NewsJam

Automatically Summarizing French News about COVID-19

Shane Kaszefski-Yaschuk, Joe Keenan, Adele Khasanova, Maxime Méloux, Mahamadi Nikiema

Institut des Sciences du Digital
Pôle Herbert Simon
13 rue Michel Ney
54000 Nancy, France

14 January 2022

Motivation

Overload of COVID News

- Many people feeling overwhelmed about all of the constant COVID-19 updates [Savage, 2020]
- One study found that more than 26 million coronavirus related articles have been posted since the beginning of the pandemic [Krawczyk et al., 2021]

Enter NewsJam

- Concise, summarized news about the COVID-19 pandemic keeps readers informed while reducing this sort of 'news fatigue'
- Our solution: create a model which automatically scrapes, classifies, and summarizes articles then posts them to Twitter

Web Scraping

Two scrapers built:

- *L'Est Républicain*[1] & *Actu*[2]
- Scraping articles under 'COVID-19', 'Coronavirus', or 'Santé' labels
- Using *beautifulsoup* and *selenium* libraries

[1] `https://www.estrepublicain.fr/`
[2] `https://www.actu.fr`

Datasets

- French version of the MultiLingual SUMmarization corpus (MLSUM) [Scialom et al., 2020]
  - contains over 400,000 articles from *Le Monde*
  - split into train, test, and validation sets
- Custom corpora of COVID-19 news articles
  - 895 articles scraped from *Actu*
  - 1,703 from *L'Est Républicain*

Article Classification

Methods used:

- Logistic Regression
- Multinomial Naive Bayes
- Support Vector Machine

Annotation

The *L'Est Républicain* portion of our built corpus was annotated manually

- Using agreed-upon annotation guidelines, 3 annotators marked articles as either 0 (global) or 1 (local)
- Annotator A annotated the entire corpus
- Annotators B and C annotated 50 articles each
- 65% local articles, 35% global
- Used as train and test sets for the classifier models

Annotation Evaluation

Coefficient	A & B	A & C	B & C
A_o	0.660	0.660	0.840
IAA_S	0.500	0.500	0.500
IAA_π	0.505	0.534	0.520
IAA_κ	0.505	0.519	0.513

Table: Inter-annotator agreement for each pair of annotators

Summarization

Two extractive methods of summarization were implemented for this project:

- Latent Semantic Indexing (LSI)
- K-Means Clustering on Contextual Word Embeddings
- Why did we choose these methods?

Latent Semantic Indexing

- Technique initially introduced for automatic document classification and information retrieval [Borko and Bernick, 1963]
- Also useful for Text Summarization
- Basic idea:
  - Bag of words (BOW) model
  - Topic modeling
- Main steps:
  - Factor the TF-IDF Matrix
  - Calculate optimal number of topics
  - Automatically assign sentences to topics
  - Sentence selection & Summary building

BERT Embeddings and K-Means Clustering

- Also topic modeling
- Used for summarization [Shetty and Kallimani, 2017]
- Typically uses TF-IDF as input
- However, our model uses BERT contextual embeddings as input:
  - FlauBERT
  - CamemBERT
- Main steps:
  - Calculate optimal number of topics via K-means clustering on BERT embeddings
  - Automatically assign sentences to topics
  - Sentence selection & Summary building

Classification Results

Method	Accuracy	Precision	Recall	F1
Multinomial Naive Bayes (MNB)	72.1	70.8	**97.7**	82.1
MNB (resampled)	67.1	88.3	38.6	53.1
MNB (tuned)	83.1	**92.2**	81.1	86.2
Logistic Regression (LR)	83.8	88	87.3	87.7
LR (tuned)	**85.7**	88.3	90.5	**89.3**
Support Vector Machine (SVM)	82.8	83.3	91.3	87.1
SVM (tuned)	85.2	87.5	90.6	89.1

Summarization Results

Method	ROUGE-L	Keyword ROUGE-L	BERTScore	Adj-BERTScore
MLSUM corpus, testset (15,828 articles)				
LSI	0.1507	0.1147	—	—
FlauBERT + k-means	—	—	—	—
CamemBERT + k-means	—	—	—	—
Built corpus (895 + 1,703 = 2,598 articles)				
LSI	**0.1589**	**0.1566**	**0.2636**	**0.7993**
FlauBERT + k-means	0.0879	0.0821	0.2374	0.7198
CamemBERT + k-means	0.0902	0.0850	0.2385	0.7231

Summarization Execution Time

MLSUM corpus	**Exec. time**
LSI	**4.79s**
FlauBERT + k-means	14.77s
CamemBERT + k-means	18.98s

Built corpus	**Exec. time**
LSI	**4.57s**
FlauBERT + k-means	15.53s
CamemBERT + k-means	19.83s

Generated Summary Example

Quality Summary | Poor Score:

- MLSUM Test Dataset Article 54[3]
- *Reference Summary*: "Le suspect principal, un employé des services de la ville, a tiré « à l'aveugle ». Il est lui aussi décédé."
- *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."
  - ROUGE-L
    - Standard F1-Score: 0.151
    - Keyword F1-Score: 0.066
  - BERTScore
    - Standard F1-Score: 0.157
    - Keyword F1-Score: 0.119

[3] `https://bit.ly/3n9qRuK`

Model Selection

Based on performance, these models were selected:

- Classification: Tuned Logistic Regression
- Summarization: Latent Semantic Indexing

LSI is the default summarizer, but BERT embeddings can be used as well

Pipeline

With all the components of our project complete, a full pipeline was implemented. It works as follows:

1. Choose summarizer and x number of articles
2. Scraper grabs the newest x articles and feeds them into the classifier
3. Valid articles classified as 'French' are fed into the summarizer
4. Generated summary & URL are posted to Twitter

Demonstration

The results of this demonstration can be seen on the Twitter account
`https://www.twitter.com/newsjam_fr`

Conclusion

- The NewsJam twitter page can be found at: `https://www.twitter.com/newsjam_fr`
- The NewsJam GitHub repository can be found at: `https://github.com/pie3636/newsjam`

Bibliography

Borko, H. and Bernick, M. (1963).

Automatic document classification.
Journal of the ACM (JACM), 10(2):151–162.

Krawczyk, K., Chelkowski, T., Laydon, D. J., Mishra, S., Xifara, D., Gibert, B., Flaxman, S., Mellan, T., Schwämmle, V., Röttger, R., Hadsund, J. T., and Bhatt, S. (2021).
Quantifying Online News Media Coverage of the COVID-19 Pandemic: Text Mining Study and Resource.
Journal of Medical Internet Research, 23(6):e28253.

Savage, M. (2020).

Coronavirus: How much news is too much?

Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., and Staiano, J. (2020).

Mlsum: The multilingual summarization corpus.
arXiv preprint arXiv:2004.14900.

Shetty, K. and Kallimani, J. S. (2017).

Automatic extractive text summarization using k-means clustering.
In *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, pages 1–9.