Newsjam: A Multilingual Summarization Tool for News Articles

Shane Kaszefski-Yaschuk
LORIA
Université de Lorraine
Nancy, France
Email: kaszefsk1u@etu.univ-lorraine.fr

Joseph Keenan
LORIA
Université de Lorraine
Nancy, France
Email: joseph-alexander.keenan@loria.fr

Adelia Khasanova
LORIA
Université de Lorraine
Nancy, France
Email: adelia.khasanova9@etu.univ-lorraine.fr

Maxime Méloux
Bureau d'Économie Théorique et Appliquée
Université de Lorraine
Nancy, France
Email: maxime.meloux4@etu.univ-lorraine.fr

***Abstract*—This paper introduces *Newsjam*, a multilingual summarization tool for COVID-19 news articles. To this purpose, two extractive summarization methods were implemented: Latent Semantic Indexing and K-means clustering on contextual word embeddings on French and English data. This tool was then evaluated using three evaluation metrics and four different corpora; two existing ones as well as two custom-built ones. Finally, the best performing methods were implemented into a complete pipeline, going from text scraping and classification to summarization, and ultimately posting the summaries to Twitter automatically.**

***Keywords*—summarization; twitter; covid-19; news.**

I. Introduction

The ongoing COVID-19 pandemic has caused people from around the world to feel fatigued from the overload of pandemic-related news articles being released every day [1]. One study collecting data from 11 different countries found that more than 26 million coronavirus related articles have been posted since the beginning of the pandemic [2].

These observations have prompted us to create *Newsjam*, a COVID-19 news summarization tool aimed at reducing news fatigue by keeping only the main points of pandemic-related articles and aggregating them in a single place, therefore reducing the amount of time and effort it takes to stay informed about the pandemic.

Newsjam consists of four interconnected parts:

- A web scraper to locate and scrape relevant COVID-19 articles from multiple French and English news websites
- An article classification model to separate these articles into four possible regions of interest
- A summarization model able to generate short summaries for selected articles
- A pipeline automating those steps by regularly fetching new articles, classifying and summarizing them, and then posting the summary as a single tweet to Twitter

In particular, the last part implies that our generated summaries must abide by the maximum tweet length of 280 characters.

The paper is organized as follows. Section II briefly surveys tools that are pertaining to *Newsjam*. Methodology is described in Section III, which details the different corpora used, a brief overview of annotation guidelines, as well as the classification and annotation approaches employed. Section IV concerns the experimental setup and evaluation protocol, and contains our empirical results. In Section VI, we provide a qualitative analysis of the results and conclude with perspectives of future work in Section VII.

II. Literature Review

This section briefly presents text summarization and classification methods and tools encountered in the literature.

A. Text Summarization

There are two main methods of automatic text summarization in the literature [3]. Extractive summarization centers around identifying key sentences in the text and putting them together verbatim, whereas abstractive summarization involves generating novel text. Extractive approaches include assigning importance scores to sentences using topic modeling, k-means clustering, and latent semantic indexing [4]. Primary approaches seen in abstractive text summarization include the use of deep learning, Recurrent Neural Network encoder-decoders, Gated Recurrent Units, and Long Short-Term Memory [5].

Text summarization comes with a few key challenges. During the testing stage, reference summaries are needed for evaluation. However, datasets often contain poor reference summaries or do not contain them at all, making evaluation unreliable or impossible [5]. Other challenges include the occurrence of Out-Of-Vocabulary (OOV) words that are absent from the training dataset but are central to understanding a document, and finding metrics able to evaluate a summarized or paraphrased fragment on both a syntactic and semantic level [6].

Given the multiple challenges encountered in text summarization, the question arises of how to validate the results. The most commonly used metric is Recall-Oriented Understudy for Gisting Evaluation (ROUGE), which compares a generated summary to a reference one (typically created by a human) and calculates the number of overlapping units [7]. ROUGE is however far from ideal: Dorr, Monz, President, Schwartz, and Zajic [8] found that the metric was sensitive to the type of summarization. ROUGE scores also tend to be higher for summaries that are longer or generated by using supervised learning approaches [9].

B. Text Classification

Text classification methods include Decision Trees, Logistic Regression (LR), Naive Bayes (NB) and Support Vector Machine (SVM) [10]. NB is one of the oldest methods of text classification, which is based on Bayes' Theorem and determines the probability that each document belongs to a given class [11]. It can operate using several probability distributions, such as the normal (Gaussian), Bernoulli and multinomial distributions. LR is a simple but effective binary classifier for text classification, which calculates the probability of a document belonging to two different classes. SVM is a supervised learning model that was also originally built as a binary classifier, but was later extended to support multiple classes [11]. Lastly, Decision Trees calculate the probability of different 'children' belonging to the 'parent' of a tree [12].

III. Methodology

This section describes how corpora acquisition and annotation were performed in our pipeline, followed by the chosen implementation of text classification and summarization.

A. Corpora

When discussing the coronavirus pandemic, many words such as *pandemic* and *vaccine* occur much more frequently than in regular news articles. Thus, after careful consideration, it was decided that new corpora should be created in order to provide better accuracy for summarizing novel articles about the pandemic compared to a more general news-based corpus.

For French, articles were scraped from the online version of the newspaper *Actu* and *L'Est Républicain*. Articles were retrieved along with reference summaries which were extracted from the highlights section or title of the article.

For English, articles were similarly extracted from the online version of *The Guardian*. This website was chosen in particular due to the fact that it hosts several versions of *The Guardian*: A USA-based version covering news from the USA and Canada, a UK-based version covering the British Isles, and an Australian one covering both Australia and New Zealand.

In both cases, articles were retrieved on an extended time frame ranging from September 2020 to March 2022.

In addition, two large corpora were selected for evaluation of our models. The French version of the MultiLingual SUMmarization corpus (MLSUM) [13], made up of news articles and summaries from *Le Monde*, was selected for French. For English, the *CNN/Daily Mail* corpus was used [14].

B. Annotation

In order to provide news articles that are relevant to readers, the scope of the tool was limited to 4 distinct geographical regions: France, English-speaking North America, the British Isles, and Australia/New Zealand. To achieve this, a classifier was implemented to determine whether or not a given news article is relevant to a particular region. To this purpose, annotation guidelines for tagging articles as local (relevant) or global (irrelevant) for each region were created. The latter category includes not only news about other countries, but also global events, such as decisions made by the World Health Organization.

Four annotators, labeled A, B, C and D, were selected for the task of annotating our custom-built corpora. For each corpus, two to three annotators tagged the entire dataset. Articles are tagged with "local" or "global" for the *Actu/L'Est Républicain* corpus, and with one of the four possible tags for the *Guardian* corpus (North America, British Isles, Oceania or Global). For each pair of annotators, the Ao, S, Scott's π and Cohen's κ inter-annotator agreement coefficients were then computed. The results are summarized in Table I.

TABLE I
Inter-annotator agreement for each pair of annotators.

Dataset	*Actu*	*Guardian*			*L'Est Républicain*		
Metric	**A-B**	**A-B**	**A-C**	**B-C**	**A-B**	**A-D**	**B-D**
A_o	0.995	0.987	0.917	0.913	0.966	0.976	0.978
S	0.990	0.974	0.834	0.827	0.958	0.952	0.955
π	0.956	0.961	0.761	0.747	0.949	0.879	0.888
κ	0.956	0.961	0.761	0.748	0.949	0.879	0.888

We observe high agreement between all pairs of annotators, for all coefficients and datasets. Furthermore, we observe the highest agreement between annotators A and B, reaching approximately 0.95 in all situations. After careful review of the results, we compute reference tags for our corpora by taking the majority vote between annotators. In case of disagreement, the priority was given to the native speaker of the language in which the article is written.

The resulting datasets, to be used for classification training and testing, proved to be rather imbalanced. The French language dataset, containing articles from *L'Est Républicain* and *Actu*, has more articles about France (58%) than about the rest of the world (42%). The makeup of the English language dataset (Guardian articles) is as follows: 48% about North America, 24% about the British Isles, 16% about Oceania, and 11% about the rest of the world.

C. Article Classification

The pre-processing of the corpus for classification begins with noise removal (punctuation and irrelevant special characters), stopword removal, and lemmatization. Three different methods for classification were implemented: Multinomial Naive Bayes (MNB), Logistic Regression (LR), and Support

Vector Machine (SVM). All three methods were used for both binary and multi-class classification. These methods are inherently probabilistic, so we decided to test several of them to see which one gives the most accurate results in our case. Two classifiers were ultimately created: one for distinguishing between French and non-French articles, and the other one for tagging English articles with the appropriate geographical region as described in Section III-B.

For model-specific issues, one way to improve the performance is hyper-parameter tuning, which can be done by exhaustive search over the parameter space. MNB may benefit from tuning the α parameter, which represents Laplace smoothing, and helps tackle the issue of zero probabilities. For the LR model, we searched for the best *C* parameter. Regarding the SVM model, we tuned the *C* parameter as well to find a balance between the minimum margin and accounting for outliers in the data.

D. Summarization

Two extraction approaches for summarization were chosen: Latent Semantic Indexing and K-means clustering on contextual word embeddings.

1) Latent Semantic Indexing: Latent Semantic Indexing (LSI) is a technique initially introduced for automatic document classification [15] and information retrieval [16], but it was later found to be efficient for automatic text summarization [17][18] and its evaluation [19][20]. LSI typically applies a matrix factorization algorithm called Singular Value Decomposition (SVD) to the Term Frequency-Inverse Document Frequency (TF-IDF) matrix of the document. Our algorithm replicates that of Gong and Liu [17], including for sentence selection. For each article, the optimal number of topics n_{topics} was chosen by measuring C_v topic coherence [21]. We then chose $n_{topics} = \arg\max_{k \in [\![2,10]\!]} C_v(k)$. 2 and 10 were arbitrarily picked as initial bounds, and further analysis would be required to determine the optimal bounds.

Sentences are then categorized by topic, and the output summary is generated by looping through the topics and choosing the best-scoring sentence for each topic until the aforementioned maximum summary length of 280 characters is reached. In practice, this length restriction allows for very few sentences in a summary. Highly-ranked sentences are therefore skipped in favor of lower-ranked ones if the former would make the summary go over the character limit and the latter would not.

We apply LSI to the TF-IDF of a list of keywords generated by removing punctuation and stopwords from the article and lemmatizing the remaining words. Our intent in using keywords rather than the raw text is to eliminate noise that could be caused by stopwords, and to apply topic modeling only to the most important words in a document.

2) Word Embeddings and K-means Clustering: K-means clustering is an alternative way to model topics within a document, which has been successfully applied to text summarization before [22].

Our implementation uses contextual word embeddings as the raw input of k-means clustering, which are generated on a sentence basis using the pre-trained models FlauBERT [23] and CamemBERT [24] for French and RoBERTa [25] for English. Embeddings were chosen with the intuition that they capture semantic information. Specifically, contextual embeddings were used with the hope that they would perform better than a bag-of-words model such as TF-IDF. We arbitrarily choose $n_{clusters} = 5$ for k-means clustering, and further research is necessary to determine the optimal value.

Output scores are generated on a word basis (and not a sentence one) by computing the cosine similarity between a word's embedding and the centroids of the article's topics. For each cluster, the top words are picked and mapped back to the original sentence that contains them. The subsequent sentence selection and summary building process is then the same as for the LSI model.

Fig. 1. LSI model.

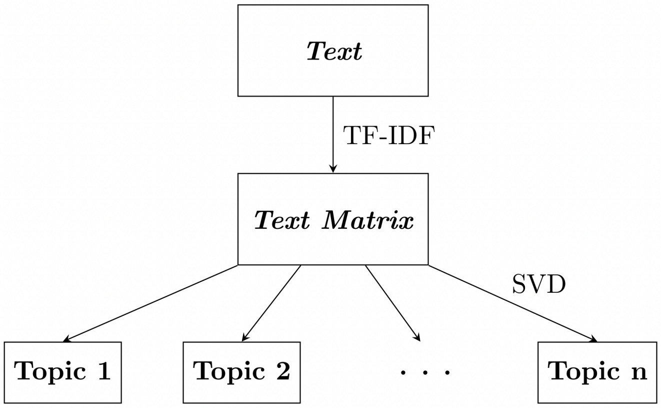
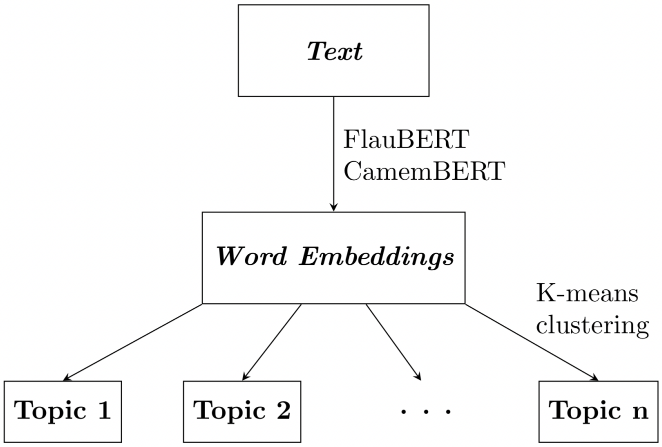


Fig. 2. K-means clustering model.



IV. Experimental Setup

This section contains quantitative information about our corpora, as well as the chosen methods of evaluation of the article classification and summarization methods.

A. Datasets

Our French corpus contains 787 articles from *Actu* and 1,803 from *L'Est Républicain*, adding up to a total of 2,503 articles along with their summaries. Our English corpus contains 2,010 articles from *The Guardian* with summaries.

The French version of the MLSUM corpus contains over 400,000 news articles. For evaluation, we restricted ourselves to the test set of that corpus, which contains 15,828 articles. Similarly, a test set of 11,490 articles was selected from the CNN/Daily Mail corpus.

B. Article Classification Evaluation

In a machine learning process, it is often insufficient to split the dataset into training and testing subsets and assume that the model will always perform well on unseen data. For this reason, k-fold cross-validation was used to further evaluate the classifier. This method divides data into a chosen number k of sets, of which $k - 1$ are used for training and the remaining one for testing. The choice of the testing set changes with each execution. A five-fold validation on all models was run and the values of all four metrics were collected: accuracy, precision, recall, and F1-score. The MNB, LR and SVM models were evaluated both before and after tuning.

C. Summarization Evaluation

To evaluate the summarization models, three different metrics were selected: the well-known ROUGE-L [26], as well as the much more recent, state-of-the-art metrics BERTScore [27] and Word Mover's Distance (WMD) [28]. These metrics calculate output scores via three different means: ROUGE-L, following a surface-level approach, attempts to align sequences of words in the generated and reference summaries. BERTScore evaluates texts on a deeper level by using contextual BERT embeddings to compute cosine similarity between two texts. Both of these metrics output three values per summary: precision, recall, and F1-score. Those are proportions ranging from zero to one, where a higher score indicates higher performance. Lastly, WMD computes a single score per summary, representing its semantic distance to the corresponding reference summary, modeled as a transportation problem. Its values range from 0 up to the size of the vocabulary in words, with a lower score indicating a higher quality summary.

We compute our metrics on two different forms of our data:

- The *standard score* is computed on pairs consisting of the generated summary and the reference summary.
- The *keyword score* is computed on pairs consisting of stemmed, keyword-only (see III-D1) versions of the generated and reference summaries.

To our knowledge, it is not common practice for text summarization models to evaluate keyword versions of summaries. The purpose behind adding this evaluation is to reduce the risk of score inflation due to stopword similarity. We also expect stemming to increase the opportunity for matching word subsequences in ROUGE-L between our generated summaries and the reference ones, which is the same motivation that led to the inclusion of a Porter stemmer module in the METEOR evaluation metric [29]. We do not expect this keyword score to be significantly different for BERTScore or WMD, as those rely on semantics through the use of embeddings.

All summarization and evaluation experiments were performed on the Grid'5000 testbed [30].

V. Results

Experimental results are split into two parts: one pertaining to article classification, and one concerning summarization.

A. Article Classification Results

Our article classification results are in Tables II and III. For the English classifier, results are expressed in terms of macro-averaged scores.

Even after tuning, the MNB consistently showed underwhelming performance compared to LR and SVM. Its performance was especially poor when it came to multi-class classification: the model barely classified any samples in the categories with lower support correctly. In contrast, LR and SVM showed comparably adequate results. They both performed slightly better on the French corpus; and the metrics for the multi-class corpus came close behind. Tuning LR and SVM using GridSearch improved their performance by approximately 1%.

For LR, the optimal value for C appeared to be a low value ($C = 100$). Lower values tend to be more fit for imbalanced datasets as they require more regularization and more weight to the complexity penalty to avoid overfitting. For SVM, we found that a larger C value ($C = 100$), which creates a smaller-margin hyperplane, improved the performance of our model.

B. Summarization Results

Our summarization results are shown in Tables IV and V.

We find that the LSI model outperforms the k-means clustering implementations for all scores and all datasets except for MLSUM. This is most noticeable for the ROUGE-L score. On MLSUM, k-means clustering performs slightly better depending on the chosen metric. We also observe that

TABLE II
Classification results for French.

Method	Accuracy	Precision	Recall	F1
Multinomial Naive Bayes	0.842	0.711	0.842	0.775
MNB (tuned)	0.845	0.732	0.845	0.781
Logistic Regression	0.934	0.934	0.934	0.934
LR (tuned)	0.943	0.943	0.943	0.934
Support Vector Machine	0.934	0.934	0.934	0.921
SVM (tuned)	0.943	0.943	0.943	0.934

TABLE III
Classification results for English.

Method	Accuracy	Precision	Recall	F1
Multinomial Naive Bayes	0.610	0.505	0.610	0.505
MNB (tuned)	0.628	0.505	0.628	0.505
Logistic Regression	0.934	0.934	0.934	0.934
LR (tuned)	0.935	0.935	0.935	0.935
Support Vector Machine	0.921	0.921	0.921	0.918
SVM (tuned)	0.932	0.932	0.932	0.932

TABLE IV
SUMMARIZATION EVALUATION FOR FRENCH (AVERAGE SCORES FOR WMD, AVERAGE F1-SCORES FOR OTHER METRICS).

Method	ROUGE-L	Keyword ROUGE-L	BERTScore	Keyword BERTScore	WMD	Keyword WMD
MLSUM corpus, test set (15,828 articles)						
LSI	**0.0652**	**0.0627**	0.5894	0.5885	**0.2517**	0.2652
FlauBERT + k-means	0.0564	0.0571	0.5905	**0.5909**	0.2558	0.2629
CamemBERT + k-means	0.0591	0.0598	**0.5907**	0.5904	0.2528	**0.2602**
Built corpus (787 + 1,803 = 2,560 articles)						
LSI	**0.1536**	**0.1538**	**0.6267**	**0.6238**	**0.2355**	**0.1959**
FlauBERT + k-means	0.1040	0.1075	0.6137	0.6134	0.2471	0.2080
CamemBERT + k-means	0.1093	0.1123	0.6153	0.6143	0.2452	0.2057

TABLE V
SUMMARIZATION EVALUATION FOR ENGLISH (AVERAGE SCORES FOR WMD, AVERAGE F1-SCORES FOR OTHER METRICS).

Method	ROUGE-L	Keyword ROUGE-L	BERTScore	Keyword BERTScore	WMD	Keyword WMD
CNN/Daily Mail corpus, test set (11,490 articles)						
LSI	**0.1207**	**0.1894**	**0.4947**	**0.4807**	**0.2178**	**0.1709**
RoBERTa + k-means	0.0839	0.1513	0.4680	0.4640	0.2342	0.1807
Built corpus (2,010 articles)						
LSI	**0.0748**	**0.1162**	**0.4822**	**0.4663**	**0.2297**	**0.2241**
RoBERTa + k-means	0.0533	0.0953	0.4702	0.4650	0.2390	0.2331

the keyword-only version of the various score is significantly better for ROUGE-L on both English corpora and for Word Mover's Distance on our French corpus and on the CNN/Daily Mail one, but that it produces either no improvement or a regression for other metrics and corpora. Finally, we observe that our English model performs significantly better on the CNN/Daily Mail corpus than on our custom one, while the opposite is true for French.

VI. DISCUSSION

Since there is no one metric that performs high and above all the others for article classification, we focus on one metric that we deem to be the most important. Therefore, we focus on the F1-score. We justify this choice with our goal to maximize the amount of true positives while minimizing the amount of false positives and false negatives, which translates to minimizing the misidentification of local and global articles in the model. Given this choice, both the tuned LR and tuned SVM methods are the best options for article classification. In the full pipeline, the tuned LR model was chosen for the classifier due to its slightly higher evaluation results.

In terms of summarization, we notice that LSI outperforms k-means clustering in nearly all metrics and situations. Although k-means clustering performs better with respect to BERTScore and Word Mover's Distance on the MLSUM dataset, the gain is of half a percentage point or less, whereas LSI can give a ROUGE-L score up to 5% higher on our French corpus. LSI has therefore been chosen as the default summarization method in the full summarization pipeline.

Our evaluation shows poor ROUGE-L scores for all datasets and summarization methods. This is easily explainable by the fact that ROUGE-L scores are computed on a purely surface level; in our case, as summaries often contain one to two sentences, it is unlikely to find large subsequences overlapping with the reference text. This coincides with the findings in [31], according to which ROUGE-L scores usually have very low correlation with human judgments.

On the other hand, the BERTScore reaches acceptable levels for English and good results for French. We observe similarly good results for Word Mover's Distance.

For French, the scores typically indicate a higher performance of the models on our corpus by several percentage points. We hypothesize that this could be due to different writing styles in the source newspapers causing changes such as different ratios of stopwords in both corpora. This may have led to articles in the MLSUM corpus being more strongly affected by the keyword scores.

For English, the opposite effect occurs: there is a significant decrease in performance across the board when switching from the CNN/Daily Mail corpus to our corpus. This could be due to a number of factors, such as our corpus containing a larger language diversity through inclusion of multiple varieties of English, and having longer average article length.

Using the keyword score seems to make little difference with respect to the BERTScore, which could be due to BERTScore's reliance on embeddings, the content of which may be more significant for keywords. The discrepancy between ROUGE-L scores for English are harder to account for, but may arise from a larger list of stopwords or more efficient stemming. It is unclear why the keyword-only version of Word Mover's Distance is significantly lower on the CNN/Daily Mail corpus and on our French corpus, while being nearly unchanged for the remaining two datasets.

We would like to draw attention to two important points. The first is that working with a maximum length of 280 characters severely restricts our models' ability to output a high-quality summary. When combined with the sentence selection algorithm, which is designed to minimize the amount of unused characters in the output tweet, it often happens that output summaries are not necessarily optimal with regard to our models. This is further reinforced by the fact that our models are extractive and typically have to work with articles consisting of a couple dozen sentences.

The second is that while looking deeper into samples of the generated summaries, we found that the summaries chosen by our model often outlined the articles well and matched a quality reference summary. For example, in the MLSUM LSI summaries, we notice a recurring issue where we feel that our model's summary matches the article, but receives a low ROUGE-L score due to a poor reference summary. This seems inevitable when working with a corpus of this magnitude, but it is important to note because it exhibits a way in which the scores may not always reflect the quality of a generated summary. To further exhibit this matter, a quick look at the generated summaries seems to show that BERTScore better represents the quality of our models. The following example, drawn from Article 54 of the French MLSUM testset, shows an instance of a quality summary given a poor score. It demonstrates that our model is able to select high-quality summaries, but that they will not always be evaluated as such since some of the reference summaries are low-quality.

- **Generated summary:** "Douze personnes ont été abattues vendredi 31 mai par un tireur dans un bâtiment municipal de Virginia Beach (Etat de Virginie), station balnéaire de la côte est américaine."
  [*On Friday, May 31st, twelve people were killed by a shooter in a municipal building in Virginia Beach (Virginia), a seaside resort on the east coast of the USA.*]
- **Reference summary:** "Le suspect principal, un employé des services de la ville, a tiré « à l'aveugle ». Il est lui aussi décédé."
  [*The main suspect, a city worker, fired "blindly". He also died.*]
- **Scores:**
  - ROUGE-L F1: **0.151** (standard), **0.066** (keyword)
  - BERTScore F1: **0.157** (standard), **0.119** (keyword)
  - WMD: **0.316** (standard), **0.348** (keyword)

To further evaluate the relevance of computed scores compared to human evaluation, one would need one or several native speakers to manually annotate each generated summary as good or bad, and to assess how those scores relate to our chosen metrics.

Finally, one issue that has affected all datasets is poor summary selection, which may be due to the performance of the summarization methods themselves.

As for the classification part of our task, we limited ourselves to traditional, supervised methods, such as Multinomial Naive Bayes, Support Vector Machine, and Logistic Regression. We are well aware that the state-of-art approaches now revolve mainly around deep learning (DL). It has also been established in scientific reviews that unsupervised models do demonstrate superior performance: XLNet-Large and XLNet showed consistently good classification results on multiple 'classic' datasets, such as IMDB and Yelp reviews [32]. For instance, for the SST-2 dataset, the best result yielded by a traditional model (Naive Bayes) was 81.8, while the best result yielded by a DL model (XLNet) was 97.0.

Even with the success of DL models today, there are a few key reasons simpler statistical models were chosen. DL models come with a handful of unique challenges that traditional supervised models do not have. For instance, most DL models cannot easily be interpreted, and poor interpretability makes it difficult to pinpoint exactly why one DL model outperforms another one [32]. Furthermore, it can be hard to decipher what a model has learned to achieve a desirable benchmark in order to use this insight later. Lastly, our custom corpora were not big enough to effectively train a DL model.

VII. Conclusion and Future Work

A web scraper was built and dynamic COVID-19 corpora covering French and English news articles were created and manually annotated for local relevance. Then, text classification and summarization models for news articles were created.

The classifier evaluation results are satisfactory, with a maximum F1-score of 93.4% (French corpus) and 93.5% (English corpus) in the tuned LR and SVM models. On the other hand, the summarization results are mixed, but we observe satisfying accuracy, especially after accounting for the restrictions put on the summarization model.

Lastly, a full pipeline was implemented that automatically selects and classifies news articles pertaining to COVID-19, summarizes them, and finally posts them to Twitter. In the future, the overall performance and usefulness of our tool could be improved by adding more news sources, and optimizing our summarization models or relaxing their constraints, such as the 280 character limit. Moreover, the framework and workflow can be easily adapted and deployed on other use-cases, such as different news topics.

All tweets produced by our pipeline can be found on the Newsjam Twitter account. The entirety of *Newsjam*'s code and results can also be found on GitHub [https://github.com/pie3636/newsjam].

Acknowledgments

The authors thank Miguel Couceiro and Maxime Amblard for their valuable guidance and time. This work was financially supported by Antoine Tabbone and the Institut des sciences du Digital, Management and Cognition, Nancy, France (IDMC).

References

[1] M. Savage, *Coronavirus: How much news is too much?* en, 2020. [Online]. Available: https://www.bbc.com/worklife/article/20200505-coronavirus-how-much-news-is-too-much Retrieved on 2022-05-10.

[2] K. Krawczyk *et al.*, "Quantifying Online News Media Coverage of the COVID-19 Pandemic: Text Mining Study and Resource," en, *JMIR*, vol. 23, no. 6, e28253, Jun. 2021, ISSN: 1438-8871. DOI: 10.2196/28253. [Online]. Available: https://www.jmir.org/2021/6/e28253 Retrieved on 2022-05-10.

[3] H. Saggion and T. Poibeau, “Automatic text summarization: Past, present and future,” in *Multi-Source, Multilingual Information Extraction and Summarization*, Berlin, Heidelberg: Springer, 2013, pp. 3–21.

[4] M. Allahyari *et al.*, “Text summarization techniques: A brief survey,” *CoRR*, vol. abs/1707.02268, 2017. arXiv: 1707.02268.

[5] D. Suleiman and A. Awajan, “Deep learning based abstractive text summarization: Approaches, datasets, evaluation measures, and challenges,” *Mathematical Problems in Engineering*, vol. 2020, pp. 1–29, 2020. DOI: 10.1155/2020/9365340.

[6] W. Soto, “X-pareval: A multilingual metric for paraphrase evaluation,” M.S. thesis, Université de Lorraine, 2021.

[7] C.-Y. Lin, “Looking for a few good metrics: Rouge and its evaluation,” in *NTCIR Workshop*, 2004.

[8] B. J. Dorr, C. Monz, S. President, R. Schwartz, and D. Zajic, “A methodology for extrinsic evaluation of text summarization: Does rouge correlate?” In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 1–8.

[9] N. Shulter, “The limits of automatic summarisation according to rouge,” in *Proceedings of the 15th Conference of the EACL: Volume 2, Short Papers*, 2017, pp. 41–45.

[10] M. K. Dalal and M. A. Zaveri, “Automatic text classification: A technical review,” *IJCA*, vol. 28, no. 2, pp. 37–40, 2011.

[11] K. Kowsari *et al.*, “Text classification algorithms: A survey,” *Information*, vol. 10, no. 4, p. 150, 2019.

[12] D. M. Magerman, *Statistical decision-tree models for parsing*, 1995. arXiv: cmp-lg/9504030 [cmp-lg].

[13] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano, “Mlsum: The multilingual summarization corpus,” *arXiv preprint arXiv:2004.14900*, 2020.

[14] K. M. Hermann *et al.*, “Teaching machines to read and comprehend,” *Advances in neural information processing systems*, vol. 28, 2015.

[15] H. Borko and M. Bernick, “Automatic document classification,” *JACM*, vol. 10, no. 2, pp. 151–162, 1963. DOI: 10.1145/321160.321165.

[16] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, “Indexing by latent semantic analysis,” *J. Am. Soc. Inf. Sci.*, vol. 41, pp. 391–407, 1990.

[17] Y. Gong and X. Liu, “Creating generic text summaries,” in *Proceedings of Sixth ICDAR*, 2001, pp. 903–907. DOI: 10.1109/ICDAR.2001.953917.

[18] M. Ozsoy, F. Alpaslan, and I. Cicekli, “Text summarization using latent semantic analysis,” *J. Information Science*, vol. 37, pp. 405–417, Aug. 2011. DOI: 10.1177/0165551511408848.

[19] J. Steinberger, K. Jezek, *et al.*, “Using latent semantic analysis in text summarization and summary evaluation,” *Proc. ISIM*, vol. 4, pp. 93–100, 2004.

[20] J. Steinberger and K. Jezek, “Evaluation measures for text summarization.,” *Computing and Informatics*, vol. 28, pp. 251–275, Jan. 2009.

[21] M. Röder, A. Both, and A. Hinneburg, “Exploring the space of topic coherence measures,” *WSDM 2015 - Proceedings of the 8th ACM International WSDM Conference*, pp. 399–408, Feb. 2015. DOI: 10.1145/2684822.2685324.

[22] K. Shetty and J. S. Kallimani, “Automatic extractive text summarization using k-means clustering,” in *2017 ICEECCOT*, 2017, pp. 1–9. DOI: 10.1109/ICEECCOT.2017.8284627.

[23] H. Le *et al.*, “Flaubert: Unsupervised language model pre-training for french,” *CoRR*, vol. abs/1912.05372, 2019. arXiv: 1912.05372.

[24] L. Martin *et al.*, “CamemBERT: A tasty French language model,” in *Proceedings of the 58th Annual Meeting of the ACL*, Online: ACL, Jul. 2020, pp. 7203–7219. DOI: 10.18653/v1/2020.acl-main.645. [Online]. Available: https://aclanthology.org/2020.acl-main.645.

[25] Y. Liu *et al.*, “Roberta: A robustly optimized BERT pretraining approach,” *CoRR*, vol. abs/1907.11692, 2019. arXiv: 1907.11692.

[26] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *ACL 2004*, 2004.

[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, *Bertscore: Evaluating text generation with bert*, 2020. arXiv: 1904.09675 [cs.CL].

[28] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, “From word embeddings to document distances,” in *Proceedings of the 32nd ICML - Volume 37*, ser. ICML’15, Lille, France: JMLR.org, 2015, 957–966.

[29] A. Lavie and A. Agarwal, “METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments,” in *Proceedings of the Second WMT*, Prague, Czech Republic: ACL, Jun. 2007, pp. 228–231. [Online]. Available: https://aclanthology.org/W07-0734.

[30] D. Balouek *et al.*, “Adding virtualization capabilities to the Grid’5000 testbed,” in *Cloud Computing and Services Science*, ser. Communications in Computer and Information Science, I. I. Ivanov, M. van Sinderen, F. Leymann, and T. Shan, Eds., vol. 367, Springer International Publishing, 2013, pp. 3–20, ISBN: 978-3-319-04518-4. DOI: 10.1007/978-3-319-04519-1_1.

[31] F. Liu and Y. Liu, “Correlation between rouge and human evaluation of extractive meeting summaries,” in *ACL*, 2008.

[32] S. Minaee *et al.*, “Deep learning–based text classification: A comprehensive review,” *ACM Comput. Surv.*, vol. 54, no. 3, Apr. 2021, ISSN: 0360-0300. DOI: 10.1145/3439726.