Classification and sentiment analysis on COVID-19 tweets regarding fake and real news

Mila Kuch
Faculty of Computer Science and Engineering,
Skopje, North Macedonia
mila.kuch@students.finki.ukim.mk

***Abstract*—With the appearance of the COVID-19 pandemic, the overall life of people worldwide suddenly and drastically changed, causing people to pay more attention to the relevance of the information they receive. In this research paper, the main accent is given on NLP data analysis and deep learning techniques. Using the CoAID Dataset, it includes different examinations and visualizations based on Twitter activity regarding COVID-19 during last year, such as presenting the number of tweets every month, determining the most frequent words, and a comparison of words used in real vs. fake tweets. It additionally includes sentiment analysis using VADER, as well as a deep learning neural network for classification, based on BERT technology and recurrent neural networks.**

Keywords—covid-19, real vs. fake, sentiment analysis, twitter dataset

I. Introduction

In times of a global pandemic, one of the most important things that improve our safety and quality of life - is the news we receive on a daily basis. The era of digital information brought us new ways to inform ourselves, and social media has become one of the main ways to share our own thoughts and opinions, as well as staying updated on everything happening around the world.

But how threatening is the accuracy of the information in our daily lives? One research conducted by MIT University [1] states that it is 70% more likely for a person to share a post containing fake news on the social network Twitter, than ones containing real information, which is highly concerning. Luckily, by analyzing current information from social media, especially text posts, we can discover patterns and recognize how we can stop the fake news from spreading, as well as capturing the overall state of mind of people regarding a certain topic.

The goal of this project is to combine those two aspects with Natural Language Processing (NLP) - to present evidence that will help people differentiate real from fake news, but also to capture the analysis of the general emotional response regarding a highly discussed topic nowadays – which is COVID-19.

In this paper, firstly, there will be a demonstration of detailed analysis of the data, as it’s very important to examine given data first, to find patterns, deductions, and the answers for some questions. Then, it continues with a deep learning model for classification of tweets using several NLP techniques and a BERT model combined with a reccurent neural network architecture.

II. Related works

According to the research of Limeng Cui and Dongwon from Pennsylvania State University [2], who published the CoAID dataset (further used in this project), with the rise of the global pandemic, the amount of fake information regarding the topic has also risen. One example is fake cures for COVID-19; an Arizona man died, and his wife was hospitalized after taking Chloroquine as a prevention against the pandemic, due to disinformation [3]. Another example is the 77 cell phone towers that were set on fire because of the conspiracy theory that 5G cell phone networks are responsible for spreading the virus. [4]

On the second topic, another research conducted by the University of South Florida, Tampa, Northern Illinois University, and Ajivar LLC [5], which focused more on the sentiment analysis and classification of sentences according to the most popular topics, shows that after the pandemic occurred, students, in general, were most worried about topics like health and family, and more often than not, mentioned those in a negative connotation, which shows the general emotional response of people during the pandemic.

Lately, there has been a lot more research regarding the recognition of real and fake news in social media [6][7][8], all of those utilizing natural language processing methods, in order to draw significant deductions.

III. Describing the problem

To find the ideal dataset, we always need to keep a certain goal in mind, specifically, what kind of analysis we want to perform. Since the topic is related to fake and real information on social media, one, in particular, has stood out to be the perfect choice to build this research upon. Twitter is a platform that is fully open for everyone, has an easily accessible API, and is currently one of the most popular and widely used social media platforms.

To obtain a more structured approach, it is always a good idea to draw out specific questions or tasks that need to be answered. In this case, there are several, for example:

1) What was the activity of Twitter users on the topic COVID-19 like in different months? Is the interest in the topic growing stronger or weaker within time?

2) What are the most frequent words occurring in tweets every month?

3) What are the most frequent words in fake tweets vs.. real tweets?

4) Performing sentiment analysis of the whole dataset

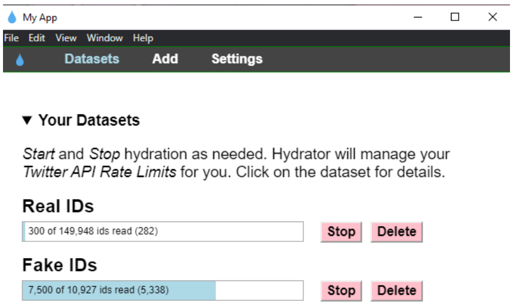
5) Creating a deep learning model for classification of fake vs real tweets.

By completing these tasks and answering these questions, we can get a solid understanding of the full picture of our dataset, which can lead us to further conclusions about the topic and further improvements.

IV. Dataset

The dataset CoAID [2] consists of 5216 news, 296752 user activities, and 958 tweets on the social media platform Twitter about COVID-19. The dataset consists of several subsets, more specifically data that is collected every other month (May, July, September, November). Furthermore, the data is divided by tweets and external data sources, provided with titles and links. This research employs only the data from tweets, and not external sources since this research focuses on the analysis of social media posts. The dataset is available on GitHub and everyone can download it from there for research purposes.

In the dataset, every tweet is identifiable by its unique ID. In order to gather the text of the tweet and other metadata, several tools can be used, one of them being Hydrator [9], a desktop application based on Electron. With Hydrator, everyone can access the data of each tweet based on its ID, and the result is brought in JSON format. The course of action is shown in Fig 1. Some additional transformations of the data were made to collect them all in one place, getting them ready to be analyzed.



▼ Your Datasets

Start and *Stop* hydration as needed. Hydrator will manage your *Twitter API Rate Limits* for you. Click on the dataset for details.

Real IDs

300 of 149,948 ids read (282)

Fake IDs

7,500 of 10,927 ids read (5,338)

Fig. 1. Gathering data using Hydrator.

V. Data analysis

For data manipulation, there are a lot of libraries that can be useful for different data science tasks; NLTK[10] is one of the most popular and successful in the field of Natural Language Processing, which has powerful functions for tokenization, embeddings, etc. This project also utilizes Pandas and Numpy as standard libraries for data science.

In this project, the data is filtered, so it only contains English tweets, even though the original dataset has tweets in several languages across the world. In the following paragraphs, the main focus is the questions and tasks mentioned in the previous part of this research.

A. *Number of tweets according to month*

The first issue is regarding the activity of Twitter users on the topic COVID-19, according to different months. The results are important to gain perspective for what the dataset is like and the interest in the topic of the general population over time.

Something that needs to be noted right away is the fact that in the latter months, there is a significant drop in the number of tweets. This doesn't mean that it's the reality of the situation, but rather such is the data set that this project uses. The results can be seen in Fig 2. In our dataset, the month with the highest number of tweets is April, namely 37651 tweets.

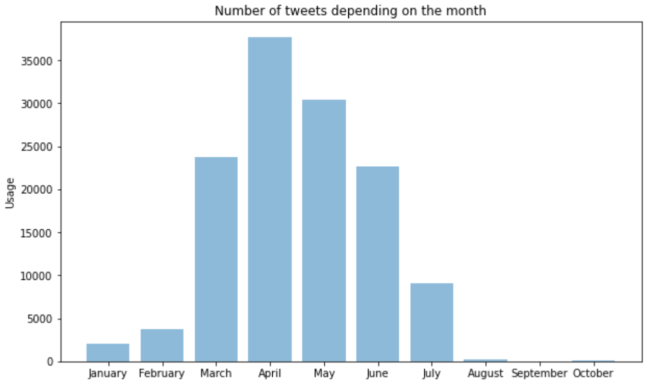


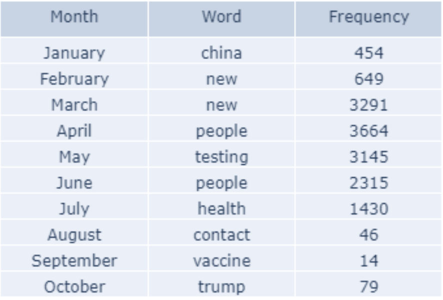
Fig. 2. Number of tweets in the dataset in given months.

B. *Most frequent words every month*

This task is necessary so we can understand the most important topic of discussion each month. In this analysis, a general vocabulary has been made for each month, out of the words used in both the fake and real tweets. Additionally, it needs to be sorted by frequency, so that the word with the maximum frequency is more apparent.

A crucial step of this process is word tokenization, and a lot of NLP libraries include such functions, one of them being NLTK Word Tokenize [10]. Additionally, to avoid repetition, the words COVID-19, COVID, and other common words were not part of the vocabulary.

The results can be seen in Fig. 3, and out of all the months, the word with the highest frequency is the word "people" in April, which appears 3664 times. It's also interesting how at the beginning of the year, the relevant topics were "China", "new", "testing”, and later on the general public was more worried about topics containing "vaccine" and "Trump".



Month	Word	Frequency
January	china	454
February	new	649
March	new	3291
April	people	3664
May	testing	3145
June	people	2315
July	health	1430
August	contact	46
September	vaccine	14
October	trump	79

Fig. 3. Table containing the most used words every month in both fake and real tweets, along with their frequency.

C. *The most prominent words in real vs.fake tweets*

Perhaps one of the most beneficial analyses is finding out the most used words in the real and fake tweets, because of its practical nature, so we can see how we can avoid fake news throughout social media.

The results are certainly interesting - in the fake tweets the presence of more political themes is more dominant (e.g. “bill”, “gates”, “trump”, “bioweapon”), and in the real ones the focus is more on the health and safety of people (e.g. “health”,” cases”, “home”, “testing”). The results can be seen in Fig. 4 and Fig.5.

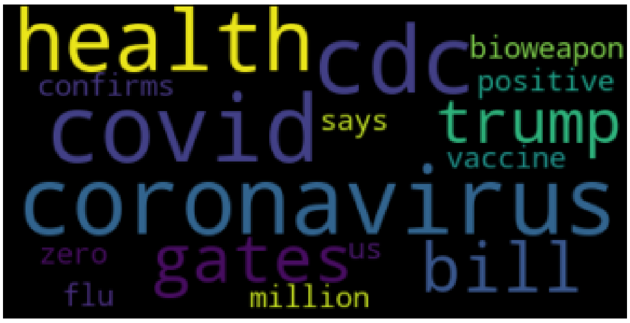


Fig. 4. Most used words in fake COVID-19 Tweets.

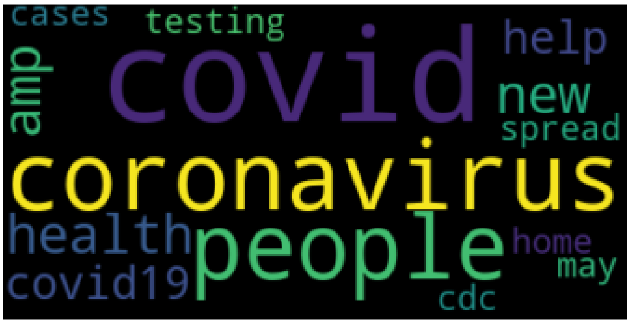


Fig. 5. Most used words in real COVID-19 Tweets.

These visualizations were generated using the library word cloud[11], which allows interesting and noticeable presentation of short textual data. In order to get that result, two separate vocabularies have been formed - one containing words from fake tweets, the other one from real tweets. After sorting each vocabulary by frequency, the top 20 words of each one are presented on the visualizations.

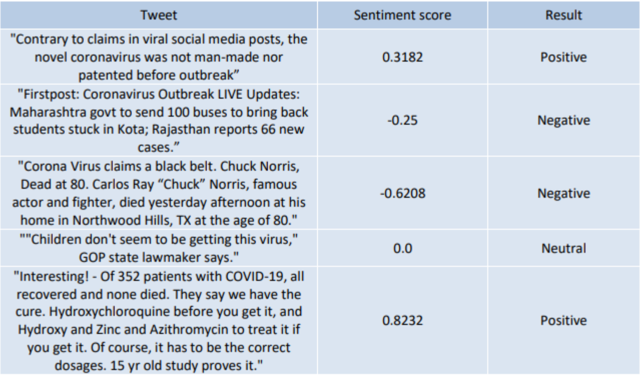
VI. Sentiment Analysis

Sentiment analysis is a common NLP task that involves classification of parts of the texts with predefined sentiments. The sentiment models of analysis usually focus on the polarity (e.g. positive, negative, neutral), but they can also refer to the emotions themselves (angry, happy, sad), etc. [12]

There are several approaches and means for sentiment analysis; one of the most prominent ones being VADER (Valence Aware Dictionary and sEntiment Reasoner)[13]. This library is based on a lexicon, and rules for sentiment analysis, being specifically useful for sentiments on social media. The library is available on GitHub. This model is sensitive to polarity (positive, negative), and to the intensity of the feeling. The final result is a sum of the intensity of every word of the text.

For example, the tweet "Contrary to claims in viral social media posts, the novel coronavirus was not man-made nor patented before outbreak" has a sentiment result of 0.3182 and is labeled as positive.

An opposite example is the tweet "Firstpost: Coronavirus Outbreak LIVE Updates: Maharashtra govt to send 100 buses to bring back students stuck in Kota; Rajasthan reports 66 new cases.” ,labeled as negative, with a sentiment result of -0.25. More examples can be seen on Fig.6.



Tweet	Sentiment score	Result
"Contrary to claims in viral social media posts, the novel coronavirus was not man-made nor patented before outbreak"	0.3182	Positive
"Firstpost: Coronavirus Outbreak LIVE Updates: Maharashtra govt to send 100 buses to bring back students stuck in Kota; Rajasthan reports 66 new cases.”	-0.25	Negative
"Corona Virus claims a black belt. Chuck Norris, Dead at 80. Carlos Ray “Chuck” Norris, famous actor and fighter, died yesterday afternoon at his home in Northwood Hills, TX at the age of 80."	-0.6208	Negative
""Children don't seem to be getting this virus," GOP state lawmaker says."	0.0	Neutral
"Interesting! - Of 352 patients with COVID-19, all recovered and none died. They say we have the cure. Hydroxychloroquine before you get it, and Hydroxy and Zinc and Azithromycin to treat it if you get it. Of course, it has to be the correct dosages. 15 yr old study proves it."	0.8232	Positive

Fig. 6. Exmples of more sentiment scores and sentiment results

The results of the sentiment analysis of the tweets are shown in Fig. 7, divided by fake vs. real tweets. One can clearly see that in the fake news section, there are drastically more tweets with a negative connotation, as fake news can often be threatening, intimidating, and aggressive sounding. On the contrary, the tweets containing real news have a more positive and neutral tone to them.

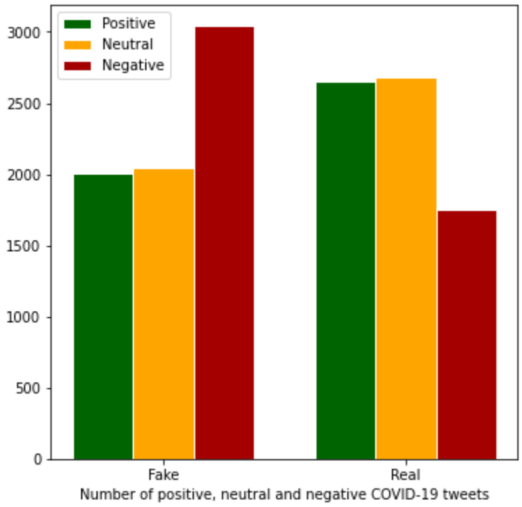


Fig. 7. Result of sentiment analysis of the CoAID dataset.

VII. Deep Learning Model

The goal of deep learning is investigate how computers and machines can use data to develop representation of complex interpretation tasks[14]

In the field of Natural Language Processing, there are several approaches when it comes to deep learning. A more recent trend are word embeddings, as well as building deep learning models based on recurrent neural networks. In the next part, both approaches will be combined for getting the best results.

As input in the neural network, the model uses the body of the tweets (the column "full_text"), represented by word vectors. The output of the neural network is defined by the labels of the classes (0 or 1 accordingly, if the tweet contains fake vs. if the tweet contains real news). The training, testing, and validation set is divided into an 80:10:10 ratio.

A. *BERT Embeddings*

In 2018 the Google AI team made a revolutionary change in Natural Language Processing by introducing BERT

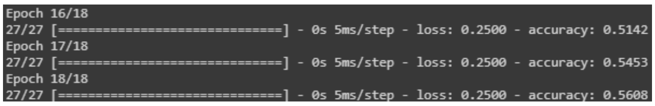
(Bidirectional Encoder Representations from Transformers)[15]. Because of its highly pragmatic approach and better performance, BERT is one of the most contemporary state-of-the-art tools for many tasks in this field. [16]

In this research, the library Sentence Transformers[17] is utilized to build feature vectors in form of word embeddings. This tool has simple methods for preparing appropriate vector presentations, on a sentence or paragraph level (also called sentence embeddings). The models based on this network are BERT / RoBERTa [18] / XLM-RoBERTa [19] and they function in a way so that the sentences with similar meanings are closer to each other in the vector space, and those with different meanings further apart from each other.

The model used in this research is 'paraphrase-distilroberta-base-v1'. This version contains a separate tokenizer, and as a result, each sentence is transformed as a vector with size 768, making the vectors ready to be used as an input for the model.

B. RNN Neural Network with LSTM layer

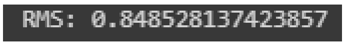
The model of choice for this research is a Keras deep learning reccurent model with an LSTM layer.[20] It consists of an input layer, LSTM layer, two hidden layers with 64 and 240 neurons respectively, and output layer, which represents the class of the tweet. By training the data in 18 epochs, with batch size 30, the results show accuracy of 0.56 in the last epoch, as shown in Fig. 8. The validation accuracy of the last epoch is 0.78, which means the model is less likely to be overfitted.



```
Epoch 16/18
27/27 [==============================] - 0s 5ms/step - loss: 0.2500 - accuracy: 0.5142
Epoch 17/18
27/27 [==============================] - 0s 5ms/step - loss: 0.2500 - accuracy: 0.5453
Epoch 18/18
27/27 [==============================] - 0s 5ms/step - loss: 0.2500 - accuracy: 0.5608
```

Fig. 8. Precision in the last epochs of training

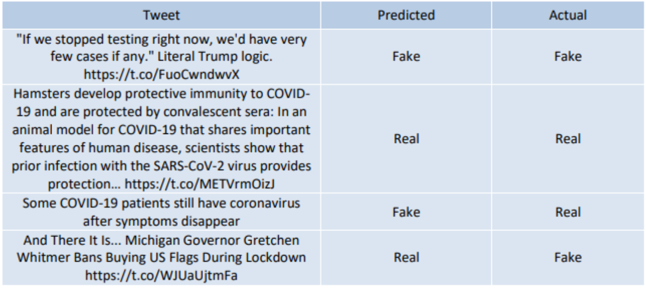
Additionally, for the predictions, RMS can be calculated as a measure for precision. For example, one prediction shows RMS of 0.848, but this number can vary depending of the data that is used for testing (Fig. 9).



```
RMS: 0.848528137423857
```

Fig. 9. RMS – Measure for precision

The model can predict the class of a tweet, specifically if it's real or fake. This model has a lot more room for improvement, but these results are satisfactory, too. Examples of the predictions can be seen on Fig.10



Tweet	Predicted	Actual
"If we stopped testing right now, we'd have very few cases if any." Literal Trump logic. https://t.co/FuoCwndwvX	Fake	Fake
Hamsters develop protective immunity to COVID-19 and are protected by convalescent sera: In an animal model for COVID-19 that shares important features of human disease, scientists show that prior infection with the SARS-CoV-2 virus provides protection… https://t.co/METVrmOizJ	Real	Real
Some COVID-19 patients still have coronavirus after symptoms disappear	Fake	Real
And There It Is… Michigan Governor Gretchen Whitmer Bans Buying US Flags During Lockdown https://t.co/WJUaUjtmFa	Real	Fake

Fig. 10. Examples of the predictions of the model

From the general results of the model, it's interesting to note that in the incorrectly predicted instances, most of them were false negative tweets - real news labeled as fake. The conclusions from the results are interesting – firstly, the model tends to label longer tweets as real, and shorter tweets as fake, which can make sense in the real world too, as real news tend to contain more detailed information. Additionally, the tweets that contain tags and replies often get labeled as fake as well. The tweets that contain more negative news also tend to get labeled as fake, due to the high amount of negativity in them, as mentioned in the sentiment analysis part of the research paper. An example would be the statement "Comparing COVID-19, Flu Death Tolls 'Extremely Dangerous' https://t.co/GGTnxSl3vB via @medscape" which is shorter in length, contains a tag, and also carries bad news, so the model labels it as fake even though it's real by nature.

VIII. Conclusion

The models and approaches for analysis of fake and real data have room for growth, like everything else in the world of Natural Language Processing. Sometimes it is interesting to look into the world of data and how it incorporates in the real world, like real information, and to draw out important conclusions from them, so we get a better future for ourselves.

References

[1] Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science 2018;359(6380):1146– 51.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Limeng Cui, Dongwon Lee: CoAID: COVID-19 Healthcare Misinformation Dataset. CoRR abs/2006.00885 (2020) Neuman, Scott. "Man Dies, Woman Hospitalized After Taking Form Of Chloroquine To Prevent COVID-19", *National Public Radio*,24 Mar. npr.org/sections/coronavirus-live-updates/2020/03/24/820512107/man-dies-woman-hospitalized-after-taking-form-of-chloroquine-to-prevent-covid-19. Accessed 24 March 2021.

[3] Hamilton Asher Isobel, "77 cell phone towers have been set on fire so far due to a weird coronavirus 5G conspiracy theory",*Business Insider Inc*., 6 May 2020, businessinsider.com/77-phone-masts-fire-coronavirus-5g-conspiracy-theory-2020-5. Accessed 24 March 2021.

[4] Ravi Sharma, Sri Divya Pagadala, Pratool Bharti, Sriram Chellappan, Trine Schmidt, Raj Goyal: Assessing COVID-19 Impacts on College Students via Automated Processing of Free-form Text. HEALTHINF 2021: 459-466

[5] N. X. Nyow and H. N. Chua, "Detecting Fake News with Tweets' Properties," 2019 IEEE Conference on Application, Information and Network Security (AINS), Pulau Pinang, Malaysia, 2019, pp. 24-29, doi: 10.1109/AINS47559.2019.8968706.

[6] Muvazima Mansoor, Kirthika Gurumurthy, Anantharam R. U, V. R. Badri Prasad: Global Sentiment Analysis Of COVID-19 Tweets Over Time. CoRR abs/2010.14234 (2020)

[7] Md Yasin Kabir, Sanjay Madria: CoronaVis: A Real-time COVID-19 Tweets Analyzer. CoRR abs/2004.13932 (2020)

[8] LamsalRabindra, "Design and analysis of a large-scale COVID-19 tweets dataset", doi: 10.1007/s10489-020-02029-z

[9] S. Bird, NLTK: the natural language toolkit, in Proceedings ofthe 21st International Conference on Computational Linguistics(ACL 2016), Sydney, Australia, 2006.

[10] MuellerAndreas, "WordCloud for Python documentation", *GitHub, Inc*., 2020, amueller.github.io/word_cloud. Accessed 24 March 2021.

[11] "Sentiment Analysis: A Definitive Guide", *MonkeyLearn*, 2020, monkeylearn.com/sentiment-analysis.Accessed 24 March 2021.

[12] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[13] Richard Socher, Yoshua Bengio, and Christopher D. Manning. 2012. Deep learning for NLP (without magic). In Tutorial Abstracts of ACL 2012 (ACL '12). Association for Computational Linguistics, USA, 5

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186

[15] RajapakshaIsanka, “BERT Word Embeddings Deep Dive”, *A Medium Corporation*, 11 October 2020, medium.com/analytics-vidhya/bert-word-embeddings-deep-dive-32f6214f02bf. Accessed 24 March 2021.

[16] Nils Reimers, Iryna Gurevych: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. CoRR abs/1908.10084 (2019)

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov: RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692 (2019)

[18] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov: Unsupervised Cross-lingual Representation Learning at Scale. ACL 2020: 8440-8451

[19] Chollet, Francois and others, “Keras”, 2015, *Github, Inc.*, github.com/fchollet/keras. Accessed 24 March 2021.