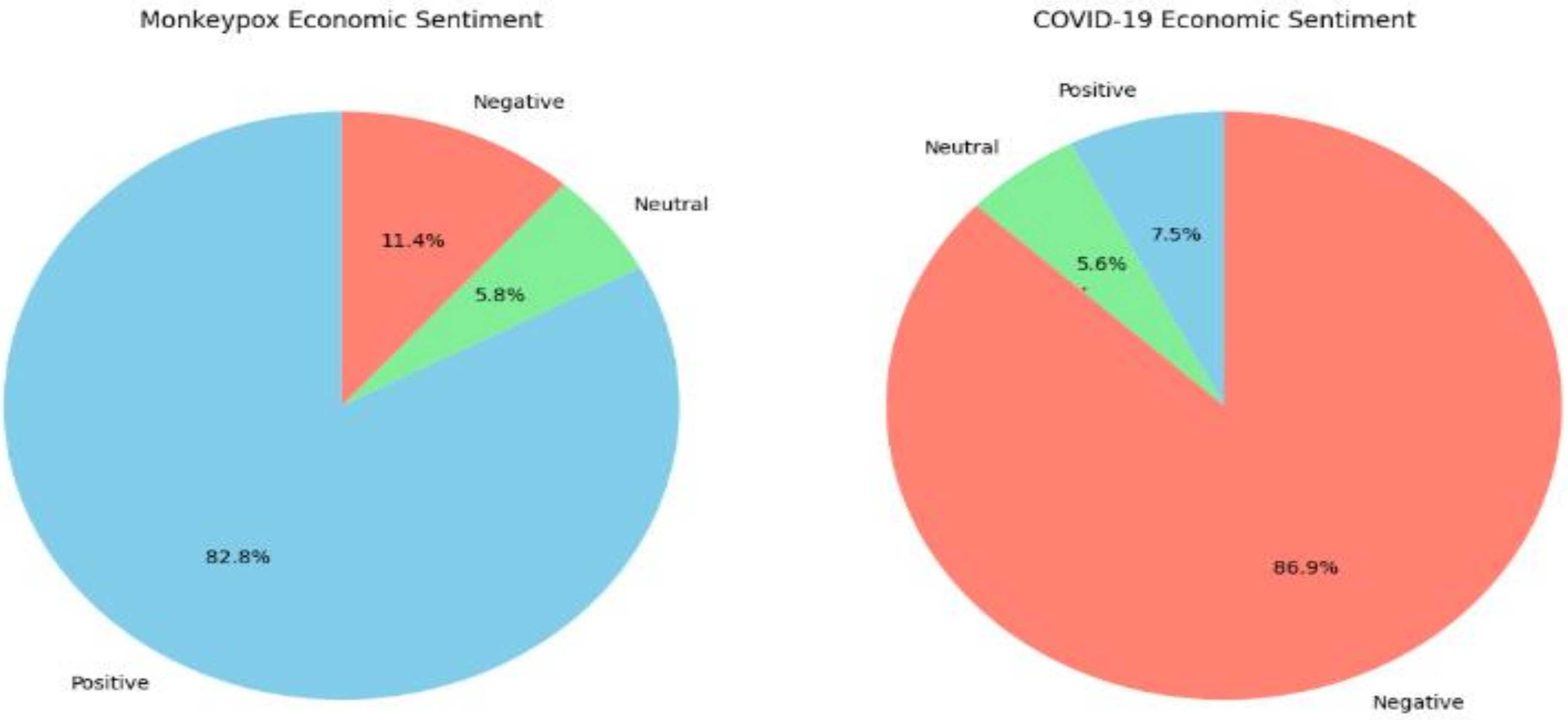

```
covid_sentiment
negative    0.924138
positive    0.039882
neutral     0.035980

mpox_sentiment
positive    0.858055
negative    0.090210
neutral     0.051735
```

Fig 2: Healthcare System Trust Analysis COVID vs Monkey pox

Correlation Cont’d

Economic Impact Perception



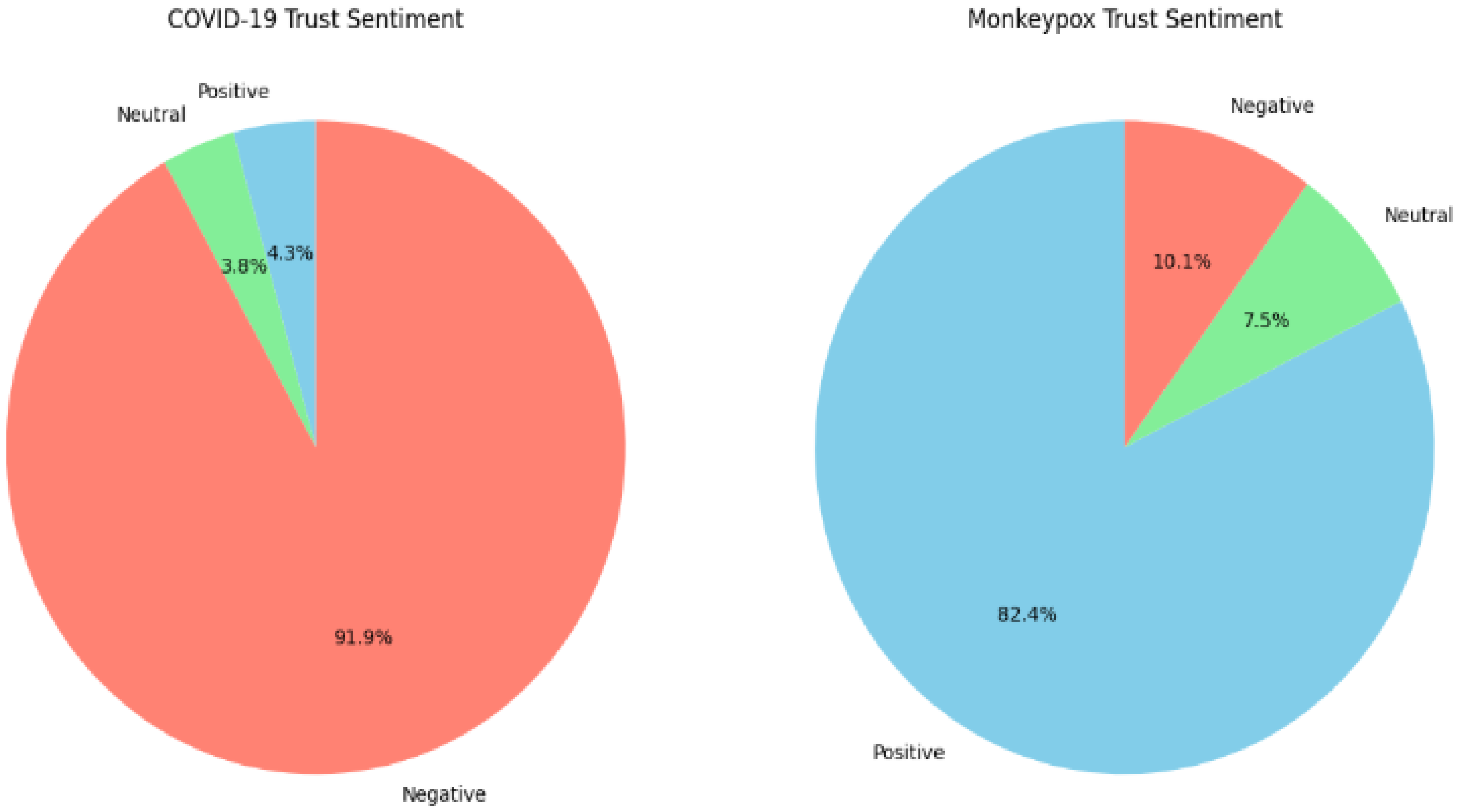
covid_sentiment	
negative	0.869037
positive	0.075158
neutral	0.055806

mpox_sentiment	
positive	0.827945
negative	0.114319
neutral	0.057737

Fig 2: Economic Impact Perception COVID vs Monkey pox

Correlation Cont'd

Information Source Trust Comparison



covid_sentiment	
negative	0.918694
positive	0.043027
neutral	0.038279

mpox_sentiment	
positive	0.824002
negative	0.100723
neutral	0.075275

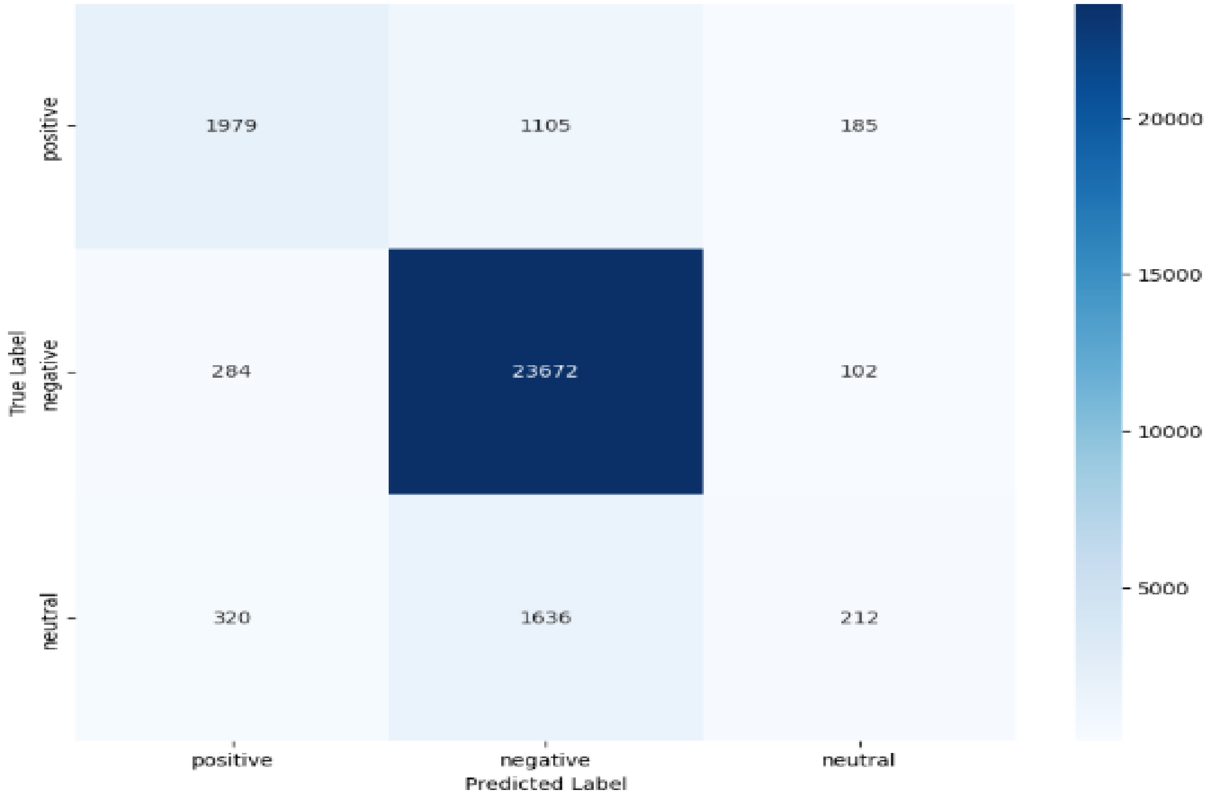
Fig 2: Information Source Trust Comparison COVID vs Monkey pox

Result and Performance Analysis Cont’d

COVID-19

```
Accuracy Score:
87.69%

Overall Model Performance:
Precision: 0.85
Recall: 0.88
F1 Score: 0.85
```



MONKEYPOX

```
Accuracy Score:
82.96%

Overall Model Performance:
Precision: 0.81
Recall: 0.83
F1 Score: 0.81
```

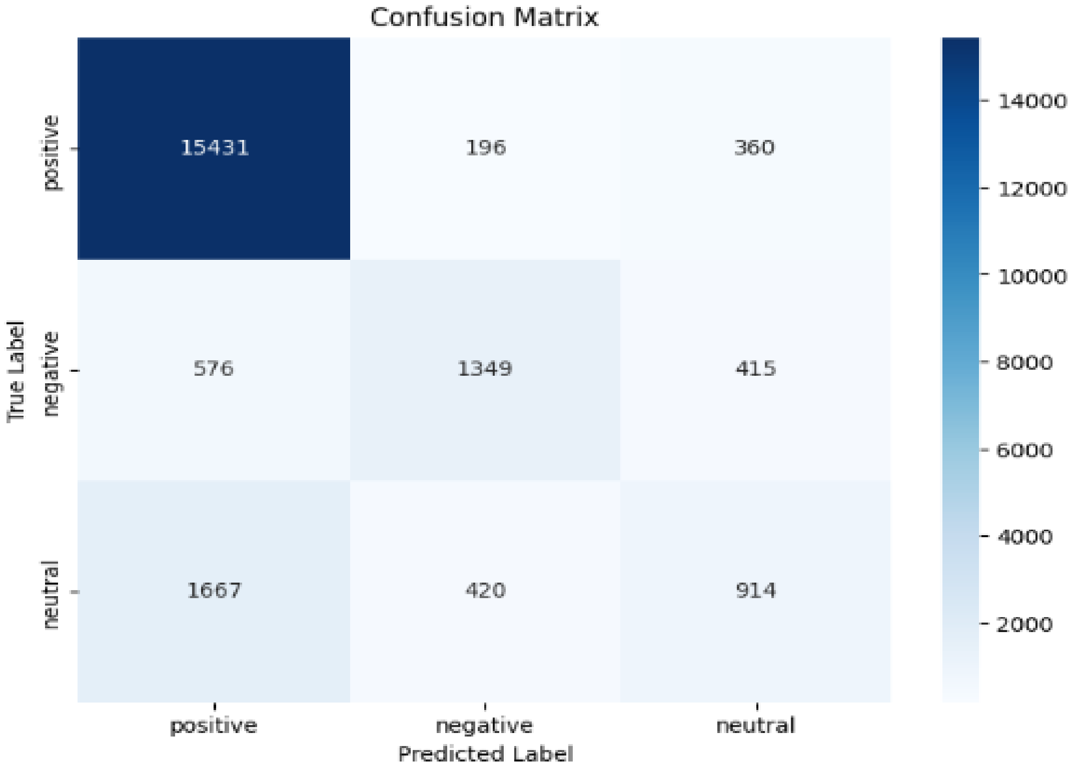
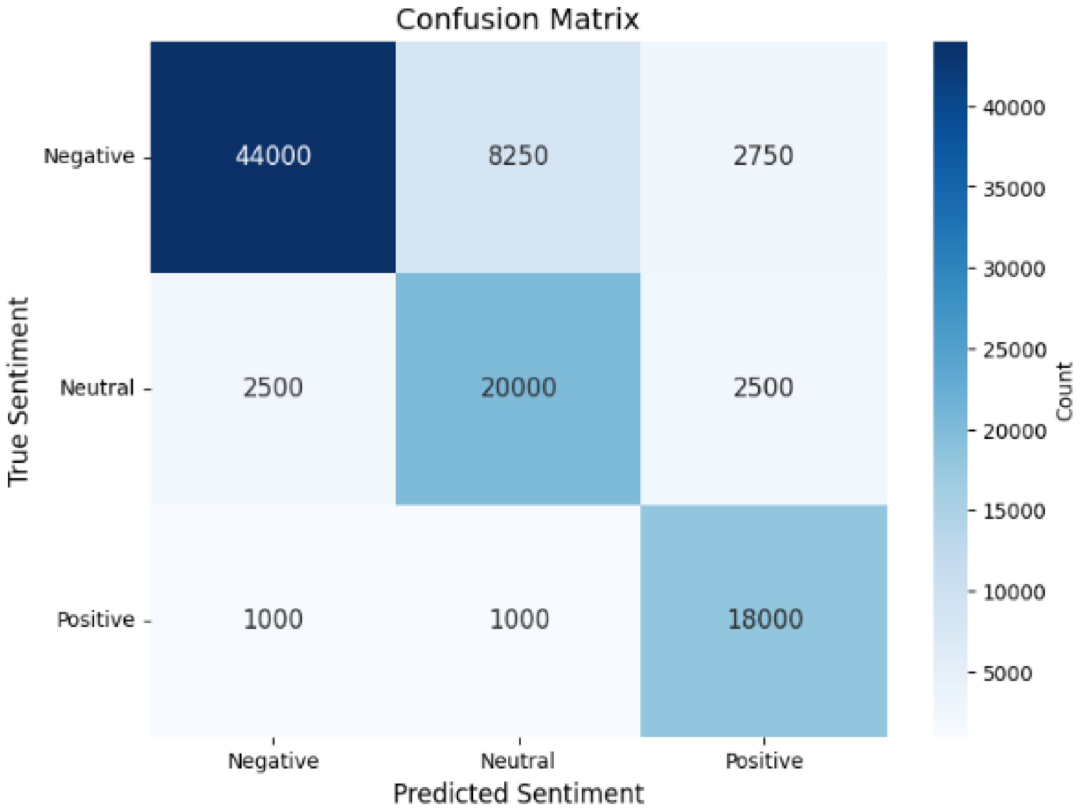


Fig 4: Classification Report of Logistic Regression

Result and Performance Analysis Cont'd

COVID-19

Accuracy: 0.70
Precision: 0.66
Recall: 0.66
F1 Score: 0.66



MONKEYPOX

Accuracy: 0.94
Precision: 0.63
Recall: 0.67
F1 Score: 0.64

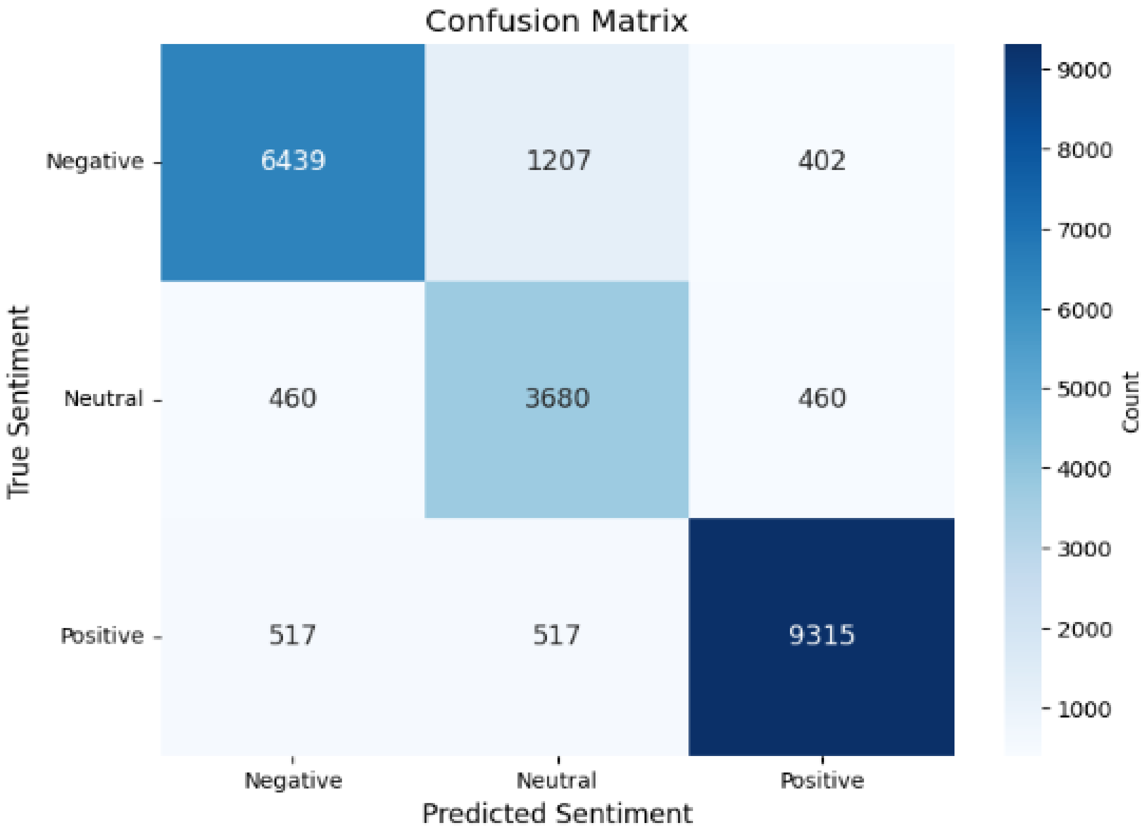


Fig 4: Classification Report of Naive Bayes

Result and Performance Analysis Cont'd

COVID-19

Accuracy: 97.99%
Precision: 0.98
Recall: 0.98
F1 Score: 0.98

MONKEYPOX

Accuracy: 95.36%
Precision: 0.95
Recall: 0.95
F1 Score: 0.95

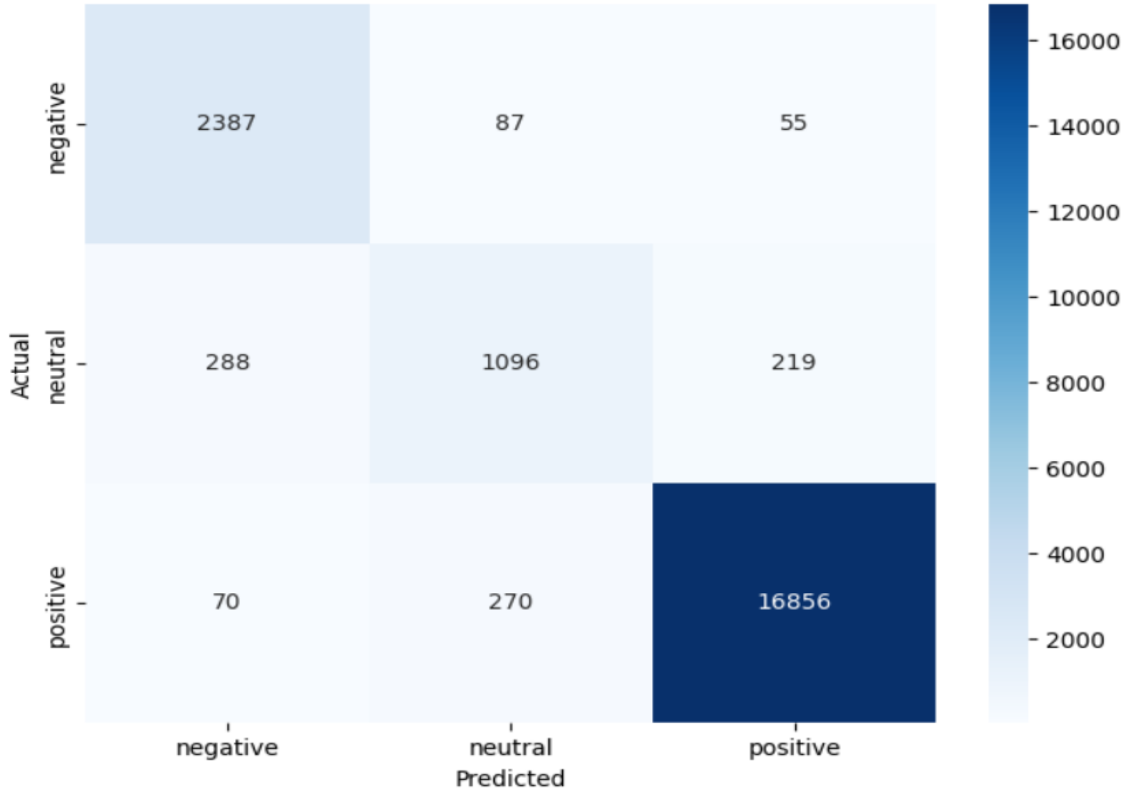


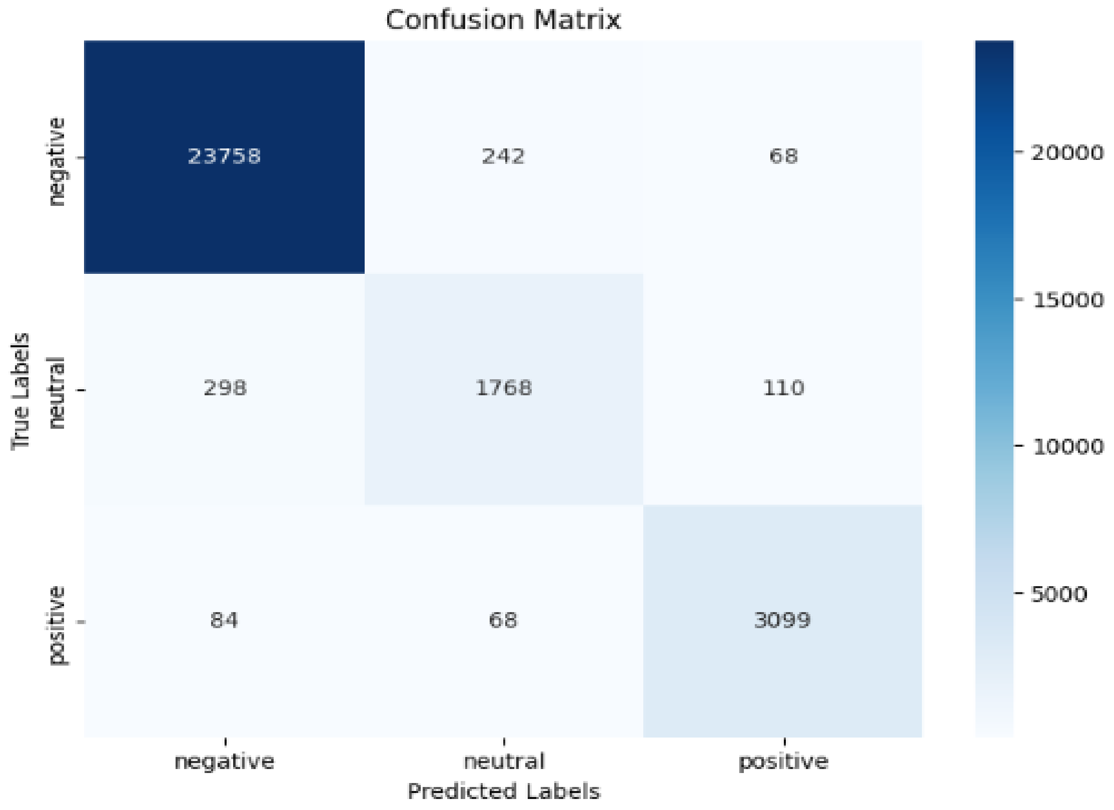
Fig 5: Classification Report of ROBERTA

Result and Performance Analysis Cont’d

COVID-19

Acuracy: 97.05%

Precision: 0.97%
Recall: 0.97%
F1 Score: 0.97%



MONKEYPOX

Acuracy: 96.82%

Precision: 0.97%
Recall: 0.97%
F1 Score: 0.97%

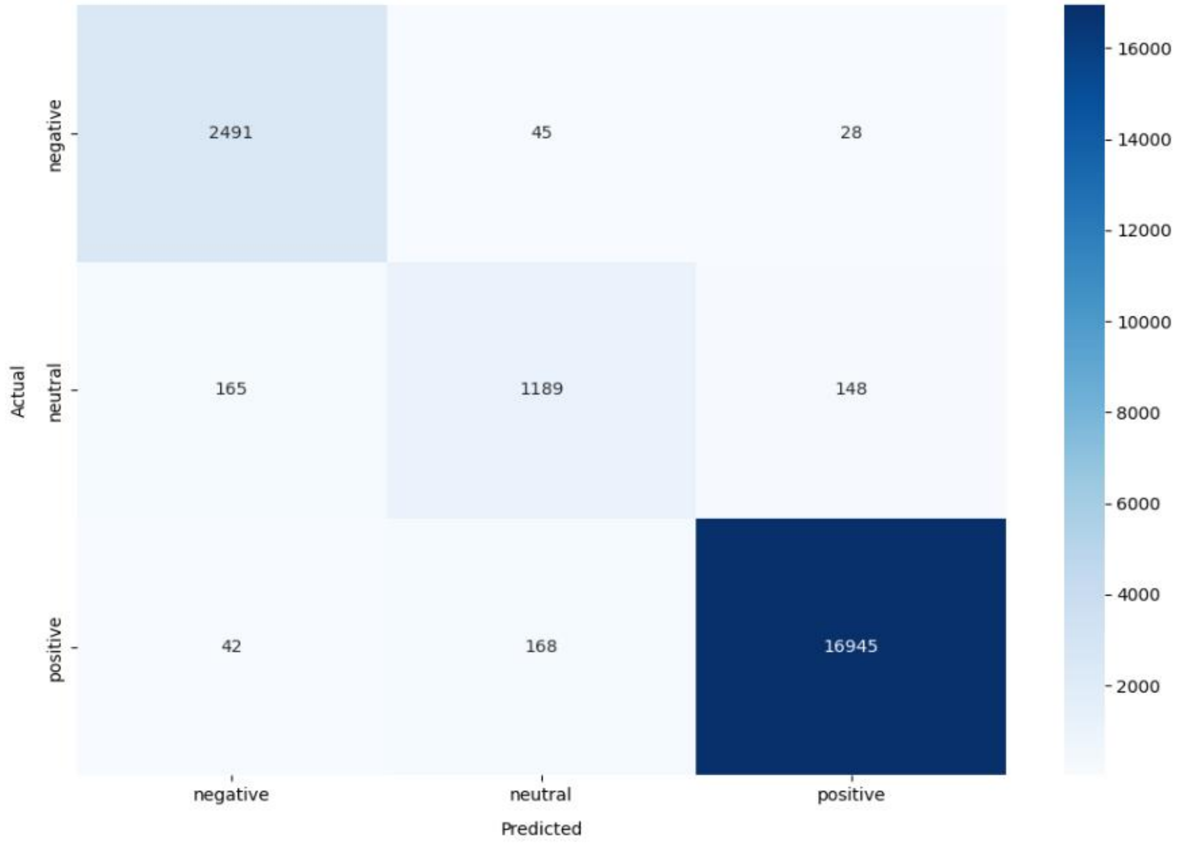


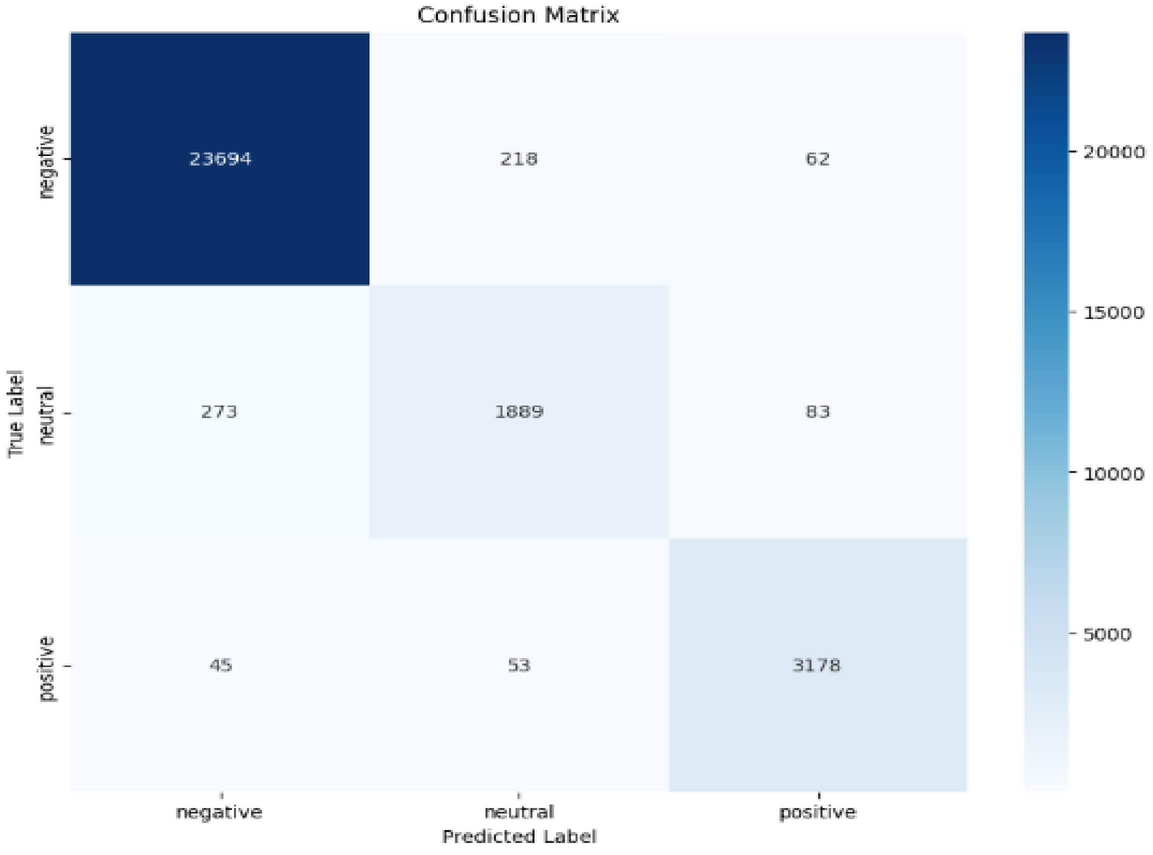
Fig 5: Classification Report of XLNet

Result and Performance Analysis Cont’d

COVID-19

Acuracy: 97.51%

Precision: 0.97%
Recall: 0.97%
F1 Score: 0.98%



MONKEYPOX

Acuracy: 95.19%

Precision: 0.95%
Recall: 0.91%
F1 Score: 0.94%

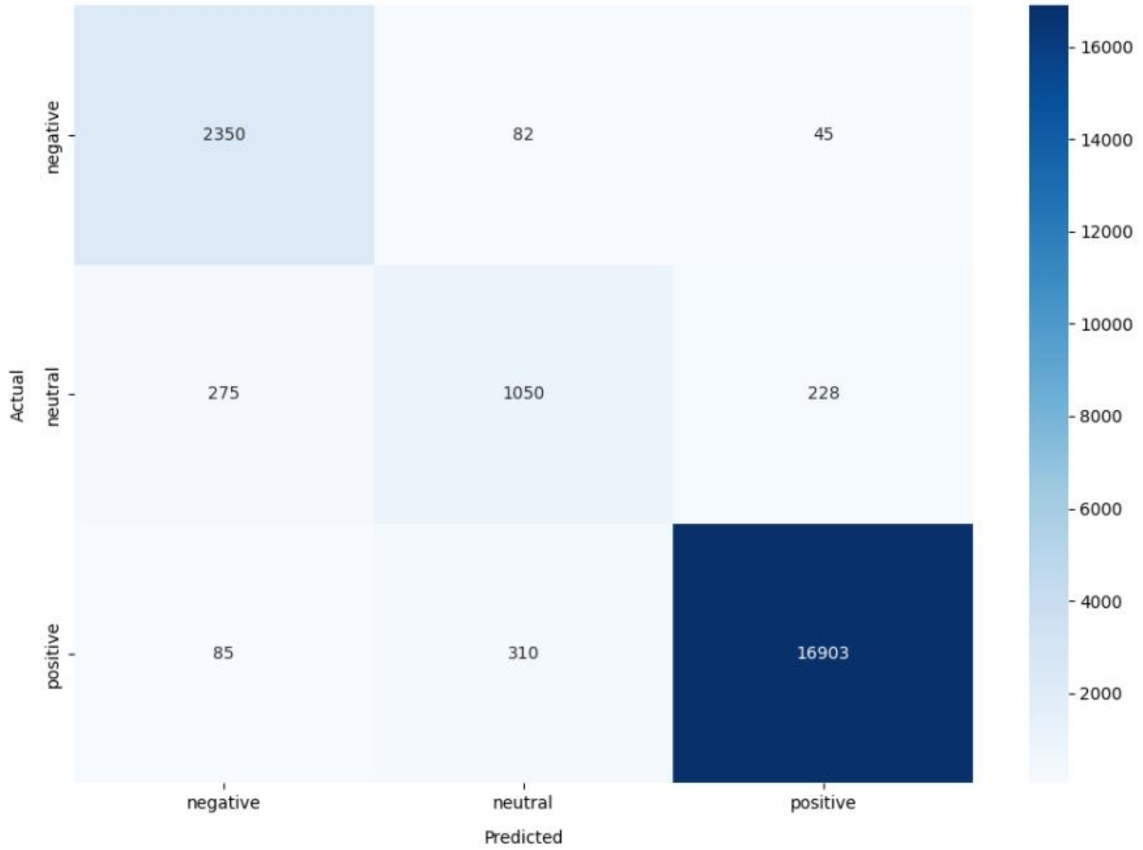


Fig 5: Classification Report of DistilRoBERTa

Future Work

- Improving Models
- API
- Correlate

Conclusion

- Analyzing Public Perception
- Sentiment and Behavior Differences
- Model Comparison
- Importance of Model Choice

Referances Cont’d

- Melton C, White B, Davis R, Bednarczyk R, Shaban-Nejad AFine-tuned Sentiment Analysis of COVID-19 Vaccine–Related Social Media Data: Comparative StudyJ Med Internet Res 2022;24(10):e40408URL: https://www.jmir.org/2022/10/e40408DOI: 10.2196/40408

- Staphord Bengesi, Timothy Oladunni and Ruth Olusegun et al. A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets. IEEE Access. 2023. Vol. 11:11811-11826. DOI: 10.1109/ACCESS.2023.3242290

- Al-Ahdal, Tareq et al. “Improving Public Health Policy by Comparing the Public Response during the Start of COVID-19 and Monkeypox on Twitter in Germany: A Mixed Methods Study.” Vaccines vol. 10,12 1985. 22 Nov. 2022, doi:10.3390/vaccines10121985

- Thakur, N. (2023). Sentiment Analysis and Text Analysis of the Public Discourse on Twitter about COVID-19 and MPox. Big Data and Cognitive Computing, 7(2), 116. https://doi.org/10.3390/bdcc7020116

- Liu, Yinhan et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” ArXiv abs/1907.11692 (2019): n. pag.

Referances Cont’d

•P. Rajapaksha, R. Farahbakhsh and N. Crespi, "BERT, XLNet or RoBERTa: The Best Transfer Learning Model to Detect Clickbaits," in IEEE Access, vol. 9, pp. 154704-154716, 2021, doi: 10.1109/ACCESS.2021.3128742.

• Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., "RoBERTa: A robustly optimized BERT pretraining approach", arXiv:1907.11692, 2019.

•P. Rajapaksha, R. Farahbakhsh and N. Crespi, "BERT, XLNet or RoBERTa: The Best Transfer Learning Model to Detect Clickbaits," in IEEE Access, vol. 9, pp. 154704-154716, 2021, doi: 10.1109/ACCESS.2021.3128742.

• Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., "RoBERTa: A robustly optimized BERT pretraining approach", arXiv:1907.11692, 2019.

•Liu, Yinhan et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” ArXiv abs/1907.11692 (2019): n. pag.

•Majid, Abdul et al. “Sentiment Analysis on Tiktok Application Reviews Using Natural Language Processing Approach.” Journal of Embedded Systems, Security and Intelligent Systems (2023): n. pag.

THANK YOU