# Newsjam

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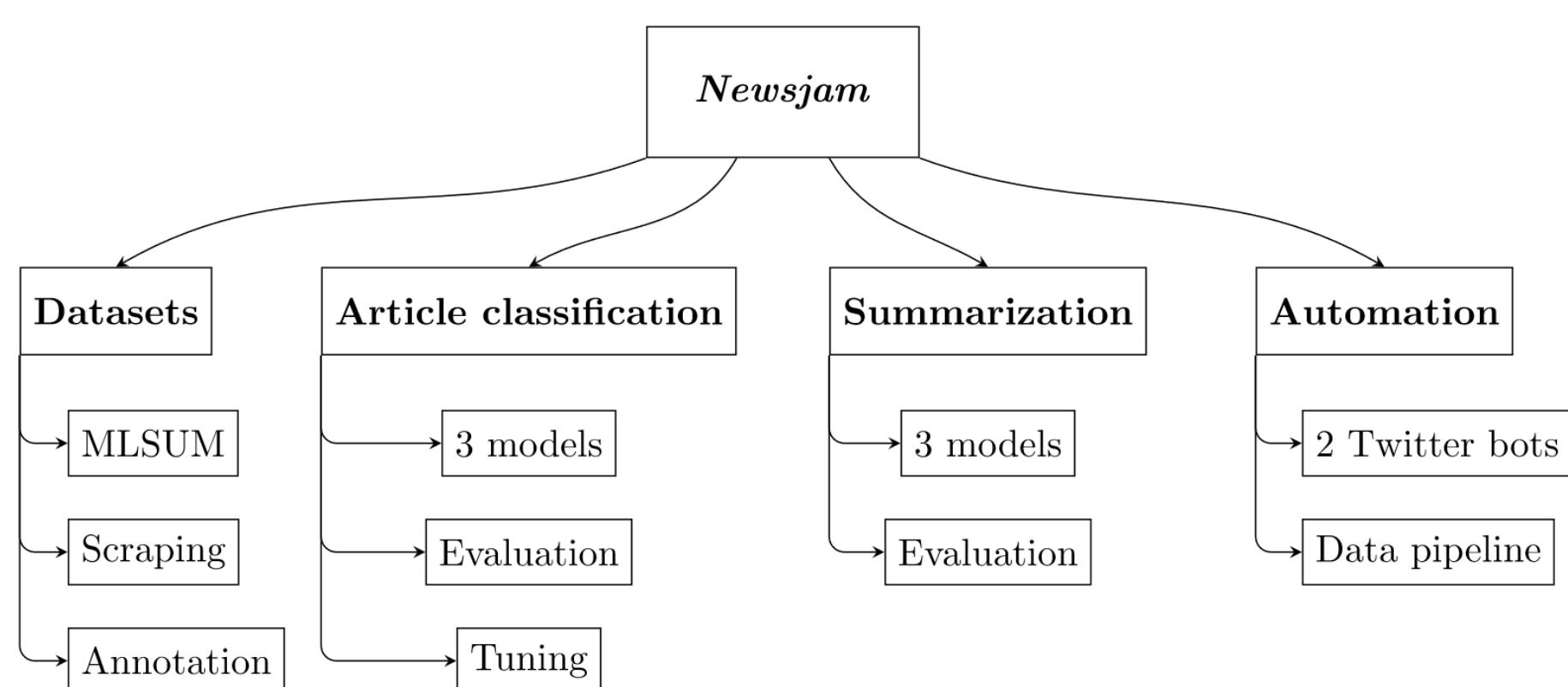## Maxime Méloux, Mahamadi Nikiema

## Presentation

- *Newsjam*: an automatic summarization tool for French COVID-19 news articles

- Articles are fetched, summarized and posted to Twitter

- Two Twitter feeds: Articles about France and global ones

## Project Overview



## Scraping and Datasets

Two French news sites were selected for scraping and another corpus was used for training and evaluating our models.

- *Actu.fr* - 895 articles scraped, including summaries

- *L'est Republicain* - Approximately 1,300 articles

- MLSUM (MultiLingual SUMmarization) Corpus - Over 400,000 articles from *Le Monde*, with reference summaries, split into training, test, and validation sets

## Summarizing

Summarization Methods:

- Latent Semantic Indexing (LSI)

- K-means clustering + FlauBERT embeddings
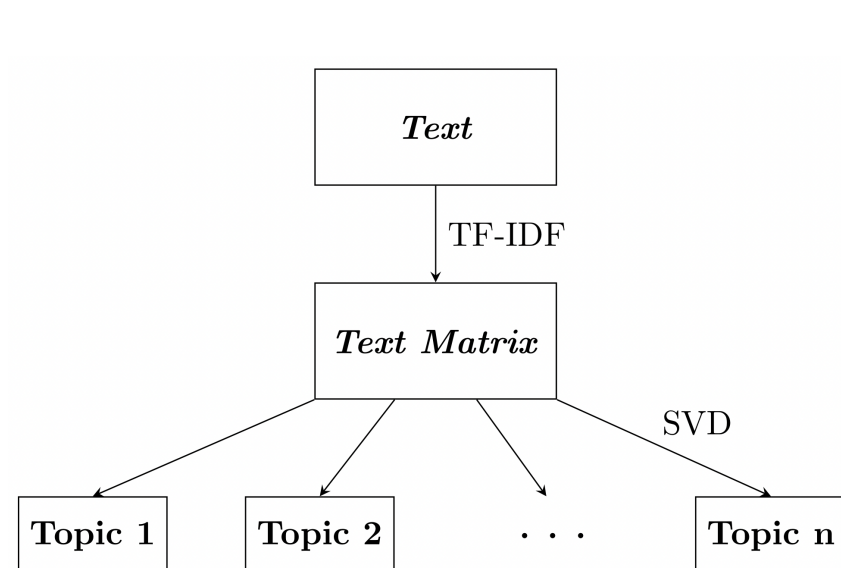
- K-means clustering + CamemBERT embeddings
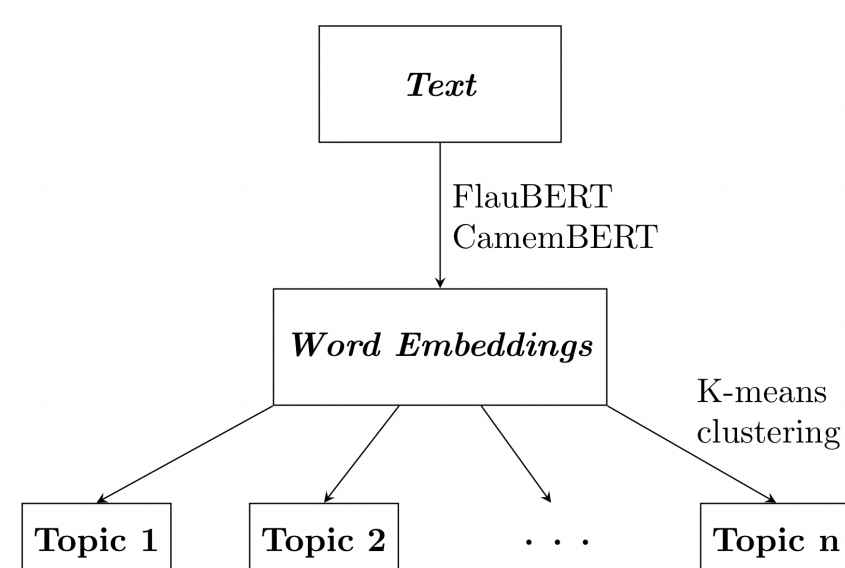


Fig. 1: LSI



Fig. 2: K-means clustering

| Scores | Sentence $s_1$ | Sentence $s_2$ | Sentence $s_3$ | Sentence $s_4$ | . . . | Sentence $s_n$ |
|--------|------|------|------|------|------|------|
| Topic $t_1$ | 0.18 | -0.54 | 0.32 | **0.67** | . . . | 0.13 |
| Topic $t_2$ | **0.87** | -0.23 | -0.50 | 0.72 | . . . | -0.20 |
| Topic $t_3$ | 0.54 | 0.02 | 0.09 | **-0.89** | . . . | -0.12 |
| . . . | . . . | . . . | . . . | . . . | . . . | . . . |
| Topic $t_n$ | 0.19 | **0.56** | 0.34 | 0.26 | . . . | -0.28 |

Table 1: Sentence selection algorithm

$$\text{Summary} = \{(\arg\max_{s_i \in S} Score(s_i, t_j)) \mid t_j \in T\}$$
$$= \{s_1, s_2, s_4\}$$

## Article Classification

To separate articles about France from those about the rest of the world, three classification methods have been compared:

- Multinomial Naive Bayes

- Logistic Regression

- Support Vector Machine

## Evaluation

- Summarization Evaluation Metrics:

  - ROUGE-L

  - BERTScore

- Two sets of our data evaluated:

  - *Standard summaries* are made up of the full-length summaries from our model.

  - *Keyword summaries* are simplified versions of the standard summaries with stopwords removed and the remaining keywords stemmed.

## 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 |

Table 2: Article classification evaluation

| Method | ROUGE-L | Keyword ROUGE-L | BERTScore | Exec. time |
|--------|---------|-----------------|-----------|------------|
| *MLSUM corpus, testset (15 828 articles)* | | | | |
| LSI | 0.1507 | 0.1147 | TBA | TBA |
| FlauBERT + k-means | TBA | TBA | TBA | TBA |
| CamemBERT + k-means | TBA | TBA | TBA | TBA |
| *Built corpus (895 + 1,753 = 2,648 articles)* | | | | |
| LSI | 0.5391 | 0.5309 | 0.5666 | TBA |
| FlauBERT + k-means | 0.2911 | 0.2862 | 0.2098 | TBA |
| CamemBERT + k-means | 0.2463 | 0.2290 | 0.2941 | TBA |

Table 3: Summarization evaluation (only F1-Scores reported)

## Future Directions

- Create Twitter bots

- Fully implement pipeline that fetches, classifies, and summarizes articles, then posts them to Twitter

## Sources

Horacio Saggion and Thierry Poibeau. 2013. Automatic text summarization: Past, present and future. In *Multi-Source, Multilingual Information Extraction and Summarization*, pages 3–21. Springer, Berlin, Heidelberg.