Investigating Urdu News Headlines of Jang News about Provincial and Federal Government measures to tackle COVID-19: using Sentiment Analysis to detect the quality of measures taken

Muhammad Qasim Khan
Computer Science
Habib University
Karachi, Pakistan
mk05539@st.habib.edu.pk

Abstract—Numerous pandemics have arrived in human history and have been recorded by the people of their time. In order to prevent the virus from causing huge amounts of casualties, the leaders elected in charge take decisions to limit the effect of the virus. In the 21^{st} century, the COVID-19 pandemic has affected many countries and has led to many deaths, many economies weakening, prompting the respective governments to take actions to contain these devastating effects. News being widespread in this generation and in the digital age being easily accessible gives us a look at the government's policies being implemented to contain the effects of the coronavirus pandemic. These measures can have either an impact - a positive or negative one - or, not even have much of an impact. This project will be assessing the quality of measures taken by the provincial and federal governments of Pakistan to help prevent the spread of coronavirus since it started, as reported in the headlines by the premier Urdu news outlet in the country, Jang News using sentimental analysis tools.

Index Terms—COVID-19, sentiment analysis, government measures, news headlines, data collection

I. Introduction

COVID-19 has taken the world by storm and as such has had many far spreading effects on different spheres of life. The social and economic spheres have been particularly impacted the most due to huge cut-down in movement from one area to another while following strict protocols on social distancing. It definitely reduces the pull of social areas such as shopping malls, cafes, restaurants, tourist spots etc. Consequently such policies were to be formed and implemented by governments across the world that would help reduce the impact that the pandemic was leaving them with, while also reducing human losses from the virus as much as possible.

Any measure/initiative taken by any government will not always be well received or widely resented. Some measures also fall in the category that they don't bring any real substantial impact or change or start of a chain that ultimately changes something for the better (or for the worse). Public access to what those measures are, what they entail, why they were taken, and how long will they remain; all this information comes in the form of news. Whether that news be a newspaper in physical print, an online article, or some social media post. It's bound to give information about a measure taken. Similarly, the measures taken by the Pakistani provincial and federal governments for containing the spread of coronavirus and the side effects it entails (health-wise, socially, and economically), have been reported in the country's premier news outlets. Examples such as Dawn, Jang, The Express Tribune, and many others have been since day one of the pandemic in Pakistan, been reporting on day-to-day basis about the total number of cases, while simultaneously using that data to present analysis and give advice on how to tackle the situation[1]. The headlines in any of these newspapers gives an idea of how the measure taken is effective - if so, was it positive or negative - , or ineffective. For this research, we will be analyzing the news headlines of Jang News, particularly the second headline, to investigate if the measure/action taken by the provinvial/federal government falls in the category of good, bad, or neutral.

Sentimental analysis, a tool of Natural Language Processing will be the most helpful in extracting the data from headlines and analyzing them to indicate a measure's quality.

II. Related Work

There are recent examples of research work into investigating news headlines during the COVID pandemic, most notably in English. IN on research, the news headlines across four nations: India, UK, South Korea, and Japan. They analyzed the country with the most focused news during the pandemic and also found the country which had the highest negative sentiment[2]. They gathered their data by scraping news articles from the web using the Beautiful Soup Library by Python, and then used topic modelling to create a corpus to later classify. They used 3 popular libraries from Python to do sentimental analysis on the headline data, to classify it as positive or negative[2].

Another research was conducted on the sentiments and emotions evoked by the news headlines of COVID-19 outbreak. They gathered news headlines from a publicly available source by Johns Hopkins and then broke down the headlines into text files and then removed numbers, white-spaces, applied lowercase. They then applied stemming and then built a document term matrix, where rows are the documents, terms as the columns, and the frequency of words as cells of the matrix [3].

A research carried out on Th Washington Post and The New York Times was done by using Lexis-Nexis to retrieve articles and then use topic modelling to break down the articles into a word corpus that would help in detecting the frames that the news best fit in regarding the COVID-19 pandemic [4].

There haven't been many notable researches on Urdu language for sentiment analysis, let alone analysis of news headlines. However a research was carried out into generating a corpus of Urdu language to act as a data-set to serve for NLP purposes. The data was gathered manually and then annotated manually by three native speakers.[5]

III. Dataset Operations

A. Data Sources

As indicated in the title of the paper, the newspaper focused on will be Jang News website with the focus on COVID-19 headlines. Narrowing down this focus to specifically headlines about measures taken by the Federal and Provincial Pakistani Government(s), against the spread of the virus will help in this research.

The reason to choose Jang News as the primary source of data is due to their reliability and popularity amongst the general masses as a news outlet. Also worthwhile adding is that their website is not a scan of news headlines like other news websites.[6]

At the beginning of data collection, it was assumed that the data collected would be sufficient fro training and testing purposes. Unfortunately the data was found to be inadequate at the time of training so instead another data source was opted for training: a data source that comprised of tweets with tags about what kind of news do they fall in: N for negative, P for positive, O for objective.[7] The dataset was prepared by a group of researchers. It can be found one of the researcher's GitHub profiles.[7]

The reason this data was chosen because it is also divided into three categories and is tagged N,P or O as the data indicates. The tweets include news headlines as well, which removes any possibility of the data being non-valid. This dataset contains 1000 points.

B. Data Preprocessing

As the focus of the research is to identify whether the headlines are indicating a positive/Good measure, a negative/Bad measure, or a insignificant/neutral measure, these tags were allotted manually by the author of this paper. This was done on the basis of what the author inferred fell in the category of a Good measure/Bad measure/Neutral measure.

This raises questions about the validity of the data as one person can't be trusted for this task and previous research on such data collecttion and tagging was done by a group of people who are domain experts and cross-checked each others tagging. In order to overcome this problem the following steps were taken.

The data taken from Jang News website was stored in Google forms and then sent to domain experts in the Urdu language in order to increase the validity of the data.These domain experts included professors of an Urdu course at Habib University (Jehan-e-Urdu) and teachers at local Madrassahs. The data verification process was done through multiple forms as each person was sent different forms in order to increase the number of responses. Each form has 7 randomly selected headlines. This is so that a variety of headlines gets tagged while also making the form short to fill for the users. The forms contain the headlines in Urdu, copied from the website, and accompanied by translations in English.This allows for a wider response audience so that any doubts on the validity of data are cleared. Each headline is given a tag by the author. The forms asks users if they agree or disagree with the tag. Upon disagreement they are prompted to select the tag they find fits the headline more accurately.

Once a sizable number of responses are there for example 20 or more, the headlines are then entered into a CSV file with the most selected tag that was given. This removes any sort of bias from the author on the tagging. In case some headlines are divisive in the sense that they are falling in two tags equally, then they are discarded so as to not let the data be affected by bias.

C. Further Processing of Data

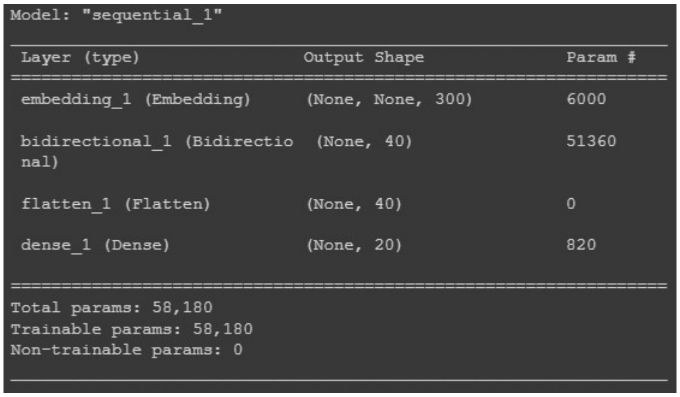
The data stored in CSV files for testing is still not suitable for the model at this point due to not being space separated. Unlike the GitHub dataset, where the tag is separated from the headline by a separator. This prevents NaN values from affecting the data. So a space separator was then placed between the tag and headline so that it can be inputted into the model.

IV. The Model Used

Considering the nature of the task at hand and the kind of data being used, an LSTM model was best suited for this as it has the capability to capture long term dependencies found in language and would also help tremendously in identifying the right tag for any headline. This model is a sequential model and consists of layers:

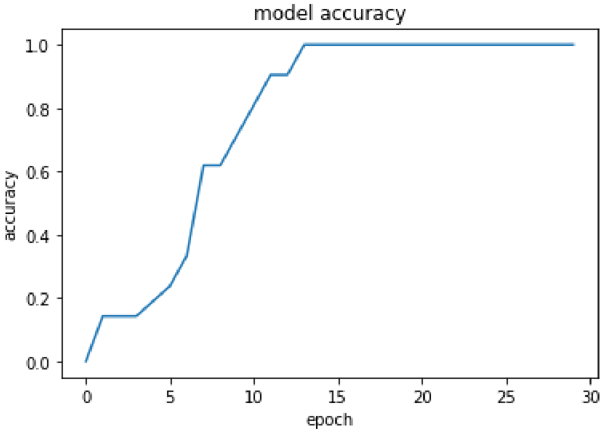
- Embedding Layer
- Bidirectional Layer
- Flatten Layer
- Dense Layer

This image here shows the layers and the parameters while training:

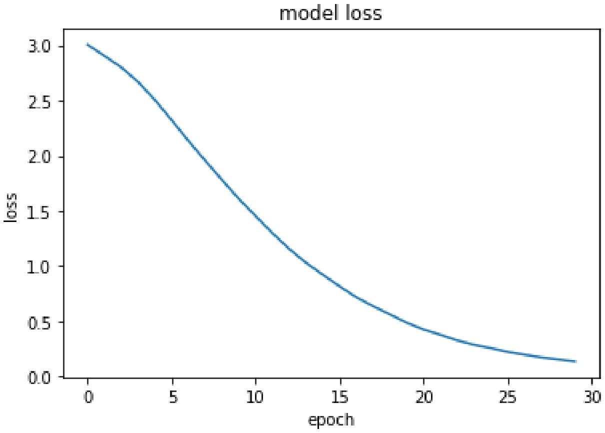


```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_1 (Embedding)     (None, None, 300)         6000

 bidirectional_1 (Bidirectio  (None, 40)               51360
 nal)

 flatten_1 (Flatten)         (None, 40)                0

 dense_1 (Dense)             (None, 20)                820

=================================================================
Total params: 58,180
Trainable params: 58,180
Non-trainable params: 0
_________________________________________________________________
```

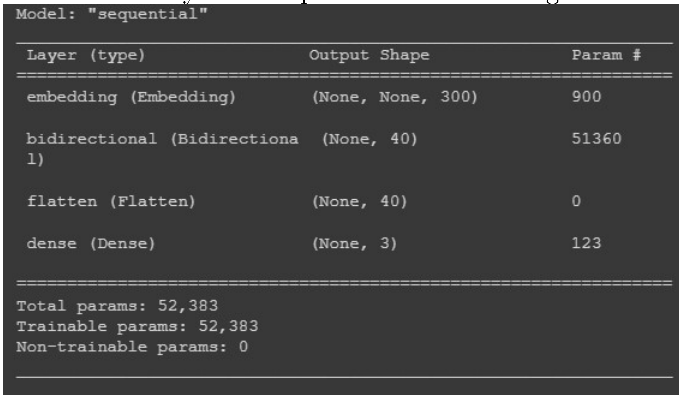
Here is the model accuracy vs epochs graph:



Here is the loss vs epochs graph:

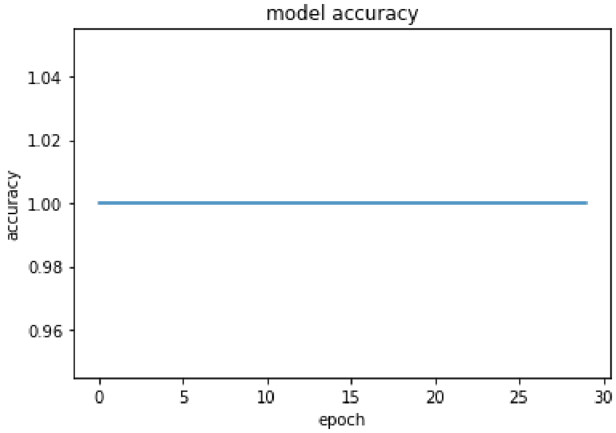


Here are the layers with parameters for testing:

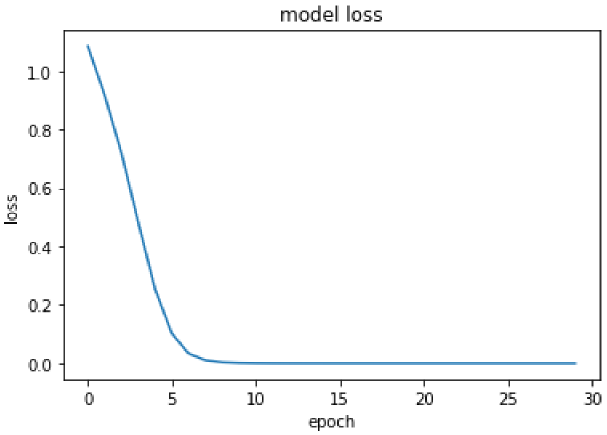


```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 300)         900

 bidirectional (Bidirectiona  (None, 40)               51360
 l)

 flatten (Flatten)           (None, 40)                0

 dense (Dense)               (None, 3)                 123

=================================================================
Total params: 52,383
Trainable params: 52,383
Non-trainable params: 0
_________________________________________________________________
```

Here is the testing accuracy vs epochs graph:



Here is the testing loss vs epochs graph:



A. Explaning the Model

The Model is an LSTM model with 4 layers. The LSTM used is Bidirectional and the Embedding layer used had a vocabulary size of the Tokenizer's word index length + 1. The loss used is Categorical Cross Entropy as we are focusing on three categories or tags: Good, Bad, and Neutral. The optimizer used is Adam with a learning rate of 0.01.

The model ran 30 epochs for both the training and testing datasets and in the training graph, it's visible that it went up in accuracy. It trained well enough to give 100 percent accuracy for the testing data. The model uses also a padding layer on the data and flattens it. The activation used is Softmax to prevent gradient vanishing.

V. Improvements to be Made

A. Data Collection

The data collection part of this research was not organized well enough and consumed most of the part of the research. Focusing on one news source narrows it down but ultimately, it does not allow for a wider variety of data to train on the model on so that it can predict from any news source. Scraping the data could have been used instead of manual collection as the Jang News website is not in scan form. An argument can be made that one needs manually annotated data before to train the model with so that when a headline is inputted into it, it makes an accurate prediction. Also to add that previous researchers have collected and made data but they are not in the form of three classes as Good, Bad, Neutral. Rather they focus more on the type of news being reported for example: local, international, sports, entertainment etc.

B. Model Selection and Training

The model chosen had 4 layers and was suitable for the task at hand. However, compared to previous research carried out, this does not seem to involve a level of complexity that would be beneficial for future applications or researches. If the purpose of the project were to predict the tone of the article or identify the kind of articles the author writes or the newspaper publishes based on one article. Then the most advanced transformers can be used to work on this and it could produce great results. Perhaps the reason this could not be carried out is due to a small time-frame for this and lack of organization by the author.

References

[1] Shazia Aziz, Dr. Akifa Imtiaz Rabea Saeed | Krisda Chaemsaithong (Reviewing editor) (2022) Framing COVID-19 in Pakistani mainstream media: An analysis of newspaper editorials, Cogent Arts Humanities, 9:1, DOI: 10.1080/23311983.2022.2043510

[2] P. Ghasiya, K. Okamura: Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach

[3] F. Aslam, T.M. Awan, J.H. Syed, A. Kashif, M. Parveen: Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak

[4] Austin Hubner (2021) How did we get here? A framing and source analysis of early COVID-19 media coverage, Communication Research Reports, 38:2, 112-120, DOI: 10.1080/08824096.2021.1894112

[5] Khan, L., Amjad, A., Ashraf, N. et al. Multi-class sentiment analysis of urdu text using multilingual BERT. Sci Rep 12, 5436 (2022). https://doi.org/10.1038/s41598-022-09381-9

[6] J. pk, "https://jang.com.pk/hot-topics/coronavirus," jang.com.pk, 2022. [Online]. Available: https://jang.com.pk/. [Accessed: 07-Dec-2022].

[7] M. Y. Khan and M. S. Nizami, "Urdu Sentiment Corpus (v1.0): Linguistic Exploration and Visualization of Labeled Dataset for Urdu Sentiment Analysis," 2020 International Conference on Information Science and Communication Technology (ICISCT), 2020, pp. 1-15, doi: 10.1109/ICISCT49550.2020.9080043.