COMP9444 Project Summary

< Sentiment Analysis towards COVID-19 on Twitter >

< Xiao Chen (z5545738), Hiking Shi (z5467071), Kunisuke Ishigaki (z5466173),

Yuen Tak LI (z5483312), Soo-Young Moon (z5419507) >

I. Introduction

Sentiment analysis of social media has been a common task in the last several years as it allows us to survey popular opinion and better understand the impact of the spread of information and misinformation. It is safe to say that social media has had a major impact in the public's perception of COVID-19; in particular, it provides a platform for quick updates, news reports and political discussions about policies.

We have been presented with the problem statement of:

> *"Create a model that classifies the sentiment of tweets involving the COVID-19 pandemic into three classes: Positive, Negative and Neutral."*

Ideally, our model would be accurate and applicable to a wide variety of contexts and conditions, thus being potentially utilized by:

- by policy makers and businesses to inform their decisions
- by news sources to inform the community
- by health and medical organisations to influence the priority of their research

In this report, we study several neural networks, a few of which performed with great accuracy, although with their own conditions for the dataset and other factors.

II. Related Work

The related work in this problem space presents a variety of models, each with their own feature extraction and classification methods. Many papers tend to study word embedding methods such as Word2Vec and GloVe, however, there is a lot more variety in the classification methods used. Other common themes are the use of transformer-based models and some hybrid implementation of different models [Naseem, U., et. al., 2021].

Although many papers have presented many models with great accuracy, this is not a basis of how well the model solves the problem at hand. The nature of the dataset means that each model requires different datasets with different pre-processing methods to perform ideally, which means it is difficult to compare different methods to determine the 'best' model:

'Hence, there is no "one-size-fits-all" classifier solution that outperforms all other classifiers.'

- Braig, N., et al. (2022)

Thus, we decided to study several common sentiment analysis models used in this problem space, acknowledging the limitations of their implementations. Although our baseline models are Word2Vec and GloVe feature extraction methods with Bidirectional Long Short-Term Memory (BiLSTM), there is an abundance of models of, or extensions of, Bidirectional Encoder Representations from Transformers (BERT). BERT models have been known to perform with great accuracy for sentiment analysis of tweets and, although there are known limitations, namely the computational recourses required, transformer-based models such as BERT are acknowledged as state-of-the-art models in this problem space [Joloudari, J.N., et. al., 2023].

III. Methods

In our experiment, we implemented two baseline architectures as shown in Fig1. These two baseline models are originally proposed in the project reference paper. These architectures utilizing pre-trained embeddings from Google News (300B) and Twitter (27B) will be our performance benchmarks. Understanding the importance of data quality in model trainings, we developed a comprehensive training framework consisting of three pre-processing techniques combined with three sampling methods, providing us with nine unique datasets to train on our baseline models. It allows us to identify the optimal combination of pre-processing and sampling techniques which can be applied to our improved models for training and used to evaluate the performance of all models.

In our literature review, the papers indicated that transformer-based architectures such as BERT have demonstrated superior performance in sentiment analysis tasks. Therefore, for the comparison with the baseline models, we have implemented six enhanced models as shown in Fig 1.

The selection of these models was majorly motivated by the difference in how each of these models is trained with different purposes. BERT is a transformer-based model. RoBERTa is a pre-trained model which is optimized based on the BERT's architecture, while BERTweet is also a pre-trained model based on social media contents. Finally, CT-BERT is a pre-trained model based on tweets messages collected from the COVID period. Implementing these models allows us to evaluate the efficiency of models with different pre-training approaches in the COVID context for our sentiment analysis task.

IV. Experimental Setup

COVIDSENTI is the dataset we used to perform classification. The dataset is sourced from https://github.com/usmaann/COVIDSentim, which contains 90,000 unique tweets and each tweet is labeled as positive, negative, or neutral. COVIDSENTI further divides equally into three sub-datasets, COVIDSENTI-A, COVIDSENTI-B, and COVIDSENTI-C, each dataset contains 30,000 tweets. In data exploration, we explore data from the following three angles.

1. Data distribution among different labels in each sub-dataset

We observe class imbalance exists in every sub-dataset in table II, where each sub-dataset contains around 2000 positive tweets, 5000 negative tweets, and remaining majority is neutral tweets. In order to address the class imbalance issue, we may adopt two data sampling methods, oversampling and undersampling, which make dataset have the same number of tweets between different labels.

2. Most frequent words

We compute the most frequent words in all sub-dataset and analyze the result in COVIDSENTI-A, result is plotted in Fig. 1 and generate a word cloud in Fig. 2. We find out the top 5 frequent words in Fig. 1 are “Coronavirus”, “China”, “virus”, “corona” and “case”. This result gives us a brief summarization on what does tweets talk about.

3. Context structure of raw tweets

We select a few raw labeled tweets in Table III to analyze and find out tweets are often short, unstructured, and noisy. We come up with two main problems among tweets which will be pre-processed before evaluation. The first problem is that we believe some context parts have no semantic value, these context parts are hyperlinks, Non-English words, Punctuations, Hashtags, @mentions, stop words, and numbers.

The second problem is different word forms like different tenses or singular/plural should have the same semantic value and we should treat them as the same, e.g. “fails” and “fail”, “viruses” and “virus”, “treatments” and “treatment” etc.

Based on the data exploration, we picked one sub-dataset COVIDSENTI-C to perform classification due to time limitation to work on all dataset. In data sampling, we have decided to go with the original dataset without any data sampling techniques, because when we do oversampling, the accuracy rate will do better, but it may cause overfit and when we do undersampling, the data size will shrink which is not enough to perform a comprehensive analysis. Thus, we use the original dataset only. In data pre-processing, to address the problems we have mentioned above, we remove all context parts with no semantic value and lemmatize different word forms into their root form. To summarize, we use original dataset COVIDSENTI-C, with the mentioned data pre-processing to experiment.

In the baseline model, we use Word2Vec and Glove to extract features, then apply BiLSTM to predict labels. The BiLSTM architecture contains two BiLSTM layers, the first layer has 128 hidden units and the second layer has 64 hidden units, each layer is followed by a dropout layer with drop rate 0.2. Then the architecture applies softmax activation to calculate the probability and predict the output label. Early. Stopping is applied on “validation_accuracy” as a performance matric, when “validation_accuracy” does not improve in three epochs consecutively, we stop the training and start evaluating the prediction.

In BERT, RoBERTa and CTBert, they both have the same parameters of sentence_length = 128, batch = 16, learning_rate = 1e-5 and epochs = 10. In BERTweet, the only diffence is learn_rate = 3e-5. Parameters of all transformer-based models is summarized in table II.

The dataset is split into 7: 2: 1 ratio, where 70% is used in training, 20% is used in testing and 10% is used in validation.

V. Results

To round up the dataset and model selection, we used pre-processed original dataset COVIDSENTI-C, and data experiments in 8 proposed models. We have summarized the results of F1 score and accuracy in Table III. Detailed explanation and discussion will be elaborated in this section based on Table IV.

To compare the baseline models’ performance, Word2Vec and Glove embeddings with BiLSTM have achieved at least 80% accuracy rate where Word2Vec embedding is performed slightly better than Glove embedding. When we analyze the f1-score among different classes, the performance in neutral class is around 0.9 while negative and positive classes are approximately 0.7 and 0.5.

To compare the BERT model with different classification, we observe integrating BERT with BiLSTM or CNN to classify only reach around 0.75 accuracy rate, which is even lower than our baseline models. On top of that, even though their f1 scores on neutral class remain a satisfactory result of around 0.85, their f1-score on negative and positive classes are unreasonably low. However, we accomplish a magnificent accuracy rate and f1-score in an end-to-end BERT model. Both accuracy rate and f1-score in all classes reach 0.95.

Based on the result, there are three main findings from the evaluation.

1. Different data-preprocessing lead to different accuracy performance

In the original paper [Naseem, U., et. al., 2021], they conduct the COVID-19 sentiment analysis with Word2Vec and Glove embeddings with BiLSTM and BERT model. The accuracy rate in the three models is 0.749, 0.743 and 0.932 respectively. Unfortunately, all accuracy rates among different models perform a bit lower than the same proposed models. Noted that original paper only lists out the accuracy rate, it is the only performance matrix to do comparison. To compare with the pre-processing techniques used, we both remove hyperlinks, punctuations, hashtags, @mentions, stop words, and transform word into its root form. However, we additionally remove non-English words and number, which leads to a better accuracy rate, probably non-English words and numbers are considered as noisy data in this classification task and interfere the prediction.

2. Underperformance of F1-score in negative and positive classes

Among all proposed models using BiLSTM and CNN classification, f1-scores in negative and positive classes are underperforming while the majority neutral class performance steadily well. The reason could be imbalance class distribution lead to not enough features are extracted from these two classes, that would result in wrongly classifying as other labels or tends to classify as the majority class. Also, In BERT integrates with CNN or BiLSTM classification, the unsatisfactory f1-score is because BERT is designed to use as end-to-end model, mixing BERT with other classification methods limit the ultilizability of transformer model, as the resulting representation of BERT is a high-dimensional semantic vector, we should not use other classification model, that may lead to a worse f1-score.

3. BERT as state-of-art-model performs outstanding result

In the end-to-end BERT model utilizing our data-preprocessing method, results in a magnificent accuracy rate and f1-scores. This indicates that the transformer model is the most suitable model for tweet sentiment analysis, even if there is class imbalance distribution, end-to-end BERT can achieve excellent performance on f1-score in minor classes.

4. Comparison on different BERT Models

In our result, RoBERTa is lower than our expectation, the reason is model used Dynamic mask mechanism and step size, with a bigger dataset from book and wikipedia, the low efficiency due to high computation resources and lower accuracy rate obtained with limited epoches indicates we should have more resources to use RoBERTa. In BERTweet and CT-BERT, the result matches our expectations, which is higher than the initial BERT model. The reason is BERTweet is pre-trained on tweets and CT-BERT is pre-trained on COVID-19 tweets, which fit into our task on COVID-19 sentiment analysis

Among all models we suggested in paper and the findings we mentioned. We believe that outstanding performance of end-to-end BERT model is good enough to be deployed in real-world applications, as we have touched on a couple of applications. For example, policy or business making, research for health and medical organization, and news source communications etc.

VI. Conclusions

The final experimental results demonstrate that our baseline models using Word2Vec and GloVe embeddings with BiLSTM architecture have achieved satisfactory performance in most of our test cases. Nonetheless, the transformer-based models that utilized BERT architecture demonstrated stronger performance in sentiment classification tasks compared to our baseline models. Nonetheless, our models were trained specifically on Twitter data collected during the COVID period. As a result, it is inevitable that the performance of the models in sentiment analysis may vary if we applied the models with datasets in different contexts and domains.

In our future work, we will focus on two key areas. Firstly, we plan to evaluate models' robustness across different varieties of datasets, particularly financial texts such as FOMC speeches. Secondly, we aim to further include additional transformer architectures such as DistilBERT and XLNet in our future testing. The additional work provides us with an opportunity to understand how our models will perform in transferability and allows us to identify the most optimal architectural choices for sentiment analysis across different contexts and domains.

VII. References

Naseem, U., et. al. (2021). COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis

Braig, N., et. al. (2023). Machine Learning Techniques for Sentiment Analysis of COVID-19-Related Twitter Data

Joloudari, J.N., et. al. (2023). BERT-deep CNN: state of the art for sentiment analysis of COVID-19 tweets

VIII. Appendix

Fig1. Word Distribution Histrogram on COVIDSENTI-A

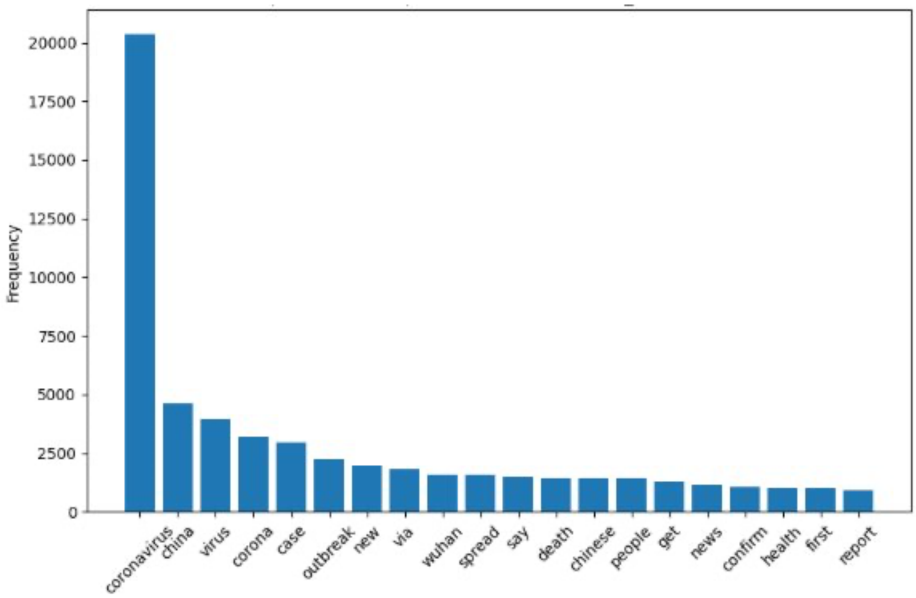


Fig2. Word Cloud on COVIDSENTI-A

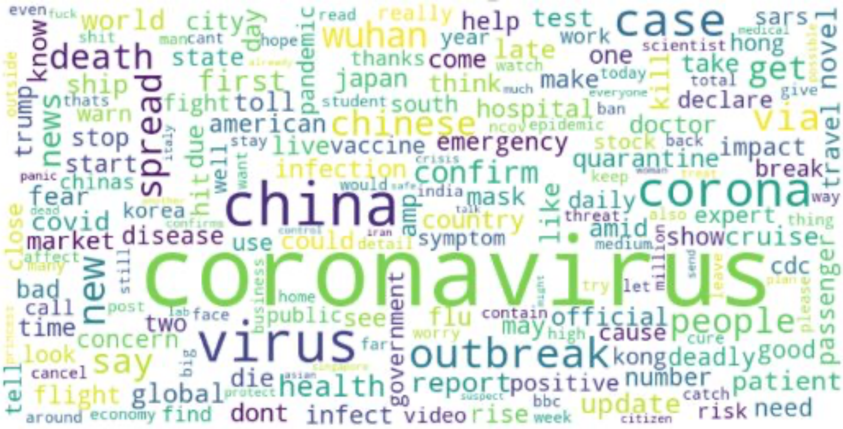


Table I. Summarized table on proposed model

Model Number	Model Type	Feature Extraction	Classifier
1	Baseline Model	Word2Vec	BiLSTM
2		GloVe	BiLSTM
3		BERT	BiLSTM

4	Enhanced Model	BERT	CNN
5		BERT (end-to-end feature extraction and classification)	
6		RoBERTa (end-to-end feature extraction and classification)	
7		BERTweet (end-to-end feature extraction and classification)	
8		CT-BERT (end-to-end feature extraction and classification)	

Table II. Data Set Distribution

DataSet\Label	**Positive**	**Negative**	**Neutral**	**Total**
COVIDSENTI-A	**1,968**	**5,083**	**22,949**	**30,000**
COVIDSENTI-B	**2,033**	**5,471**	**22,495**	**30,000**
COVIDSENTI-C	**2,279**	**5,781**	**21,940**	**30,000**
COVIDSENTI	**6,280**	**16,335**	**67835**	**90,000**

Table III. Example raw tweets

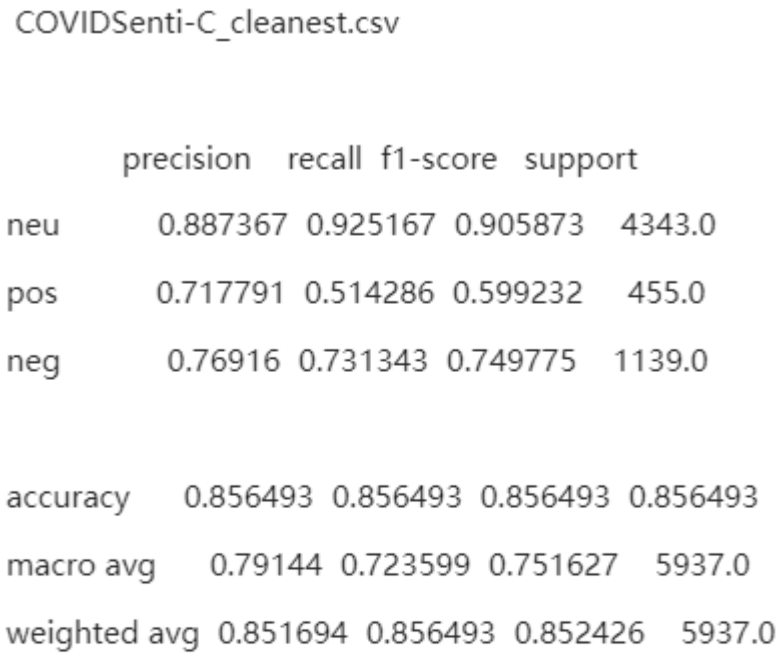
tweet	**Label**
Morrison Government fails to stop boats, despite coronavirus https://t.co/gx7f8AT8mt @IndependentAus	Negative
@Plan_Prep_Live @firefoxx66 @InfectiousDz The #coronavirus attack nervous system, today's report. So it's not just,Ä¶ https://t.co/40HoyXJ7KH	Neutral
Best Treatments For New Coronavirus Now An Urgent Target For Biotech Labs : Shots - Health News SURVIVE SEE DETAILS,Ä¶ https://t.co/H3JczrJgSQ	Positive
Why has Trump put new travel restrictions on Iran when there are 56 nations with #coronavirus? Sounds more politica,Ä¶ https://t.co/9qdV6j24f0	Neutral
@shehryar_taseer That,Äôs ⍰üíØ true , \nCorona virus \nswine flue \nBird flu in December when whole Pk is busy in Marriage,Ä¶ https://t.co/6JWBIymnyo	Neutral

Table IV. Comparison of Models

Models	**Accuracy**	**F1-score on Neu**	**F1-score on Neg**	**F1-score on Pos**
Word2Vec/BiLSTM	**0.85**	**0.91**	**0.75**	**0.60**
Word2Vec/BiLSTM (original paper)	**0.749**	**-**	**-**	**-**
GloVe/BiLSTM	**0.82**	**0.88**	**0.63**	**0.52**
GloVe/BiLSTM (original paper)	**0.743**	**-**	**-**	**-**
BERT/BiLSTM	**0.74**	**0.85**	**0.16**	**0.00**
BERT/CNN	**0.78**	**0.87**	**0.42**	**0.14**
BERT	**0.96**	**0.94**	**0.97**	**0.98**
BERT (original paper)	**0.932**	**-**	**-**	**-**

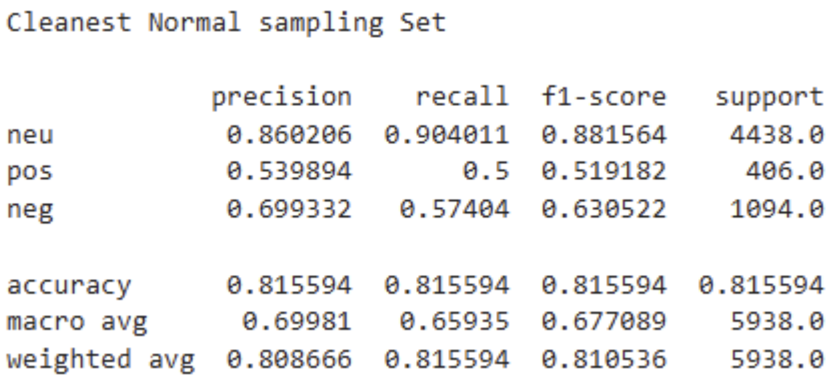
Table V. Confusion Metrix of different combinations of models

Word2vec_BiLSTM



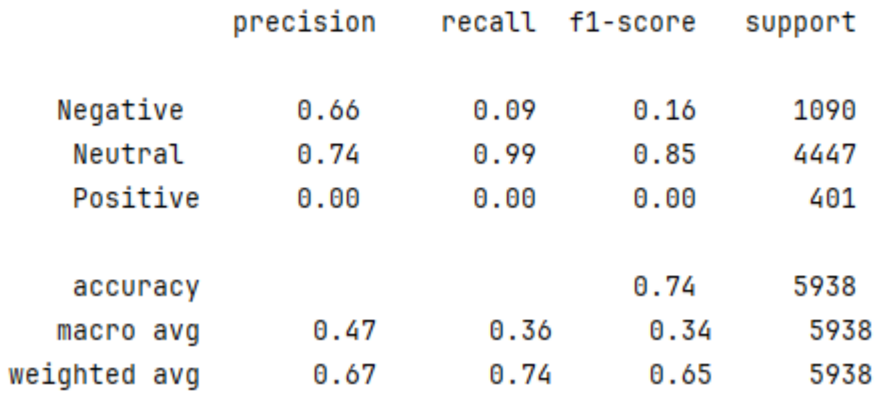
```
COVIDSenti-C_cleanest.csv


              precision   recall  f1-score  support

neu            0.887367  0.925167  0.905873   4343.0

pos            0.717791  0.514286  0.599232    455.0

neg             0.76916  0.731343  0.749775   1139.0


accuracy       0.856493  0.856493  0.856493  0.856493

macro avg       0.79144  0.723599  0.751627   5937.0

weighted avg   0.851694  0.856493  0.852426   5937.0
```

Glove_BiLSTM



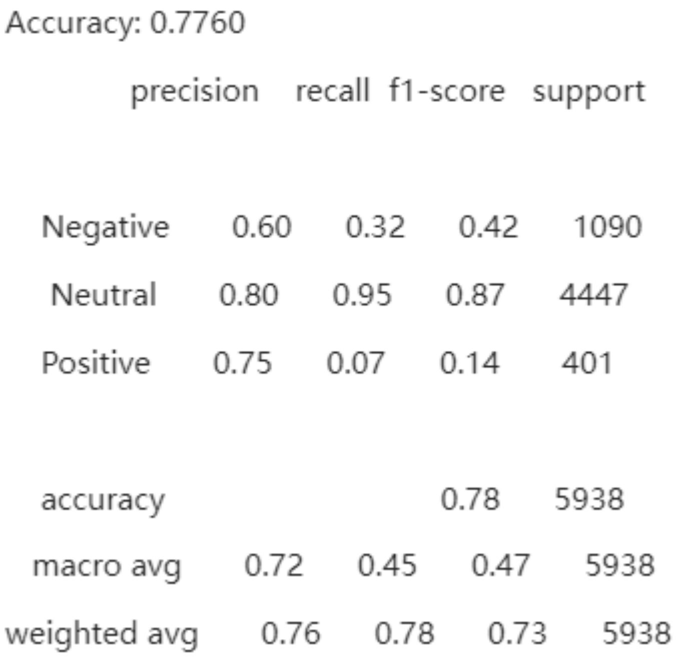
```
Cleanest Normal sampling Set

                precision    recall  f1-score   support
neu              0.860206  0.904011  0.881564    4438.0
pos              0.539894       0.5  0.519182     406.0
neg              0.699332   0.57404  0.630522    1094.0

accuracy         0.815594  0.815594  0.815594  0.815594
macro avg         0.69981   0.65935  0.677089    5938.0
weighted avg     0.808666  0.815594  0.810536    5938.0
```

BERT_BiLSTM



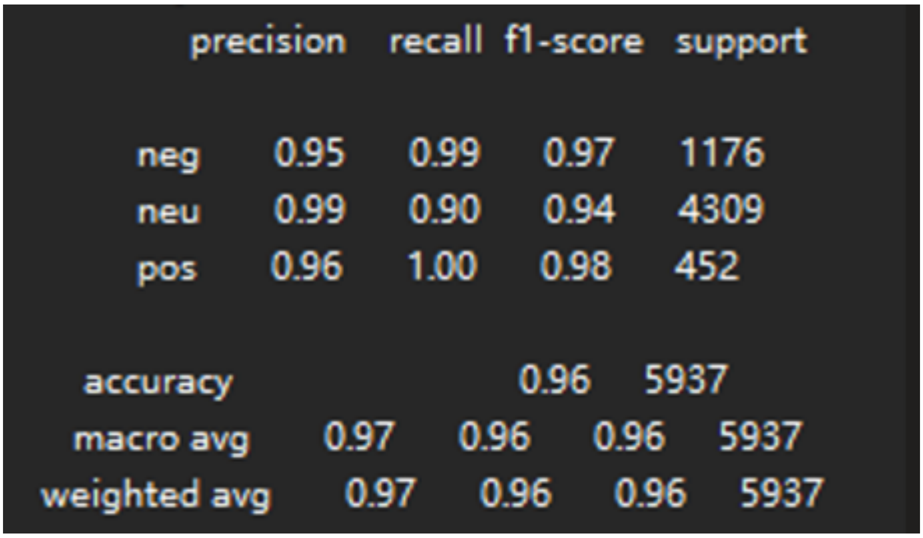
```
              precision    recall  f1-score   support

    Negative       0.66      0.09      0.16      1090
     Neutral       0.74      0.99      0.85      4447
    Positive       0.00      0.00      0.00       401

    accuracy                           0.74      5938
   macro avg       0.47      0.36      0.34      5938
weighted avg       0.67      0.74      0.65      5938
```

BERT_CNN



```
Accuracy: 0.7760
              precision    recall  f1-score   support


    Negative       0.60      0.32      0.42      1090
     Neutral       0.80      0.95      0.87      4447
    Positive       0.75      0.07      0.14       401


    accuracy                           0.78      5938
   macro avg       0.72      0.45      0.47      5938
weighted avg       0.76      0.78      0.73      5938
```

BERT



```
              precision    recall  f1-score   support

         neg       0.95      0.99      0.97      1176
         neu       0.99      0.90      0.94      4309
         pos       0.96      1.00      0.98       452

    accuracy                           0.96      5937
   macro avg       0.97      0.96      0.96      5937
weighted avg       0.97      0.96      0.96      5937
```