Newsjam

Text Summarization Meets Social Media

Shane Kaszefski-Yaschuk – Joseph Keenan – Adelia Khasanova – Maxime Méloux – Mahamadi Nikiema

Introduction/ Motivation

- **65%** increase in social media use during the pandemic[1]
- Online news sources and social media are the main sources of information[2]
- **77%** of American adults report experiencing news fatigue[3]

[1]Statista Research Department (2021), Additional daily time spent on social media platforms by users in the United States due to coronavirus pandemic as of March 2020, Statista

[2]Nielsen et al. (2021), An Ongoing Infodemic: How People in Eight Countries Access and Rate News and Information about Coronavirus a Year into the Pandemic, Reuters Institute

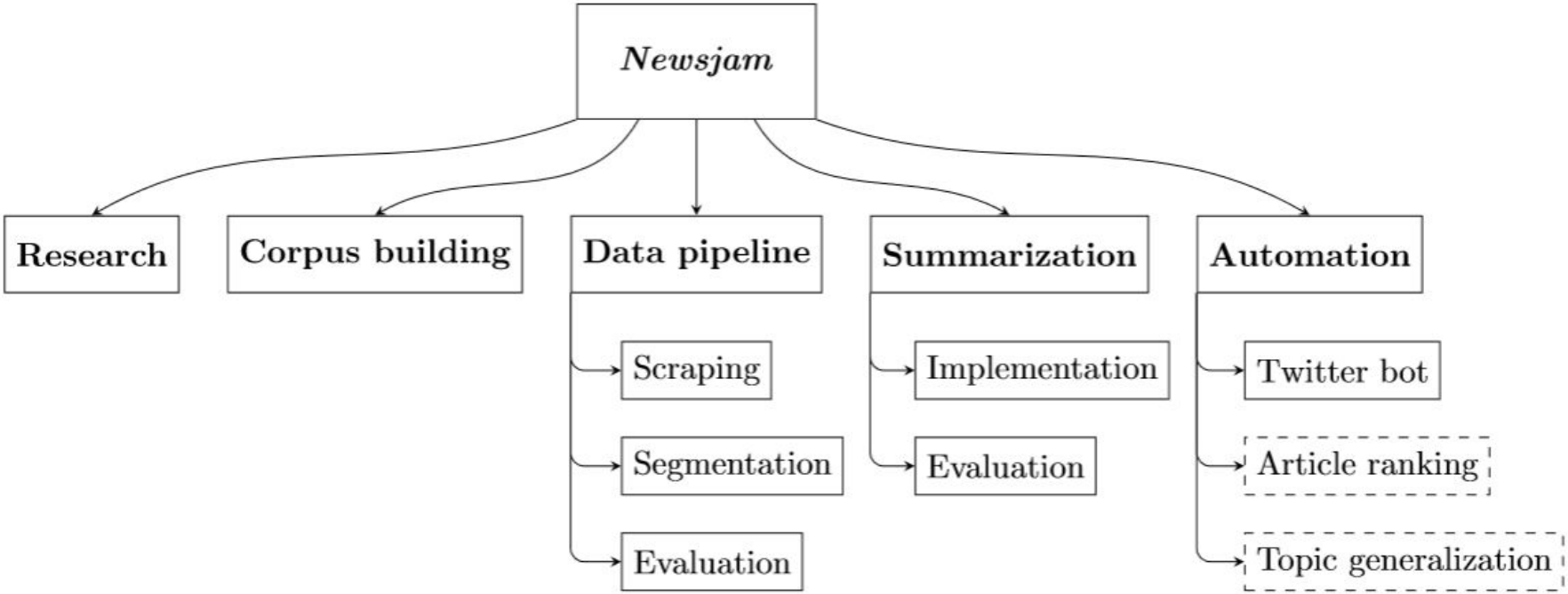
[3]Watson, A. (2020), Share of adults who feel overwhelmed by the amount of information in coronavirus news coverage in the United States as of April 2020, by age group, Statista

Enter **Newsjam**

- Concise, summarized news about the COVID-19 pandemic reduce news fatigue while keeping readers informed;

- Our solution: to create a **Twitter bot** which automatically scrapes, summarises, and posts tweets about recent COVID-19 news.

Project Overview



Methodology - Corpus building

- Libraries: `beautiful soup, selenium`
- Strategy: scraping articles under categories “Covid-19” and “Santé”
- Results: 47 articles from Actu and 1,820 from L’est Republicain

Sought news category

Covid-19

Headline

Vaccination : tout savoir sur la dose de rappel en sept questions

Reference summary

Une dose de rappel pour qui, pour quand? Quelle différence entre une "dose de rappel" et une "troisième dose"? Le président a dessiné mardi soir les contours de la stratégie de lutte contre le Covid-19 pour les prochains mois.

Par **J. C. avec AFP** - Hier à 15:00 - Temps de lecture : 6 min

Full text

Dès le 15 décembre, le pass sanitaire des plus de 65 ans ne sera plus valide si le rappel vaccinal n'a pas été effectué. Photo Alain JOCARD/AFP

Other Corpora

- MultiLingual SUMmarization corpus (MLSUM) (Scialom et al, 2020)
    - French version
        - Over 400,000 news article/reference summary pairs
        - Articles from *Le Monde* between 2010-2019
        - Reference summaries come from highlights section
    - Split into train, test, and validation sets
        - Train: 392,902 articles
        - Test: 15,828 articles
        - Validation: 16,059 articles

Other Corpora

```
pprint(mlsum['train'][4])
```

```
{'date': '01/01/2010',
 'summary': 'Cinq personnes sont mortes, et treize autres ont été blessées à '
            "Nîmes, dans le Gard, dans un incendie qui s'est déclenché "
            'vendredi 1er janvier au petit matin.',
 'text': 'Cinq personnes sont mortes, et treize autres ont été blessées à '
         "Nîmes, dans le Gard, dans un incendie qui s'est déclenché vendredi "
         "1er janvier au petit matin. Le feu, dont on ignore l'origine pour "
         'l\'instant, a pris au sixième et dernier étage d\'un immeuble. "A '
         "l'arrivée des pompiers, trois personnes étaient décédées dans un "
         'appartement, et deux autres dans un appartement voisin par '
         'intoxication", a expliqué, sur i-Télé, le directeur de cabinet du '
         'préfet du Gard. On dénombre également "treize blessés, dont trois '
         'graves. Une personne dans le coma a été transférée à Marseille", '
         'a-t-il ajouté. Les secours ont été prévenus vers 5 heures du matin, '
         'mais "l\'incendie avait déjà bien démarré", a-t-il expliqué. France '
         'Info précise que les victimes sont trois adultes et deux enfants. '
         '"L\'origine de l\'incendie est indéterminée mais a priori '
         'accidentelle", a déclaré le procureur adjoint de la République de '
         'Nîmes, cité par Europe 1.',
 'title': 'Cinq morts dans un incendie à Nîmes',
 'topic': 'societe',
 'url': 'https://www.lemonde.fr/societe/article/2010/01/01/cinq-morts-dans-un-incendie-a-nimes_1286657_3224.html'}
```

Methodology - Summarisation 1

Abstractive vs Extractive methods

Latent Semantic Analysis/Indexing (LSA/LSI)

- Technique initially introduced for automatic document classification and information retrieval (Borko and Bernick, 1963)
- Also useful for Text Summarization
- Bag of words (BOW) model
- Topic modelling
    - Number of topics? -> Topic coherence
- Main idea: Factoring the TF-IDF matrix
- Sentence selection
- Summary building

Methodology - Summarisation 2

K-means clustering

- Also topic modelling
- Used for summarization (Shetty and Kallimani, 2017)
- Typically uses TF-IDF as input (BOW)
- Our model uses contextual BERT-type embeddings (FlauBERT, CamemBERT)
    - Now contextual
- Number of topics?
- Sentence selection and building summary

Evaluation methods: quick comparison

ROUGE-L

- Initially created specifically for evaluating summaries;
- Longest Common Subsequences (LCS) of words between a reference summary and a generated summary;
- Computes scores based on the number and length of LCS' between a pair of summaries

BERTScore

- Initially created for evaluating text generation, but can be used for summarization as well;
- Evaluates on a deeper semantic level by utilizing contextual word embeddings;
- BERTScore measures are computed via cosine similarity between embeddings

Results

Method	Long ROUGE-L F1	Keyword ROUGE-L F1	Long BERTScore F1
MLSUM corpus, testset (15 828 articles)			
LSI	0.1507	0.1147	TBA
FlauBERT + k-means	TBA	TBA	TBA
CamemBERT + k-means	TBA	TBA	TBA
Actu corpus (47 articles)			
LSI	0.5391	0.5309	0.5666
FlauBERT + k-means	0.2911	0.2862	0.2098
CamemBERT + k-means	0.2463	0.2290	0.2941

Sample results

Good summary/Low score

- Reference summary:
  - «Le suspect principal, un employé des services de la ville, a tiré « à l'aveugle ». Il est lui aussi décédé.»
- Our 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.»
- F-measure: 0.1509

Future directions

- Improve the performance of summarization models/Test out other methods
- Explore more evaluation metrics
- Automate scrapers and include more news sources
- Create the Twitter bot that will post news recaps
- Legal aspect?

Conclusion

What we have accomplished so far:

- Built two corpora
- Implemented two different extractive summarization methods
- Applied two evaluation metrics
- Obtained the best result of 56% F1-score on Actu corpus of 47 articles
- Laid a solid foundation for future work

Sources (partial list)

Borko and Bernick, 1963. Automatic document classification. *Journal of the ACM (JACM)*, 10(2):151–162.

Deerwester et al. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Devlin et al, 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Gong and Liu, 2001. Creating generic text summaries. In *Proceedings of Sixth International Conference on Document Analysis and Recognition*, pages 903–907.

Gupta, 2020. Understanding Text Summarization using K-means Clustering.

Sources (partial list)

Röder et al, 2015. Exploring the space of topic coherence measures. WSDM 2015 - *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 399–408.

Scialom et al, 2020. Mlsum: The multilingual summarization corpus. arXiv preprint arXiv:2004.14900.

Shetty and Kallimani, 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.

Steinberger et al, 2004. Using latent semantic analysis in text summarization and summary evaluation. Proc. *ISIM*, 4:93–100.

Zhang et al, 2020. Bertscore: Evaluating text generation with bert.