**Sentiment analysis on the impact of coronavirus**

**in social life using the BERT model Based on**

**Real Time Tweet Data**

Duc-Toan Nguyen1,2, Thanh Nguyen1,2, Thanh-Trung Nguyen1,2, Hoang-Phuc Do1,2, and Trong-Hop Do1,2

{19522362, 19522233, 19522431, 19522027}@gm.uit.edu.vn

1 University of Information Technology, Ho Chi Minh City, Viet Nam

2 Vietnam National University, Ho Chi Minh, Viet Nam

**Abstract.** When the whole world is immersing in the Industrial Revolution 4.0, it is impossible not to mention our Information Technology field - a field that requires great development and much more. As the volume of stored data is getting larger and larger, it is increasingly difficult to store data and process data, and the data processing speed of current technologies cannot meet it. Since then, techniques and technologies of big data processing, machine learning and natural language processing are promoted. This becomes an exciting area for research, business, or social service. In particular, the data source from social networks is a rich source of data being exploited. Within the course project, the team learns and builds an application to analyze the community's emotions about COVID-19 based on real-time Twitter data. Currently, the whole world is facing an infectious disease called coronavirus. No country is left untouched in this pandemic situation. Due to the lack of correct treatment, the disease has become a serious problem for both the government and the public. Because social distance is considered the most effective way to stay away from this disease. So, to address people's eagerness about Corona pandemic and express their views, people's trend has turned to social media very quickly. Twitter has emerged as one of the most popular among those social media platforms. By studying people's eagerness and opinions to understand their mental state, we performed sentiment analysis using the BERT model on tweets. In this paper, we perform sentiment analysis on two data sets; one dataset is collected by the tweets of people from all over the world and the other dataset contains the tweets of people in India. We have confirmed the accuracy of emotion classification from the GitHub repository.

**Keywords:** NLP . YouTube streaming data API . Covid-19 sentiment.

**1. Introduction**

Today, the Internet is gaining popularity all over the world and it is serving as a cost-effective platform for information providers thanks to the rapid expansion of social media. Some social media platforms like blogs, reviews, posts, tweets are being processed to extract people's opinions about a particular product, organization or situation. Attitudes and feelings that make up an essential part of assessing an individual's behavior are called affectives. These emotions can be further analyzed against an entity, known as sentiment analysis or opinion mining. By using sentiment analysis, we can interpret the sentiments or feelings of others and categorize them into different categories that help an organization to know people's feelings and act accordingly. This analysis depends on its expected results, for example, text analysis depending on its polarity and emotion, feedback on a particular feature, and text analysis in different languages require. the corresponding language detection request. It requires a large amount of data that may not be properly structured. Therefore, several preprocessing techniques are used to construct the final dataset from the extracted data. Moreover, real-time analytics helps us to look at the current situation and make decisions to achieve better results. The COVID-19 or Corona virus has exploded in many parts of the world and people are affected on a huge scale. This leads to a great loss of life even though people have different views on the Corona Virus outbreak. In the literature, different researchers have performed sentiment analysis on COVID-19 with different views on different datasets to classify people's opinions (Sailunaz and Alhajj 2019; Dubey 2020; Muthusami et al 2020; Rajput et al 2020; Pastor 2020; Alhajji et al 2020; Medford et al 2020; Prabhakar Kaila et al 2020).

In this article, the authors analyze the opinions of people around the world including India to understand what situations are favorable or unfavorable for them. Therefore, our main focus is to perform sentiment analysis on COVID-19 to draw some conclusions about people's opinions. Recently, it can be seen that the number of people actively participating in social media such as facebook, twitter, ... However, this work uses tweeter, a social media platform, to collect Collect public comments in the form of reviews, comments, posts on COVID-19. In this proposed model, we rely on collecting data from twitter using existing author twitter APIs and prepare two datasets i.e., world specific dataset and world-specific data. India. Sentiment analysis was then performed using different matrices like Average Likes and Re-visits over a period, Strength Analysis, Polarity & Subjectivity and Wordcloud. Along with that, the Bidirectional Encoder Representation from Transformer (BERT) model was also used to classify public opinion about coronavirus.

This article is structured as follows: Part 1 provides an overview. Section 2 describes the research related to the proposed system. Section 3 presents detailed information on the proposed system design and implementation. Part 4 provides experimental results, evaluates the accuracy of machine learning algorithms applied to the proposed system. Finally, end this paper with conclusion and future development direction.

**2. Related work**

**2.1. Deep Learning Approaches**

Deep learning is a branch of machine learning that uses multiple layers of a Neurol network to generate a mathematical model on available data. Deep Learning or Deep Learning is often referred to with Big Data and Artificial Intelligence (AI) There have been many practical applications, which are thriving under the development of special computer speed. is the computing power on the GPU and the proliferation of data along with frameworks (TensorFlow or Pytorch) make model building easier. Deep learning is part of a broader family of machine learning methods based on the learning representation of data. An observation (for example, an image) can be represented in many ways as a vector of intensity values ​​for each pixel, or more abstractly as a set of edges, areas. specific shape areas, etc. Some representations make it easier to learn tasks (for example, facial recognition or facial expressions) from examples. One of the promises of deep learning is to replace manual features with algorithms that are efficient for unsupervised or semi-supervised learning and hierarchical features. Various deep learning architectures such as deep neural networks, deep convolutional neural networks, deep belief networks, and recurrent neural networks have been applied to fields such as computer vision, automatic speech recognition, and processing. Natural language processing, speech recognition, and bioinformatics have been shown to produce very good results for a wide variety of tasks. In addition, deep learning has become a trendy word, or a trademark of neural networks.

**2.2. Apache Spark**

Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It provides development APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads—batch processing, interactive queries, real-time analytics, machine learning, and graph processing. You’ll find it used by organizations from any industry, including at FINRA, Yelp, Zillow, DataXu, Urban Institute, and CrowdStrike. Apache Spark has become one of the most popular big data distributed processing framework with 365,000 meetup members in 2017.

**3. Methodology**

**3.1. Apache Pyspark**

Apache Spark is one of the most widely used frameworks when dealing with and working with Big Data and Python is one of the most widely used programming languages for Data Analytics, Machine Learning and more.

Apache Spark is written in Scala programming language. PySpark has been released in order to support the collaboration of Apache Spark and Python, it actually is a Python API for Spark. In addition, PySpark, helps you interface with Resilient Distributed Datasets (RDDs) in Apache Spark and Python programming language. This has been achieved by taking advantage of the Py4j library. Py4J is a popular library which is integrated within PySpark and allows python to dynamically interface with JVM objects. PySpark features quite a few libraries for writing efficient programs. Furthermore, there are various external libraries that are also compatible.

A picture containing text, clipart

Description automatically generated

Figure : Apache Pyspark

**3.2. BERT (Bidirectional Encoder Representations from Transformers)**

BERT stands for Bidirectional Encoder Representations from Transformers, which is understood as a pre-trained model, which learns two-dimensional contextual representation vectors of words, which are used to transfer to problems. in the field of natural language processing. BERT has succeeded in improving on recent work in finding representations of words in digital space (space that can be understood by computers) through its context. We use BERT to build our model.



Figure : BERT Model

**3.3. Pytube - free Python library**

I use Python's pytube library to get data from comments in Youtube videos related to keywords about the Covid epidemic, create json files then combine the json files into an excel csv file to generate test file for our trained model.

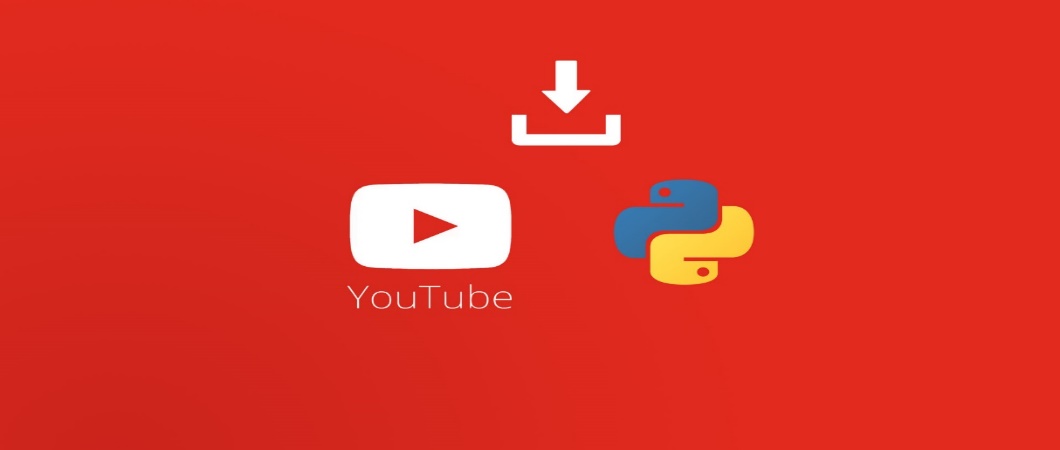


Figure : Pytube

**4. Methodology**

**4.1. Dataset**

To train the model we use a labeled dataset that we collect on github. The data file to train the model includes 2 files: nlp\_train.csv and nlp\_valid.csv and one nlp\_test.csv file so that we can compare our model with the author's labeled data file from which it shows get our model level of accuracy.

Graphical user interface, text

Description automatically generated with medium confidence

Figure : The dataset to train moel

The data file includes the columns: id, text, and emotion labels (with yes - 1 and no - 0)

A picture containing text, indoor

Description automatically generated

Figure : dataset details

**4.2. Experiment procedure and settings**

In this phase, Bi-directional Encoding Representation for a Transformer (BERT) model is used for emotion classification. The meaning of a word in a given sentence depends on the other words surrounding it. The BERT feeds all input at once to handle dependencies among words, and it is of two types: BERT-base model and BERT-large model. The BERT-base model uses 12 transformer encoders while the BERT-large model uses 24 transformer encoders. We can easily fine-tune the BERT model to get the desired results. We follow the following steps to create a mask and the encoder representation of the BERT using hugging face with pytorch library for emotion classification:

* Dividing the collected data into training and testing sets using train-test split.
* Converting the training set into respective torch tensors for the model.
* Defining the batch size to create tensors and iterators to fine-tune the BERT model.
* Training the BERT model using the model parameters and validating its accuracy.
* Evaluating the performance of the model

**4.3. Result**

We discussed how to fine-tune our BERT and prepare data for training. We fetched pre-labeled data from a Github repository that worked with similar data. The model that we built achieved a fairly high accuracy rate through the confusion matrix when compared to the author's labeled test file. Overall, the evaluated accuracy reveals the strength of the proposed model in the sentiment analysis.

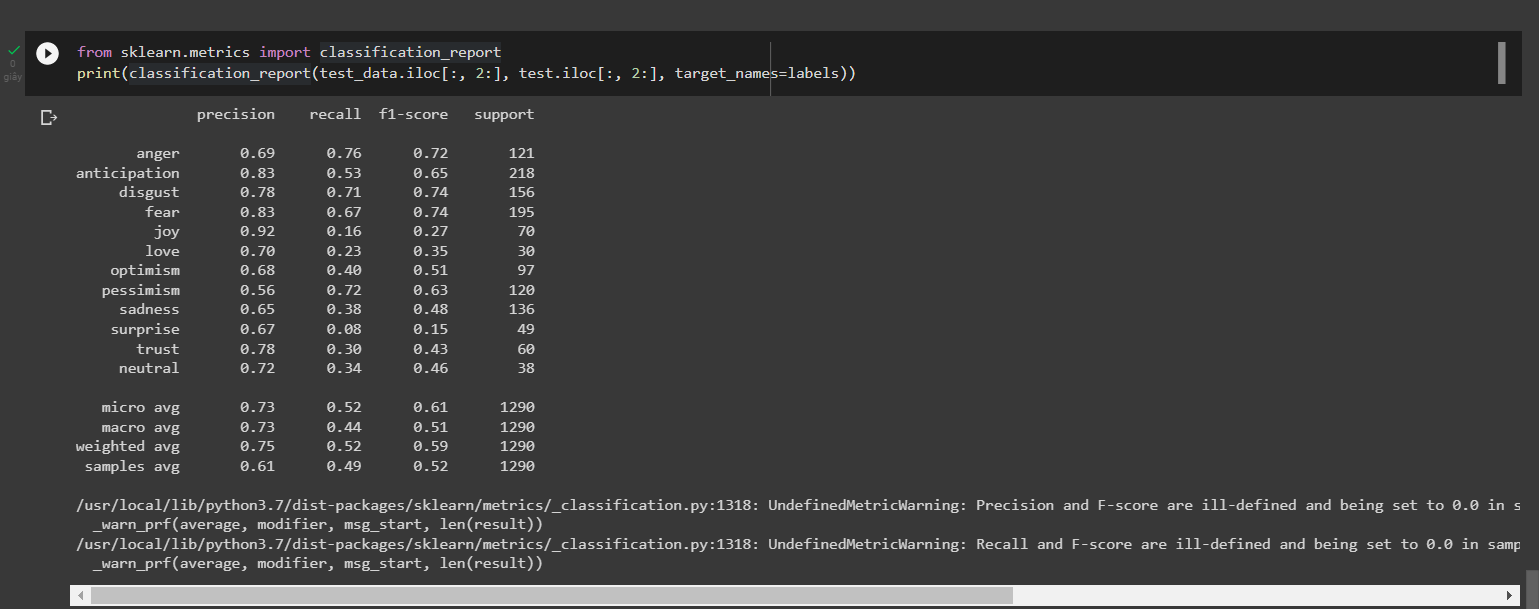


Figure : Confusion matrix

**4. Applications**

**4.1. Crawl Data**

I use Python's Pytube library to crawl the data. We use the keywords 'pandemic', 'covid-19', 'vaccine' to collect covid related data from all comments in video links from Youtube. The obtained data files are JSON files, then we combine the JSON files into CSV files to predict with the trained model.

Graphical user interface, application, icon

Description automatically generated

Figure : JSON files into CSV files

The csv file once created includes the labels. Then we process the csv file with pyspark including only selecting the id and text columns for prediction and we remove the special characters in the text column so that we can easily predict for the most accurate results.

A screenshot of a computer

Description automatically generated with medium confidence

Figure : Original csv file



Figure : Word cloud

**4.2. Prediction**

We use the trained model to predict our dataset with relatively accurate predictions of human emotions on social networks.

A screenshot of a computer

Description automatically generated with medium confidence

Figure : Prediction crawl data

**4.3. Comment**

After predicting emotions through data file crawl with the trained model, it helps us to recognize human emotions during the Covid-19 pandemic on social networks. We've found that people are enjoying life at home during the lockdown, but people's negative rates are increasing as the negative rate outweighs the positive rate. Thereby seeing the diverse emotions of people to have social solutions. Especially, the highest anticipation rate for me is that everyone expects the pandemic to pass, and everyone will have a better life.

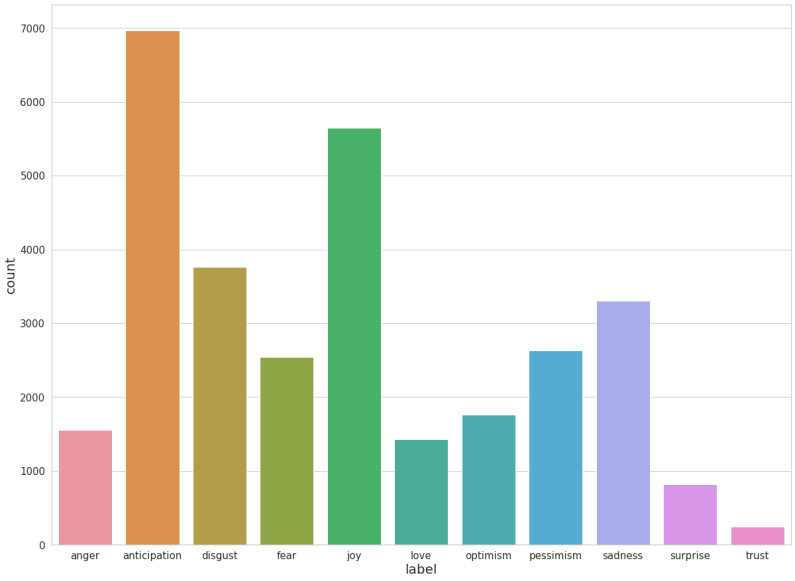


Figure : bar charts with predictable emotional labels

**5. Conclusion and future scope**

The BERT model is used in this study to do sentiment analysis on Twitter data sets. The data collection is divided into sections based on where tweets from users in India and the rest of the globe were posted. The gathered tweets were written during a period when the coronavirus was negatively impacting people's personal and professional life worldwide. Additionally, it has been shown that Twitter users from India tend to communicate more positively and less frequently propagate negativity. The examination of emotions shows if a country's government's actions were successful or unsuccessful under certain conditions. Further, it can be observed that the efficacy of taken measures for the people of a country that can support the government in taking more significant decisions to tackle novel coronavirus. The overall performance of the proposed model in terms of the validation accuracy on the collected data sets is approximately 80%.

**6. References**

1. BERT Sentence Embeddings (Base Cased)- Spark NLP Model. (n.d.). Johnsnowlabs. <https://nlp.johnsnowlabs.com/2020/08/25/sent_bert_base_cased.html>

2. Singh, M. (2021, March 19). Sentiment analysis on the impact of coronavirus in social life using the BERT model. SpringerLink. <https://link.springer.com/article/10.1007/s13278-021-00737-z?error=cookies_not_supported&code=d24f8669-a758-4b74-8b38-b6c6af189e4a>

3. Savan V (2020) savan77/EmotionDetectionBERT. GitHub. <https://github.com/savan77/EmotionDetectionBERT>

4. What is Apache Spark? | Introduction to Apache Spark and Analytics | AWS. (n.d.). Amazon Web Services, Inc. <https://aws.amazon.com/vi/big-data/what-is-spark/#:%7E:text=big%20data%20workloads-,What%20is%20Apache%20Spark%3F,against%20data%20of%20any%20size>.