1st International Conference and 6th National Conference
on Computers, information technology and applications of artificial intelligence

February 22th, 2023

Generating and Summarizing worldwide Covid-19 reports using Abstractive and Extractive NLP

Arman Feili
Student
Islamic Azad University of Shiraz
Shiraz, Iran
email: armanfeili7@gmail.com

Amin Feili
General Practitioner
Shiraz University of Medical Sciences
Shiraz, Iran
email: aminf.master@gmail.com

Iman Badrooh
Lecturer
Islamic Azad University of Shiraz
Shiraz, Iran
email: iman.badrooh@iau.ac.ir

***Abstract*— Since the beginning of the Covid-19 outbreak, each country has experienced substantial changes regarding the burden of the disease. Accordingly, the volume of Covid-19 data has expanded in size and variety.[1] Authorities, news services, and the public find it too challenging to stay aware of new happenings about the disease. It is, therefore, vital to help the public and policymakers catch up with the latest news about the pandemic using an easy-to-understand report. This project aims to produce an accessible application in Python that can retrieve Covid-19 data, generate a well-structured text report, and create two reviews based on the Extractive and Abstractive NLP (Natural Language Processing) summarization approaches without any grammatical errors.**

Keywords*— *Covid-19, NLP , data2text, Abstractive Text Summarization, Extractive Text Summarization, grammar-check

I. Introduction

Since the beginning of the Covid-19 pandemic, each country has undergone dramatic changes regarding the burden of the disease. Several subtypes have emerged, and each nation has experienced multiple lockdowns. People, news services, and healthcare decision makers desperately need to stay up-to-date on covid-19 alterations in different regions. However, the volume of covid-19 data has become too big in size and variety to handle.[2] Consequently, analyzing and interpreting such massive data is both time and energy-consuming, which has led the general public to neglect long reading reports and navigate the vital parts every day. Therefore, people do not stay aware of the new subtypes and cannot get a sense of severe conditions, which makes them reluctant to respect social distancing and other health policies. In addition, the availability of a summarized report about the specific statistics and condition of the outbreak in any time section since the epidemic's beginning can help researchers recognize historical trends and assist policymakers in seeing the consequences of local or national decisions at any time period. It is, therefore, vital to help the public and authorities catch up with the latest news about the pandemic using an easy-to-understand report.

In order to create a simple, daily report on the status of Covid-19 in each region, we need a digital platform that can fetch data from valid resources and transforms them into easy-to-understand text reports. Then the report should be summarized by NLP (Natural Language Processing), a branch of Artificial Intelligence and Linguistics devoted to making computers understand the statements or words written in human languages.[3] Summarization is implemented using extractive or abstractive techniques. In the Extractive process, the critical phrases of sentences from the actual content are combined to form a summary.[4] The abstractive procedure, on the contrary, focuses on paraphrasing the corpus, adds novel words or phrases into the summary if necessary, and simultaneously keeps the original meaning alive.[5]

Many obstacles arise during the data-to-document conversion process, and various factors must be considered. The first is clarity, i.e., readers should appropriately understand the underlying data in detail. If data is gathered from multiple sources, it may lead to a lack of coherence. In addition, using manual data entry while there are too many items to cover increases the chance of text fragmentation and repetitive descriptions in the generated text.[6] This project provides the facility of generating a quick and meaningful summary to make it possible for all readers to get notified about their country's Covid-19 condition just by taking a glance at a single paragraph.[7]

The actual code for this project is accessible at: https://github.com/armanfeili/covid-19-nlp-summarization

II. Related Work

A. Previous Efforts

Previously, researchers have covered many aspects of Covid-19 studies. Taking the COVIDSum research into account, scientists focused almost exclusively on abstractive techniques in datasets like the COVID-19 Open Research Dataset, including thousands of scientific articles, to achieve a summarized version of the Abstract part of an article.[8], [9] They used the mentioned dataset to train their SciBERT-based model for abstractive summarization. Other NLP papers covered further aspects by training models to summarize a massive amount of Covid-19 news.[10] Another article that worked on summarizing Covid-19 medical records used the Extractive technique on existing Covid-19 text datasets.[11]

B. How Our Proposed Article Is Different

Since multiple publications tried gathering Covid-19-related big data, training their models, and summarizing all articles into one paper or paragraph using NLP techniques, the main focus of our proposed scientific paper was to acquire a different approach for summarization.[12] As a result, we generated our raw text using a straightforward data2text approach and performed the first attempt at text summarization for a unique report explaining the Covid-19 statistics. Accordingly, our project is not quantitatively comparable to previous researches on existing datasets since

we neither proposed any new model nor used the same dataset for our summarization.

Nonetheless, we used one of the most popular pre-trained models for the Abstractive summarization named 'Facebook/ Bart-Large-CNN' as the final summary result was closer to the context of the first raw text.[13] The quantitative evaluation is available in the Abstractive Summarization section.

III. Methods

A. Report Structure

Initially, we scrutinized some of the official reports of WHO[1], NHS[2], and CDC[3], along with the news articles about the Covid-19 pandemic in European and American magazines. We only searched for texts in the English language. We used the following keywords for searching in the google engine: Covid-19 daily cases, Covid-19 news, and the Covid-19 statistics report. We searched for the most frequent information repeatedly reported in the news and articles about the burden of the disease in different countries since the beginning of the pandemic. Essential epidemiological variates which appeared to be more focused in daily news reports were found as follows: (1) The frequency of the disease by region among different age groups and sexes (2) The incidence of new cases (3) Vaccination rates (4) Hospitalized patients' rates (5) Governmental policies on distinct situations affected by the outbreaks.[14], [15]

Furthermore, we searched for well-written sentence structures in news articles that can vividly deliver the message and involve humanized semantic themes.

B. Data Gathering

For our generated report to be comprehensive and trustworthy, we created selection criteria for datasets: (1) Being validated by international health authorities, (2) Updating daily, (3) encompassing nationwide epidemiological covid-19 statistics. As a result, we chose our datasets from Oxford World Data and Google Covid-19 Open Data.[1], [2] We wrote the code in Python, used its Pandas library to read 7 CSV files, and converted them into data frames. We incorporated the following datasets:

- Oxford's main Covid-19 statistics (OWID)

The following datasets are gathered from google open data:

- Covid-19 cases among different sexes and ages
- Demographics for all countries
- Hospitalizations
- Health data set
- Governmental response dataset

These datasets facilitate making queries and fetching desired reports based on two arguments, i.e., country code and date. For instance, one can enter 'it' or 'ita', as the alpha-2 or alpha-3 country code for Italy and '2022-01-09' as the date or leave the field of 'date' blank to get the most recent report for that country. (Fig. 1)

We found that worldwide statistics need to be added to our datasets to compare the regional data with the global status in the reports to give a deeper perception of information. The regional data comparable to the global data in our report include infection rates, case fatality rate (CFR), new case rates, new death rates, and vaccination rates.[1]

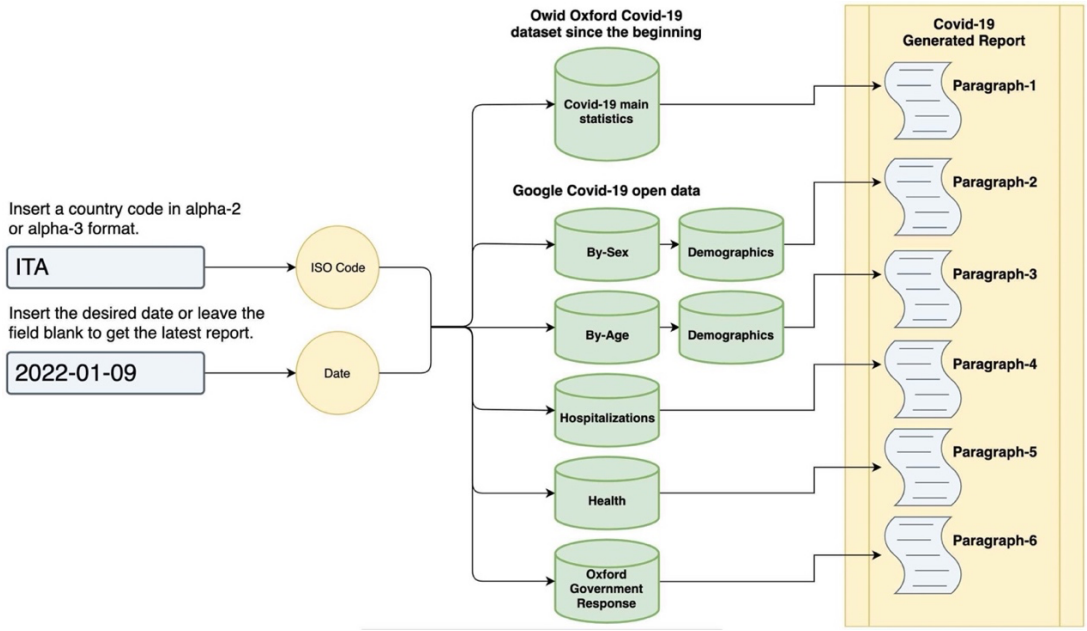

Fig. 1. Generating Covid-19 report.

[1] World Health Organization
[2] the National Health Service (UK)
[3] Centers for Disease Control and Prevention(USA)

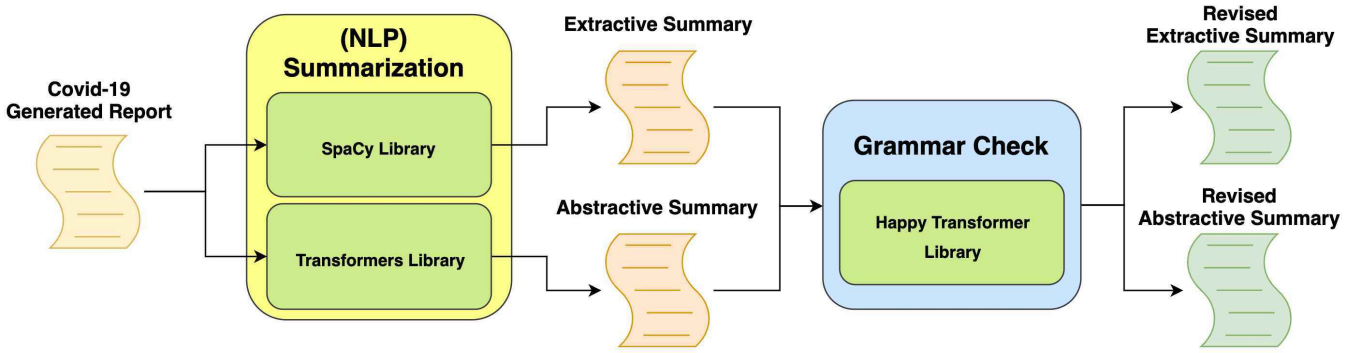

Fig. 2. The process of generating extractive and abstractive summaries and creating revised versions by checking grammar.

C. NLP Process

Among all accessible NLP frameworks in Python, we chose SpaCy and Transformers as the most advanced NLP open-source pre-trained libraries known for their rapid parsing and understanding of large volumes of text. We used these libraries for extractive and abstractive summarization and checking the grammar. (Fig. 2) [16]

D. Extractive Summarization

Extractive summarization is a method that delivers a summary by extracting sentences, showing the most substantial or relevant information within the original content. The information about the content is protected via this approach.[17]

The process starts with tokenizing the text. Tokens are smaller units of a sentence like words, keywords, phrases and symbols. Secondly, we filter all tokens to omit stop-words (e.g., the, a, an, so, what) or punctuation marks. As a result, only the meaningful keywords remain. Thirdly, we create a list of keywords and value them based on their frequency in the text. Besides, we find the most-common keyword in the text as the most valuable token for placing in the denominator of the following equation:

$$V = \frac{f(W)}{f(Wmc)} \quad (1)$$

Where V represents the token's value, which can be achieved by dividing the frequency of a single word f(W) by the frequency of the most-common word f(Wm).[17] Next, we loop over each sentence to determine its value by summing up the values of its tokens. Eventually, we sort the sentences by their final values. We declared a 'limit' argument for the limited number of sentences that return in summary. Hence, if we set 5 as the limit, we will get the first five noteworthy sentences in summary. We used SpaCy, which creates a processing pipeline for applying several functionalities to our text report, including (1) tokenization, (2) assigning part-of-speech tags and dependency labels to tokens, and (3) detecting and labeling named entities. (Fig. 3).[18]

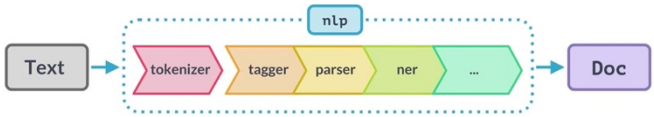

Fig. 3. SpaCy pipeline.

E. Abstractive Summarization

Abstractive summarization means generating an overview in the computer's own words. Instead of selecting sentences from the initial text to create the outline, it paraphrases the main contents of the given text, using a vocabulary set different from the original document. We used Transformers as the most popular Python library for Natural Language Understanding (NLU) and Natural Language Generation (NLG), providing thousands of pre-trained models to perform machine learning tasks. For summarization, we used a transformer encoder (seq2seq) pre-trained model named 'Facebook/ Bart-Large-CNN'. [19], [20], [13] The following quantitative evaluations demonstrate the power of this model among other well-known models based on the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric which is used to evaluate automatic summarization.

Models	*ROUGE-1*	*ROUGE-2*	*ROUGE-L*
Bert	42.79	17.78	35.66
BioBert	43.22	17.94	36.29
Facebook/Bart	**42.94**	**20.81**	**30.61**

Fig. 4. Precision scores of summarization models at the ROUGE metric.

Where ROUGE-1 and ROUGE-2 are the overlaps of unigram (each word or phrase) and bigrams (a pair of words) between the system and the reference summaries, respectively. ROUGE-L, however, is a score based on the length of the longest common subsequence (LCS) that is shared between both the reference text and the model output. As can be seen here, the precision of the Facebook/Bart model is better than the other two models in ROUGE-2 and is better than Bert in ROUGE-1. [21]

First, we divide the original text into four-sentences paragraphs. Then, we summarize each one just by exploiting the mentioned model. Although the software sometimes suggests websites as references to the summarized text, which is interesting, the general meaning of a summary created in this method may be far from the actual report.

F. Checking Grammar

Most of the time, making an automatic summary from a massive report leads to grammatical and semantic errors. One way to resolve such a problem is to check the text's spelling

and grammar using the Happy Transformer package, built on top of Hugging Face's Transformer library. The model used for this purpose is "T5", which generates a revised version of the text to contain fewer grammatical mistakes. It takes several arguments, including the minimum and the maximum number of generated tokens, plus the algorithm's sensitivity. The final result is two extractive and abstractive paragraphs explaining the Covid-19 condition for the desired country and date, without any grammatical errors. (Fig. 3) [22], [23]

IV. Result & Discussion

For easier comparison, we show separate figures for the original Covid-19 report (Fig. 5) vs. the summarized and grammar-checked versions for extractive and abstractive summaries. (Fig 6 and Fig 7, respectively).

> Here is the comprehensive covid-19 report in Italy on date: 2022-08-09:
> Due to the difference in reporting times between states, territories, and the federal government, it can be challenging to get a current picture of the pandemic in Italy. Here we have brought together data on cases, deaths, hospitalizations, and vaccinations. As the latest data suggests, 21368480 total cases have been reported. Unfortunately, Italy has more infection rates than the worldwide infection rates. (36.071 for Italy compared to 7.415 worldwide) The newly reported cases are around 328.
> In comparison to the world data, Italy has a higher new case rate. Regarding the decedents, Italy has lost a total of 173426 residents due to Covid-19 until this date which indicates that Italy has more death rate than the world data. 177 new deaths in the country were reported from national data. Unfortunately, Italy has a higher new death rate than the worldwide data.

Fig. 5. A sample for the generated Covid-19 report.

> (36.071 for Italy compared to 7.415 worldwide) the newly reported cases are around 328. In comparison to the world data, Italy has a higher new case rate. Regarding the decedents, Italy has lost a total of 173426 residents due to covid-19 until this date which indicates that Italy has more death rate than the world data. Here is the comprehensive covid-19 report in Italy on date: 2022-08-09: Due to the difference in reporting times between states, territories, and the federal government, it can be challenging to get a current picture of the pandemic in Italy. Unfortunately, Italy has a higher new death rate than the worldwide data. Unfortunately, Italy has more infection rates than the worldwide infection rates.

Fig. 6. A sample for the revised extractive summary. As can be seen, it selected just the first five significant sentences as we set the limit equal to 5.

> The covid-19 report in Italy on date: 2022-08-09: Due to the difference in reporting times between states, territories, and the federal government, it can be challenging to get a current picture of the pandemic.
> The number of people living in Italy was 36.071 for Italy compared to 7.415 worldwide. The number of adults living in Italy was 6.4 million, compared to 8.2 million worldwide.
> Italy has lost a total of 173426 residents due to Covid-19. Regarding the decedents, Italy has lost more than 17,000 people.

Fig. 7. A sample for the revised abstractive summary.

This project has aimed to obtain two goals. The first was to generate a comprehensive and easy-to-read report (Fig. 4), and the second was to summarize it in a single paragraph using state-of-the-art NLP libraries. (Fig. 5 and Fig. 6) Nevertheless, there were some issues with the final summaries.

A. *Qualitative Evaluation*

The abstractive summary is almost far from the primary Covid-19 text semantically. Sometimes it omits the vital parts of the report and adds unnecessary information. (Fig. 5) The same happens for the extractive summary. It ignores the critical segments of the text while choosing the sentences' significance based on the frequency of the words and creates the summary by skipping the rest of the sentences. (Fig. 6) This is in line with the previous efforts for creating summarized reviews out of data. [17], [19] Generating the automatic and original Covid-19 report was the most reliable part of our results compared to the two summaries. We wrote more than four thousand lines of code, including many 'if-else' statements that indicate which sentences should be added to the report considering the queried numbers or their comparison to the world records. [7] By trying to fetch reports for different countries and dates, it has been revealed that the more the actual text of the report is well structured and similar to real-world news, the more the summaries are accurate and close to the central concept of the report and maintain the importance.

B. *Challenges*

Concerning the limitations of this project, opting for the essential Coronavirus data available for all countries to generate an automatic report was the biggest challenge. Many countries lacked reported datasets. Some data were not comparable to international information. [1] In addition, the algorithms used in the two summarizing approaches still cannot recognize the real significance of some sentences. To make more organized reviews, we suggest recapping each part of the original text separately using an appropriate method. For instance, it is possible to use extractive summarization for the most critical text elements and reserve the abstractive approach for the least important parts. E.g., we might care about new cases of infection by the disease, but at the same time, we care less about the number of tests people have taken. Consequently, we can take advantage of the Extractive approach for summarizing new case data and use the Abstractive technique for summarizing the report related to the number of tests.

Conclusion

Covid-19 data are highly suitable for text generation. This work has attempted to create an automated exclusive Covid-19 report for every country from various datasets and recap it in 2 ways using Abstractive and Extractive Natural Language Processing (NLP) techniques for the first time. Despite this, both summaries had individual issues by ignoring the central part of the report.

Future work should give priority to (1) enhancing the Abstractive and Extractive NLP models, (2) using each summarization approach on separate parts of the text, and (3) summarizing other reports related to Covid-19 to make it faster to inform people.

Acknowledgment

We would like to thank the Oxford and Google Open Data teams for providing such an up-to-date Covid-19 dataset. We further thank the three anonymous reviewers.

REFERENCES

[1] Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina, and Joe Hasell, "Coronavirus disease (COVID-19)–Statistics and research," *Our World in data, 2020 - sipotra.it*, 2020. https://scholar.googleusercontent.com/scholar?q=cache:svhjPXr8nNIJ:scholar.google.com/+covid-19+statistics&hl=en&as_sdt=0,5 (accessed Aug. 26, 2022).

[2] O. Wahltinez *et al.*, "COVID-19 Open-Data a global-scale spatially granular meta-dataset for coronavirus disease," *Sci Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1038/s41597-022-01263-z.

[3] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges," *Multimed Tools Appl*, Jul. 2022, doi: 10.1007/s11042-022-13428-4.

[4] C. Zhu, "Applications and future of machine reading comprehension," in *Machine Reading Comprehension*, Elsevier, 2021, pp. 185–207. doi: 10.1016/B978-0-323-90118-5.00008-4.

[5] R. Paulus, C. Xiong, and R. Socher, "A Deep Reinforced Model for Abstractive Summarization," May 2017.

[6] S. M. Shieber, A. M. Rush, and S. Wiseman, "Challenges in Data-to-Document Generation," 2017. [Online]. Available: http://aclweb.org/anthology/D17-1238

[7] Arman Feili, Amin Feili, and Iman Badrooh, "Generating and Summarizing worldwide Covid-19 reports using Abstractive and Extractive NLP," 2022.

[8] X. Cai *et al.*, "COVIDSum: A linguistically enriched SciBERT-based summarization model for COVID-19 scientific papers," *J Biomed Inform*, vol. 127, p. 103999, Mar. 2022, doi: 10.1016/j.jbi.2022.103999.

[9] Resources on AWS, "COVID-19 Open Research Dataset (CORD-19)." https://registry.opendata.aws/cord-19/ (accessed Dec. 10, 2022).

[10] N. Hayatin, K. M. Ghufron, and G. W. Wicaksono, "Summarization of COVID-19 news documents deep learning-based using transformer architecture," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 19, no. 3, p. 754, Jun. 2021, doi: 10.12928/telkomnika.v19i3.18356.

[11] D. S, L. K. N, and S. S, "Extractive Text Summarization for COVID-19 Medical Records," in *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Nov. 2021, pp. 1–5. doi: 10.1109/i-PACT52855.2021.9697019.

[12] M. Afsharizadeh, H. Ebrahimpour-Komleh, and A. Bagheri, "Automatic Text Summarization of COVID-19 Research Articles Using Recurrent Neural Networks and Coreference Resolution," *Frontiers in Biomedical Technologies*, Feb. 2021, doi: 10.18502/fbt.v7i4.5321.

[13] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," Oct. 2019.

[14] "Covid-19 Australia data tracker: coronavirus cases today, deaths, hospitalisations and vaccination." https://www.theguardian.com/australia-news/datablog/ng-interactive/2022/sep/08/covid-19-daily-cases-australia-today-vaccine-data-tracker-deaths-per-day-hospitalisations-coronavirus-tracking-stats-live-update (accessed Sep. 09, 2022).

[15] A. Hoseinpour Dehkordi, M. Alizadeh, P. Derakhshan, P. Babazadeh, and A. Jahandideh, "Understanding epidemic data and statistics: A case study of COVID-19," *J Med Virol*, vol. 92, no. 7, pp. 868–882, Jul. 2020, doi: 10.1002/jmv.25885.

[16] Bhargav Srinivasa-Desikan, *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*, 6th ed. 2018. Accessed: Aug. 26, 2022. [Online]. Available: https://books.google.de/books?hl=en&lr=&id=48RiDwAAQBAJ&oi=fnd&pg=PP1&dq=spacy&ots=R3A3Lbo3f7&sig=OTYBLVPls2zmCmInlna4UrEKxII&redir_esc=y#v=onepage&q=spacy&f=false

[17] S. JUGRAN, A. KUMAR, B. S. TYAGI, and V. ANAND, "Extractive Automatic Text Summarization using SpaCy in Python & NLP," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Mar. 2021, pp. 582–585. doi: 10.1109/ICACITE51222.2021.9404712.

[18] Honnibal, Matthew and Montani, and Ines, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017. https://spacy.io/ (accessed Aug. 27, 2022).

[19] S. Gupta and S. K. Gupta, "Abstractive summarization: An overview of the state of the art," *Expert Syst Appl*, vol. 121, pp. 49–65, May 2019, doi: 10.1016/j.eswa.2018.12.011.

[20] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[21] C.-Y. Lin, *ROUGE: A Package for Automatic Evaluation of summaries*. 2004. Accessed: Jan. 19, 2023. [Online]. Available: https://aclanthology.org/W04-1013.pdf

[22] Eric Fillion and Ted Brownlow, "Happy Transformer," 2019. https://happytransformer.com/ (accessed Aug. 27, 2022).

[23] T. Wolf *et al.*, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," Oct. 2019.