Sentiment Analysis to Classify Depression from Twitter Using Tweets Before and After COVID-19 Through Different NLP Approaches

Camille Porter, Olof Johansson, Savya Sachi Gupta
Chalmers Institute of Technology, Sweden
gusportca@student.gu.se, olojohan@student.chalmers.se, savya@student.chalmers.se

Abstract

We aim to understand the affect of the COVID-19 pandemic on mental health, specifically depression, by performing sentiment analysis on 'tweets' shared on the social media service Twitter. We selected the United Kingdom and Ireland for our analysis as they had a government instituted lockdown across the country, which provides us with a definitive date as a reference point to gauge trends before and after. In order to understand how a lockdown affects depression, we sampled tweets from these locations and trained two different models to detect depression in tweets—an LSTM model, and the DistilBERT model, which a condensed form of BERT. We scraped 5,000 tweets for each week that we measured during multiple time periods. The LSTM model performed better than DistilBERT yielding an accuracy of 0.94 as compared to 0.90 for DistilBERT. We found a 2-3% bump in the level of depression two weeks after lockdown started, but no long-term changes.

Note: Figure sizes in the paper are small, so we have also attached a larger version of them as an appendix.

Introduction

The COVID-19 pandemic quickly spread around the world during 2020, causing large changes to the way people all over the world are living their lives. One of the most dramatic changes was the government instituted lockdown, where everyone was required to stay indoors for an extended period of time and only leave home to buy food or carry out essential businesses. This sudden change in lifestyle and restrictions could possibly affect a persons physical and mental well-being. People could possibly develop feelings of loneliness, isolation, boredom, frustration and could potentially lead to depression. It would be interesting to explore how the pandemic affected people and whether a lockdown led to an increase/decrease in feelings of depression.

In the current digital age, many of people take to social media to express their feelings, so it was determined to be a good starting point for our analysis, more specifically, the micro-blogging service, Twitter. Twitter is the ideal scrape-able source of time-stamped information about current world events. Here, users post about the news, their opinions, and things that are currently happening in their lives.

Now, in order to answer our question, we source short messages called 'tweets' from Twitter, and analyze them to detect if the tweet is indicative of depression or not. Our hypothesis is that being in a nation-wide lockdown will raise the amount of depressive tweets posted to Twitter by a noticeable amount.

There is a lot of research done previously regarding sentiment analysis using Twitter and it has been further extended to specifically detect depression through tweets (Singh and Wang). Social media and mental health are connected together in different ways (Kumar, Sharma, and Arora 2019). For example, active social media users could be distressed by negative interactions or alternatively use social media to vent their feelings. Therefore, services like Twitter are a good sample for our experiments.

Since we are trying to understand trends for depression in relation to the pandemic, it is useful to analyze tweets before the pandemic and after a state of emergency is declared by the government and restrictions of some form are imposed on the people. To simplify our task, we wanted to select a location that was mostly English speaking, as most language processing methods are written for English. We also wanted a location with a definitive lockdown date. This ruled out the United States of America, where the rules were different for every state (creating further confusion). Thus, the United Kingdom and Ireland were selected. There was a suggested lockdown starting 16 March 2020, when the Secretary of State for Health and Social Care informed the House of Commons that all unnecessary social contact should cease (Fact 2020). But on 23 March 2020 Boris Johnson, the Prime Minister, told everyone they must stay home and asked businesses to be closed. Ireland had a slightly different lockdown date–27 March 2020 ((Wire), Chaney/Collins), and REUTERS) 2020). This date is close enough for us to consider 23 March 2020 as the official lockdown date that we will use. We are expecting to detect some rising depression scores a little earlier because some people began to travel less before the official lockdown. For our experiments, we will be comparing the depression trends for a period of three months before and after the beginning of lockdown (23 March 2020), as well as the same time period the previous year. This will give us a control to see whether the season makes a difference in the depression score.

To source tweets from Twitter, we use the 'Twint' package (C. Zacharias) in Python that enables us to source data based on cities, or a specified distance radius around a longitude/latitude, language, keywords, date range etc. and the collected data was passed through neural network models.

Method

Collecting the Dataset

In order to maximize the locations included in our analysis, we used the longitude/latitude method. Using Google Earth, we determined that the Isle of Man is approximately at the center of the UK. Using the measurement tool on Google Earth, we found that a 550 kilometer circle around the Isle of Man covered all of UK and Ireland without touching France.

In order to train the model we sourced tweets in two halves. One half of tweets related to depression, and another half of tweets that were non-depressive. This enables accurate labeling of data that helps train the model. The words we used to label depressive tweets were : depressed, lonely, sad, depression, tired, and anxious. The words that were labeled non-depressive were: happy, joy, thankful, health, hopeful, and glad. For each word specified above, we scraped 1,000 tweets, resulting in a training set size of 12,000 tweets. 80% of the tweets were selected for training and 20% for testing. Subsequently, for analyzing, we sourced 5,000 tweets per week for three different time periods; three months before and after the initial UK lockdown of 23 March, 2020, same period the year before and then a six months period starting from three months after the initial lockdown up to 17th of December, 2020.

Preprocessing

Prior to the training process, the collected tweets had to be cleaned, annotated, and combined as many raw collected tweets contain emojis, URLs, mentions and non-alphabetic tokens. Thus, a pre-processing step cleaned the tweets as above and also removed possible punctuation marks. Next, a vocabulary based on words from the training data was built which consisted of two dictionaries for encoding and decoding the input text data. Moreover, the encoding process also padded the input sequences to be the same length by adding a specific padding token. Additionally had the labels also be encoded as the training data were labeled with either *depressive* or *not-depressive*. These two categories were encoded into corresponding integers of 0 and 1.

LSTM Model

The initial model used for the sentiment analysis was a standardized embedding, that converts input integer tokens into real-valued data, followed by a *Long-Short-Term-Memory* (LSTM) model with an added feed-forward layer with dropout as output layer. As RNN models has been widely used in natural language processing due to the dependencies through time they obtain and thus memorize the dependency between the different words in the input text sequences (Patel and Tiwari 2019), this approach was more favorable than methods like Naive Bayes and CNN's. The choice of LSTM was motivated by it addressing the issue of vanishing gradients better than a basic RNN or GRU network. It does this by incorporating nonlinear controls into the RNN cell that are data dependent (Sherstinsky 2020). This, in combination with it being better for longer dependencies than the other two architectures, motivated the use of the LSTM architecture in particular.

The LSTM architecture consists of a RNN cell (hidden gate) and three other gates called input gate, forget gate and output gate that regulates the flow of information inside the cell. These gates seperates the LSTM from other RNN models such as the gated recurrent units (GRUs) which lacks the output gate (Murthy et al. 2020). The algorithm for the forward pass of a LSTM, when implemented in the pytorch framework, is defined as

$$\begin{aligned}
i_t &= \sigma\left(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}\right)\\
f_t &= \sigma\left(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}\right)\\
g_t &= \tanh\left(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}\right)\\
o_t &= \sigma\left(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}\right)\\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t\\
h_t &= o_t \odot \tanh\left(c_t\right),
\end{aligned}$$

where $\odot$ is the element-wise dot product (Hadamard product), W are weight matrices for corresponding gate connections, b their corresponding bias vectors, x_t input at time step t, h_t the hidden states and σ the activation function (Torch Contributors 2019). The initial states of h_0 and c_0 were zero initiated.

The output of the feed-forward output layer was then fed into a sigmoid function, also used as the activation function in the LSTM network, defined as

$$\sigma(x) = \frac{1}{1+\exp\left(-x\right)}.$$

As the problem is binary classification, the loss was then computed by the binary cross entropy loss (BCE) defined as

$$\mathcal{L}_{BCE}\left(y, \hat{y}\right) = -\frac{1}{N_B}\sum_{i=1}^{N_B} y_i \log\left(\hat{y}_i\right) + \left(1-y_i\right)\log\left(1-\hat{y}_i\right),$$

where N_B is the batch size, y_i is the true label and $\hat{y}_i$ the output value. Lastly, the decoding of the output from the predictions on the test data was done as

$$f_{decode}\left(\hat{y}_i\right) = \begin{cases} \textit{depressive}, & \text{if } \hat{y}_i < 0.5 \\ \textit{not-depressive}, & \text{if } \hat{y}_i \geq 0.5 \end{cases}$$

The hyperparameters used for the training process, including those for the optimizer, epochs and batch size were

```
emb_dim = 64
rnn_size = 64
nr_layers = 1
dropout = 0.5
lr = 1e-3
batch_size = 64
n_epochs = 30
decay = 1e-5
patience = 5
loss_fn = nn.BCELoss()
optimizer = torch.optim.Adam
```

where patience is the parameter adjusting the early stopping function used to prevent overfitting.

DistilBERT Model

In addition to our LSTM model, we wanted to try a state-of-the-art transfer learning model. We decided to use DistilBERT, a smaller version of Bidirectional Encoder Representation from Transformers (BERT). BERT is a bidirectional LSTM with multi-head attention.

The 144 attention heads of BERT focus on certain relevant parts of the text. The input for attention is the hidden layer $h = [h_1 ... h_n]$. Each h_i is transformed into query (q_i), key (k_i), and value (v_i) vectors. Then the softmax-normalized dot product of the query and key vectors becomes the attention weights α.

$$\alpha_{ij} = \frac{exp(q_i^T k_j)}{\sum_{l=1}^{n} exp(q_i^T k_l)}, \qquad o_i = \sum_{j=1}^{n} \alpha_{ij} v_j.$$

The output of the attention head is weighted by the value vector. (Clark et al. 2019) This allows the model to focus on the most important words.

BERT begins with pretraining on two tasks (masked language modeling and next sentence prediction) simultaneously. This gives the model a good grasp of language. We then fine-tune the model for our data. This consists of adding at least one more layer to the output layer to fit our task, but minimally changing the weights inside the BERT layers.

DistilBERT uses the same architecture as BERT, but it has fewer layers. It is 40% smaller and 60% faster than BERT, while retaining 97% of BERT's performance. (Sanh et al. 2019) A bidirectional LSTM is supposed to work well with longer texts, because it can remember words from far away. However, we have very short texts, so we are not sure this will be so crucial for our data.

For this method, we used a binary cross entropy with logistic loss, which combines a sigmoid layer with a BCE loss. We used the same method of decoding the output as above.

The hyperparameters used for the training process were:

```
batch_size = 200
learning_rate = 1e-05
num_epochs = 8
loss_fn = nn.BCELogitsLoss()
optimizer = torch.optim.Adam
```

Even though DistilBERT is smaller than BERT, it was still slow to run on the Google Colab GPU. The final layers we added on top of DistilBERT base were a dropout layer with dropout probability of 0.1 and a linear layer that enabled us to end with 2 classes. We tried having the dropout be 0.2, but it lowered the validation accuracy.

Experiments

We conducted experiments using two different machine learning approaches - an LSTM network and the pre-trained DistilBERT model as described in the above section to analyze the change in percentage of depressive tweets over time, before and after the initial UK lockdown date. The experiments involved two processes of data collection of tweets, first for the training data and secondly for the test data used for the analysis. After the data went through the pre-process of cleaning and tokenizing, the two models were trained and evaluated on the training data and used to analyze the sentiment on the test data for the collected time periods. These are discussed in detail in the next section and we also use the LSTM model to further perform analysis for different time periods to understand the trends and look out for any patterns that stand out.

Results

Training

The results of the training progress and accuracy metrics for the LSTM model are shown in Figure 1. The highest validation accuracy for this model was 0.94.

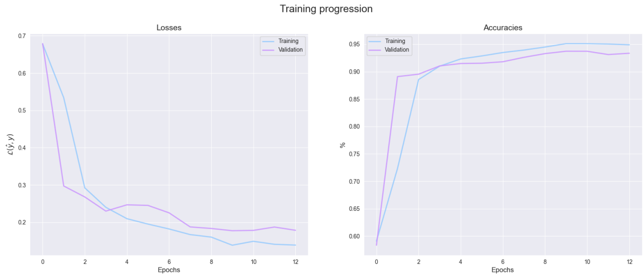

Figure 1: *Training progression of the LSTM model with the losses of the training and validation (**left**) and and the corresponding accuracies (**right**).*

The results of training the DistilBERT model are shown in Figure 2. The highest validation accuracy achieved by the DistilBERT model was 0.90.

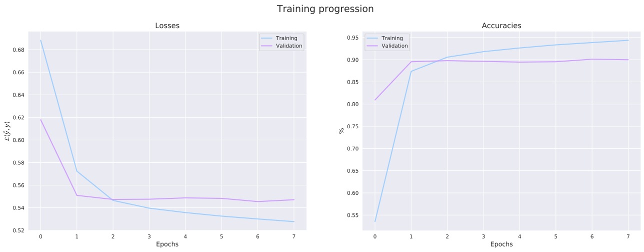

Figure 2: *Training progression of the DistilBERT model with the losses of the training and validation (**left**) and and the corresponding accuracy (**right**). We can see the model becomes over-trained very quickly.*

Since our LSTM model has a greater validation accuracy, we will select that model performing our time period analysis in the next section.

Weekly Analysis

Using the LSTM Model, we analyzed different time periods. Fig. 3 shows the percentage of depressive tweets for a time period starting from 3 months before the UK lockdown date to 3 months after the lockdown date was announced (2019-12-24 to 2020-06-23). Each bar in these graphs represent a time-span of one week. This is the year when the Covid-19 pandemic was causing problems. The middle column is the beginning of the lockdowns. The trend isn't smooth, but we can see some weeks have a rise in the level of depression immediately following the lockdown. When the lockdown begins on 03/24, the level of depression is 21.64%. The week after is lower, 20.68%. Then the next week it rises to 23.1% and there is another rise two and three weeks after that to the highest level of the year, 23.32%. After this the levels of depression stay low. It looks like there is a small immediate rise in the levels of depression, but it goes away as people become used to the situation.

Similarly, we also found the percentage of depressive tweets for the same set of six months but for the previous year, i.e 2018-12-24 to 2019-06-23, represented in Fig. 4. This is a control group so that we can see if the changes that we saw in the level of depression normally happen at the same time of year as they happened during 2020. The levels of depression during this time period are highest at the beginning (during the winter), but they do not rise during March, April, or May. This is probably due to seasonal depression.

An additional analysis for the remaining time period in 2020 was also performed to understand if there were any changes in percentage of depressive tweets after an extended period of restrictions, pandemic, and change in normal routine. This is represented in Fig. 5 for the time period 2020-06-23 to 2020-12-17.

It is also interesting to be able to visually compare the weekly percentages side-by side as represented in Fig. 6. Each color represents one of the time periods described above. We can see that 2019 had the highest levels of depression overall, which was unexpected.

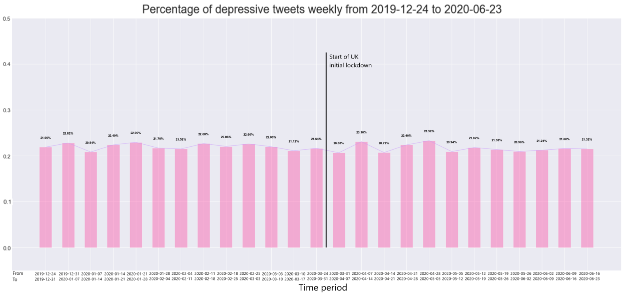

Figure 3: *The percentage of depressive tweets, predicted by the LSTM model, from the 24th of December, 2019, to the 24th of June, 2020. This is the lockdown time period.*

It is interesting to note that the overall levels of depression during 2019 were higher than 2020. This is quite unexpected, given the state of the world during both years.

Conclusion

We are able to see a rise in the level of depression after the lockdowns began. The levels of depression did not rise during the same time period the previous year, so it is unlikely to be due to seasonal change. The trend was not as large as we expected it to be, given how traumatic living through a lockdown feels.

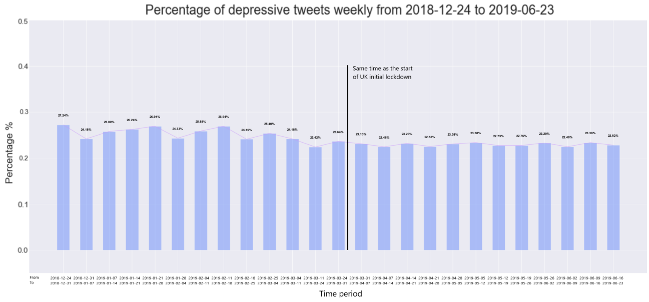

Figure 4: *The percentage of depressive tweets, predicted by the LSTM model, from the 24th of December, 2018, to the 24th of June, 2019. This is the control year.*

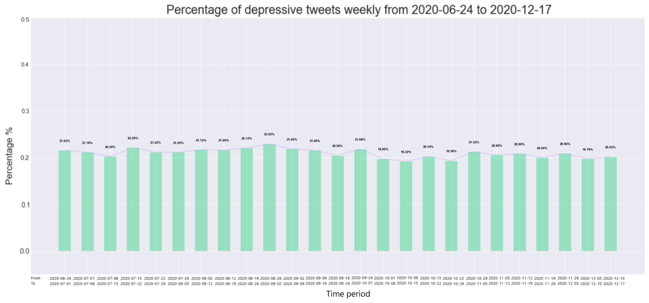

Figure 5: *The percentage of depressive tweets, predicted by the LSTM model, from the 24th of June, 2020, to the 17th of December, 2020.*

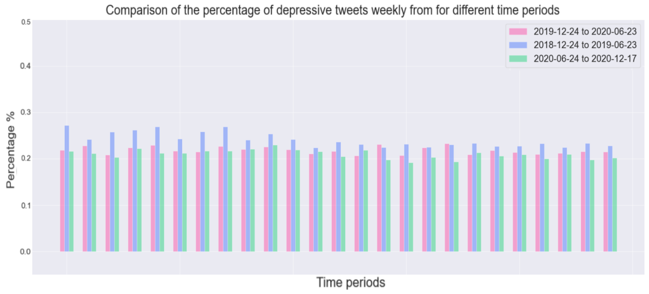

Figure 6: *Depressive tweet percentages for the three investigated time periods, predicted by the LSTM model.*

It was interesting to see that the fancier DistilBERT model did not perform as well as a unidirectional, smaller LSTM. Perhaps the more complicated model was not useful because of the short length of Tweets and the smaller size of the dataset. This model would probably need more training data to improve its accuracy, but that could make it even more slow.

The average level of depressive tweets during the lowest time period is about 21.8% which is much higher than we expected.

Further Work

As evidenced by previous research, there are multiple ways to collect and clean data which could impact the results. Since data collection is majorly driven by keywords based search, it is possible that there are a lot of false positive tweets getting marked as depressive in the initial dataset. Additional cleaning steps such as looking for specific hashtags, mental health related profiles (of organizations, health services etc.) could be incorporated to narrow down the results and get a higher quality of data.

There are some things that could be causing us to miss the true signal. The keywords that we search for are non-exhaustive; there are more words we could use, as well as much more data for each word. Another thing that could obscure our results is that when we search for 5,000 tweets per week, that is only a small subset of the tweets available. Therefore, random chance could change whether we get depressive tweets or not. This is a weakness of our method. Twitter is a very dense and informal format of language, where people shorten their words, misspell, and use different forms of slang frequently. It is easy to imagine that the network might have trouble picking up all of the meaning packed into one tweet. In some cases it could also be important to understand the context of the tweet as it could be sarcastic or satirical which may result in false positives.

It could also be interesting to see the level/intensity of depression using a regression based model. Predicting depression could be extended to observe patterns in sentiment of the tweet, timing, frequency between tweets etc. (Kumar, Sharma, and Arora 2019) and develop a score that can be used to diagnose early onset of depression.

As the pandemic is still ongoing, once it is over, it could be interesting to explore the long terms effects on mental wellness of individuals and locations at a larger scale and compare against available infrastructure to potentially develop recommendations to improve mental health services.

References

C. Zacharias, F. P. *TWINT - Twitter Intelligence Tool.* Viewed: 2020-12-07.

Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What does BERT look at? An analysis of BERT's attention. *arXiv*.

Fact, F. 2020. When did lockdown begin in the uk?

Kumar, A.; Sharma, A.; and Arora, A. 2019. Anxious depression prediction in real-time social data. *SSRN Electronic Journal*.

Murthy, D.; Allu, S.; Andhavarapu, B.; and Bagadi, M. 2020. Text based sentiment analysis using lstm. *International Journal of Engineering Research and* V9.

Patel, A., and Tiwari, A. K. 2019. Sentiment analysis by using recurrent neural network. *SSRN Electronic Journal*.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* 2–6.

Sherstinsky, A. 2020. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena* 404:132306.

Singh, D., and Wang, A. Detecting depression through tweets. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6879557.pdf.

Torch Contributors. 2019. Lstm. https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html. Viewed: 2021-01-15.

Wire), I. S. H.; Chaney/Collins), I. G.; and REUTERS), I. 2020. When did lockdown in ireland start? timeline of coronavirus restrictions.

Appendix

A. Larger version of images

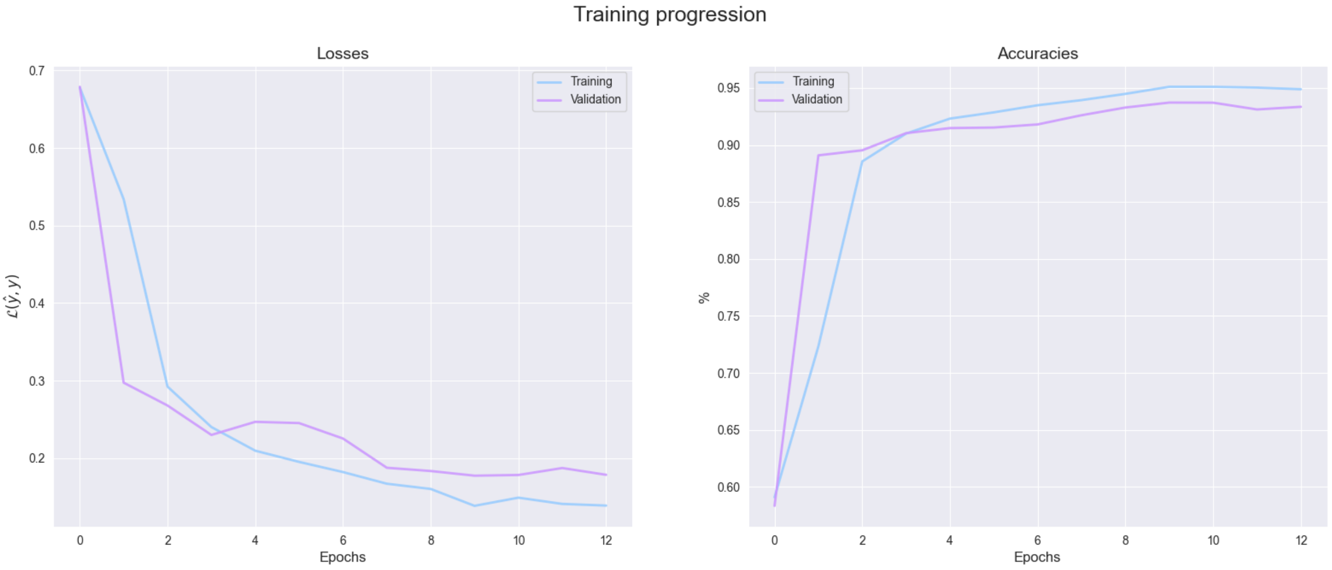

Figure 7: *Training progression of the LSTM model with the losses of the training and validation (**left**) and and the corresponding accuracies (**right**).*

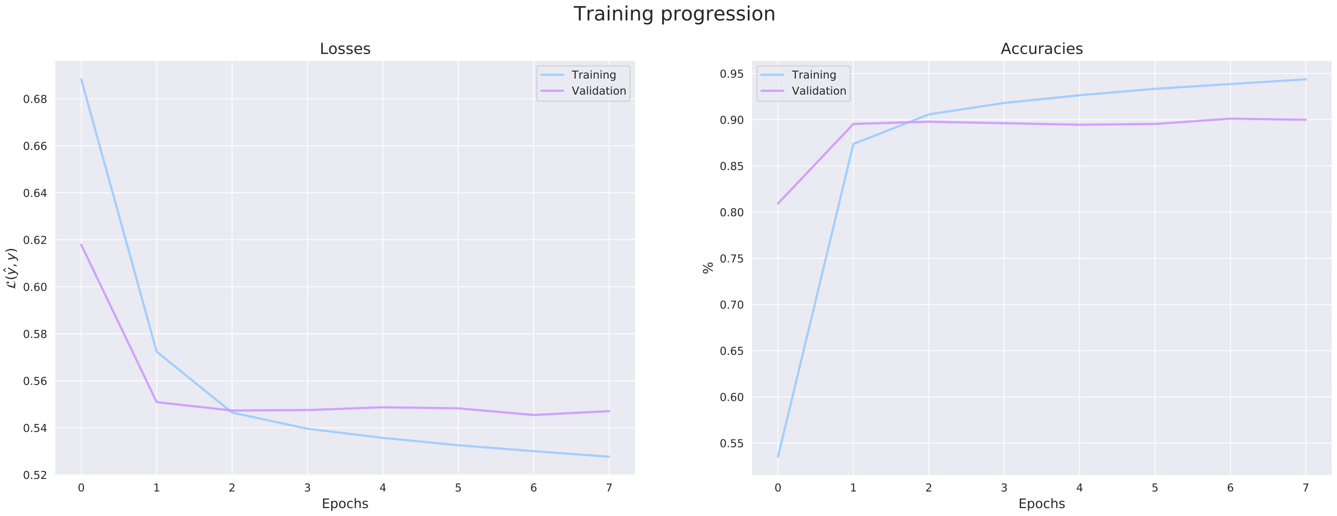

Figure 8: *Training progression of the DistilBERT model with the losses of the training and validation (**left**) and and the corresponding accuracy (**right**). We can see the model becomes over-trained very quickly.*

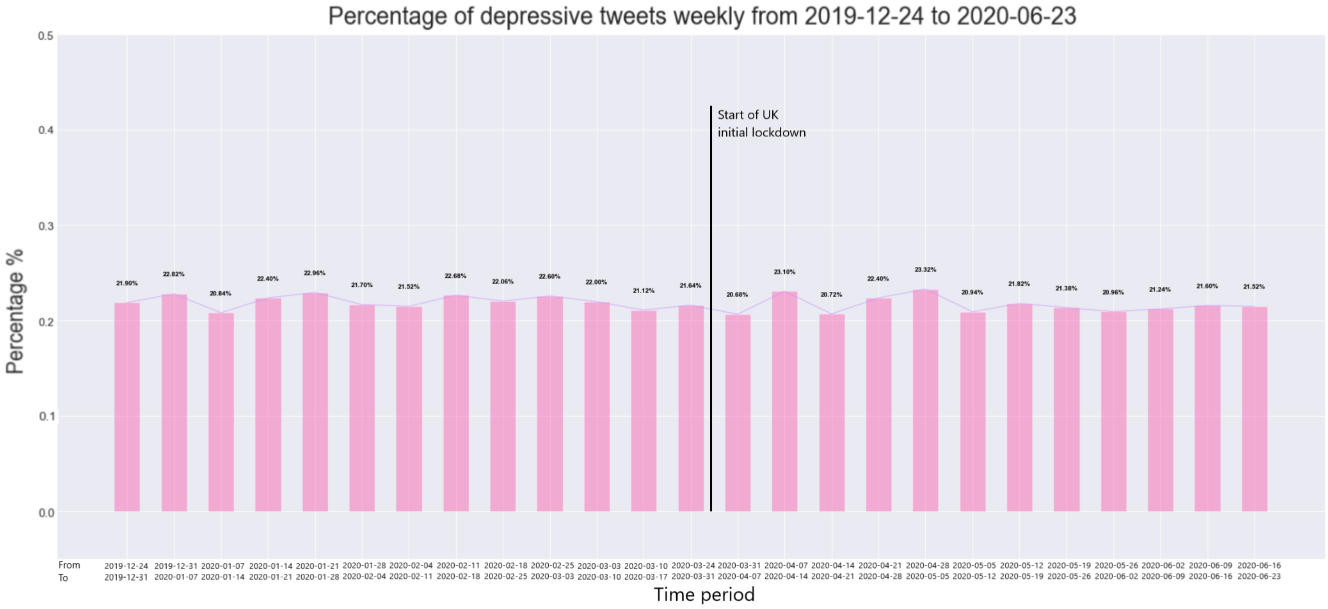

Figure 9: *The percentage of depressive tweets, predicted by the LSTM model, from the 24th of December, 2019, to the 24th of June, 2020. This is the lockdown time period.*

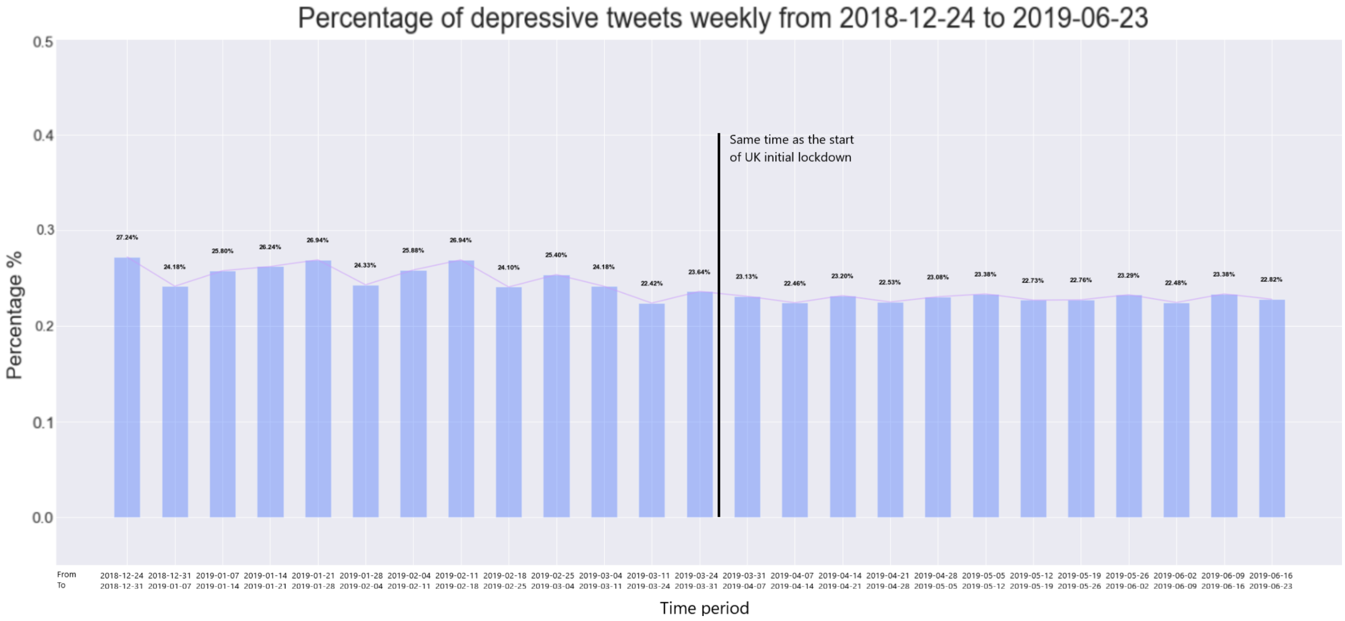

Figure 10: *The percentage of depressive tweets, predicted by the LSTM model, from the 24th of December, 2018, to the 24th of June, 2019. This is the control year.*

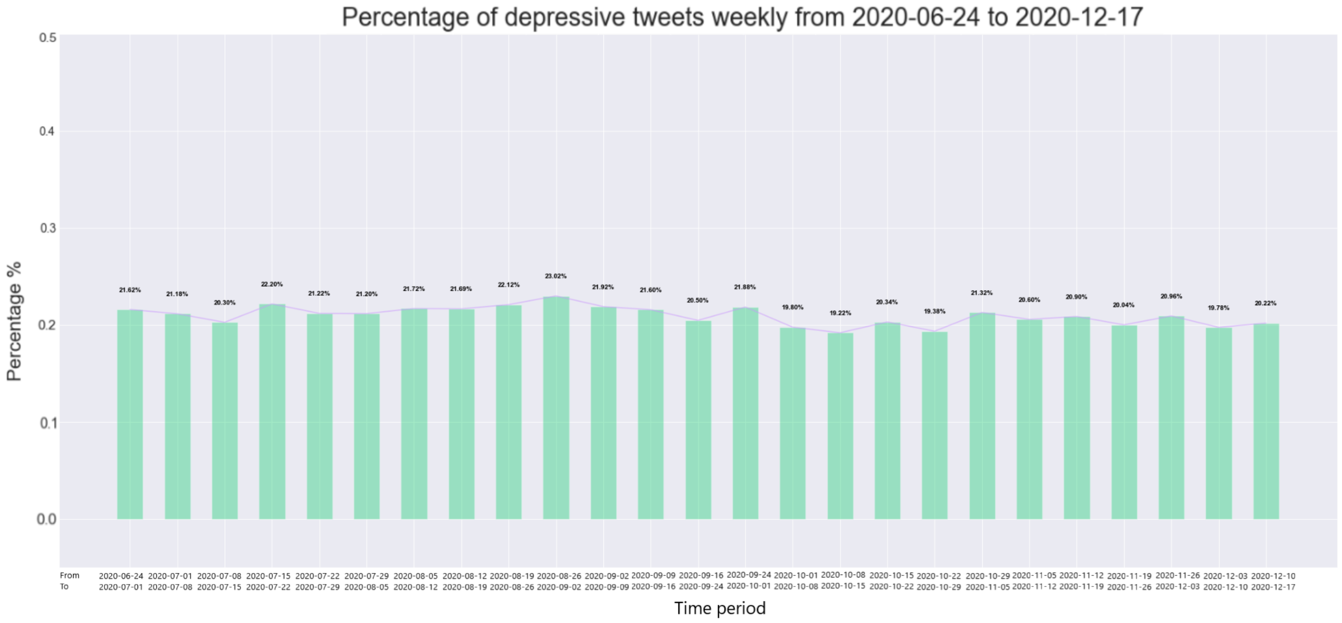

Figure 11: *The percentage of depressive tweets, predicted by the LSTM model, from the 24th of June, 2020, to the 17th of December, 2020.*

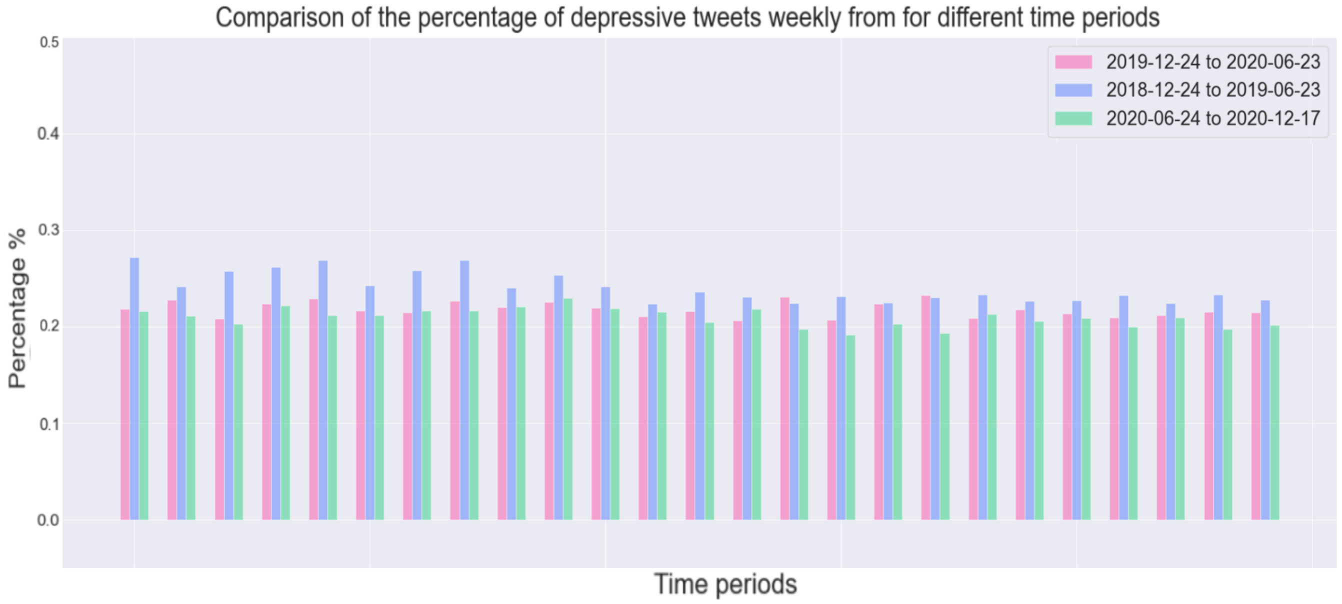

Figure 12: *Combined time periods of the percentage of depressive tweets predicted by the LSTM model combined into one graph.*