UCLouvain

CENTAL
centre de traitement automatique du langage

ORM
OBSERVATOIRE DE RECHERCHE SUR LES MÉDIAS ET LE JOURNALISME

Detecting opinion in news:
An automated analysis of linguistic subjectivity in French-language press articles

Louis Escouflaire – Antonin Descampe – Cédrick Fairon

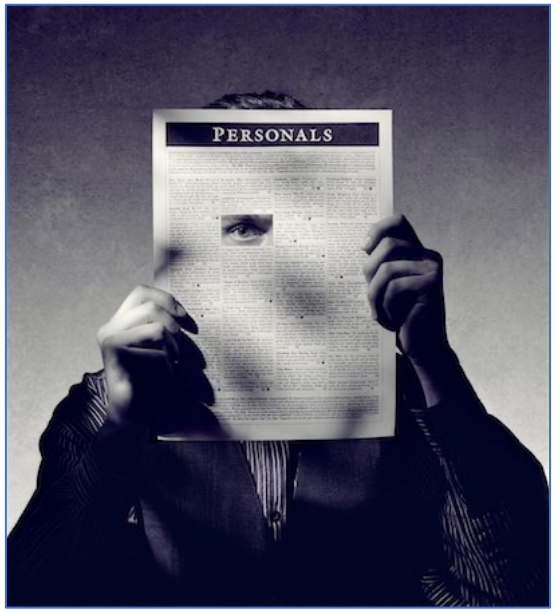
fnrs
LA LIBERTÉ DE CHERCHER

8th December 2022

BIJU 4th Biennial conference
A True and Fair View

State of the art

- **Journalistic objectivity** is altogether a structural norm, a set of editorial practices, and a **writing style**. [Koren, 2004]
- Through a variety of writing techniques (strategic ritual of objectivity), journalists can **mask their enunciative subjectivity** [Tuchman, 1972] :
    - Impersonal verbs and phrases
    - Less adjectives and adverbs
    - More citations and numbers
    - Neutral punctuation
    - …
- Most journalists admit that complete objectivity is an **unattainable ideal** towards which they still must strive. [Lagneau, 2002]



State of the art

- In **linguistics**, a sentence can be considered objective if it does not contain any subjective unit. [Wiebe et al., 2005]
- **Subjective units** are « linguistic cues through which the speaker/writer inscribes him/herself in the statement » [Kerbrat-Orecchioni, 1980]
- They can be morpho-syntactic, lexico-semantic and stylistic. [Krüger et al., 2017]

Morpho-syntactic	Lexico-semantic	Stylistic
• Pronouns : *I, you, we*...	• Words with positive/negative connotation	• Expressive punctuation : *? !* ...
• Negations	• Word complexity	• Sentence length
• More adjectives, less verbs	• ...	• Less citations
• ...		• ...

UCLouvain

Context

- **Opinions** have been shown to be more and more present in North-American news. [Alhindi et al., 2020]
- In Europe, 50% of users (75% of users under 30) inform themselves primarily through **social media**. [Matsa et al., 2018]
- **Recommandation algorithms** tend to favor emotional and polarizing content. [Diakopoulos, 2019]

Objectives :

- Understand the **linguistic mechanisms** of subjectivity in news articles.
- Build an **efficient and explainable algorithm** for automated subjectivity detection.
- Provide **contextualization cues** for readers of online information.

UCLouvain

Experiment n°1

Automated classification of *opinion* vs *information* articles.

- Data : **13.400 French articles** from RTBF (Belgian public service)
  - 6.700 **opinion** pieces (editorials, reviews…)
  - 6.700 **information** pieces (news, press agency dispatches…)
- **30 linguistic features** of subjectivity were extracted from each article. [Krüger et al., 2017]
- A **logistic regression classifier** was trained on the features :
  - Trained on 90% of data (10-fold cross-validation), evaluated on 10%.
  - 18 best features selected in order to build the best classifier.
- Results : **89% accuracy**

rtbf.be

UCLouvain

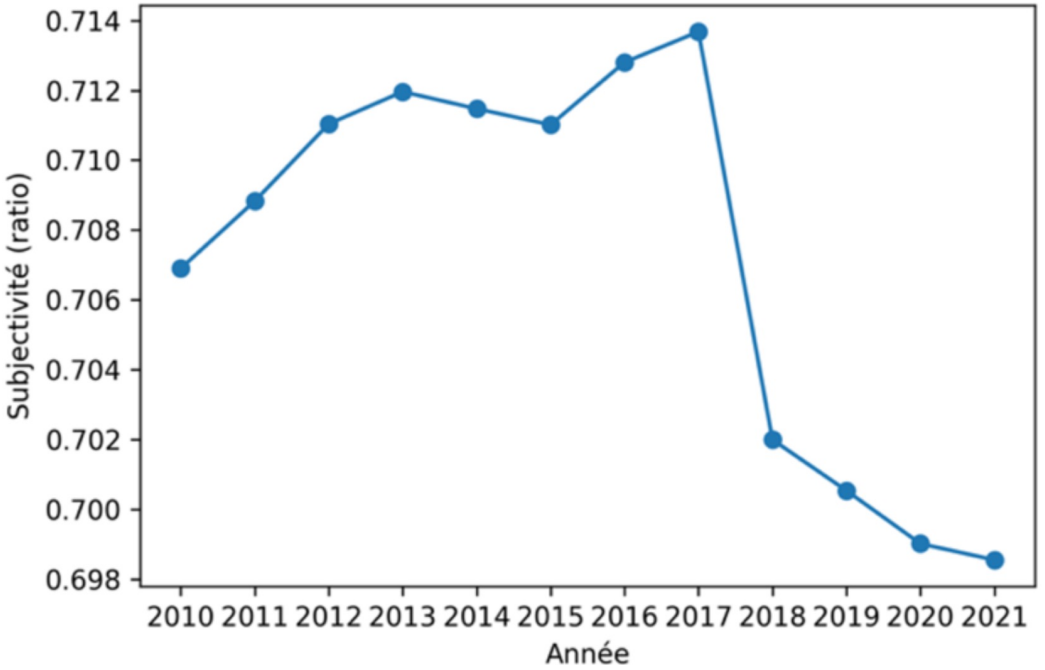
Experiment n°2

How has subjectivity evolved over 12 years?

- Data: **120.000** RTBF ***information*** articles
  - 10.000 articles/year (2010 ➔ 2021)
  - 36 M tokens in total
- A **subjectivity score** was assigned to each article (0 < *score* < 1)

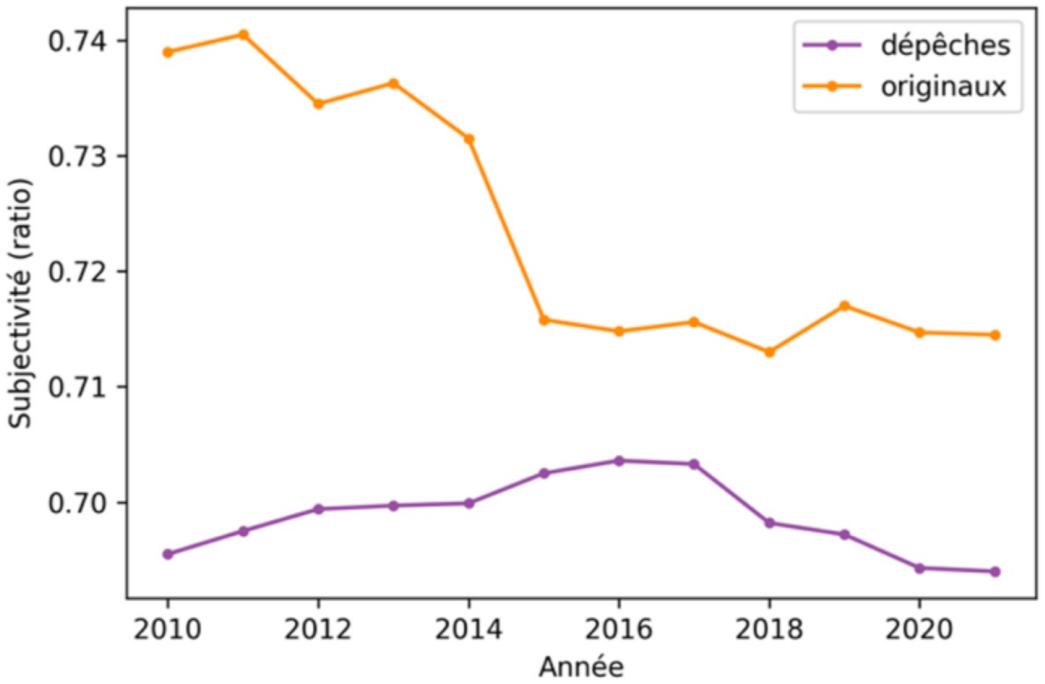
- **Interviews** with 2 chief editors from RTBF to interpret our results:
  - **2017 ➔** complete replacement of the redaction team

  + dispatch publication automated



Mean subjectivity score of RTBF articles between 2010 and 2021

Experiment n°2



Mean subjectivity score of dispatches and original RTBF articles from 2010 to 2021.

Comparative analysis of *dispatches* vs *originals*

- Press agency dispatches are significantly **less subjective** than articles written by RTBF journalists.
- **2014 ➔** all journalists (TV, radio…) need to produce content for the RTBF website.
  - ➢ **Uniformization** of writing (guidelines) and **optimization** for algorithmithic visibility (Google and social media)
  - ➢ Dispatches need to be "**enriched**" before publication.

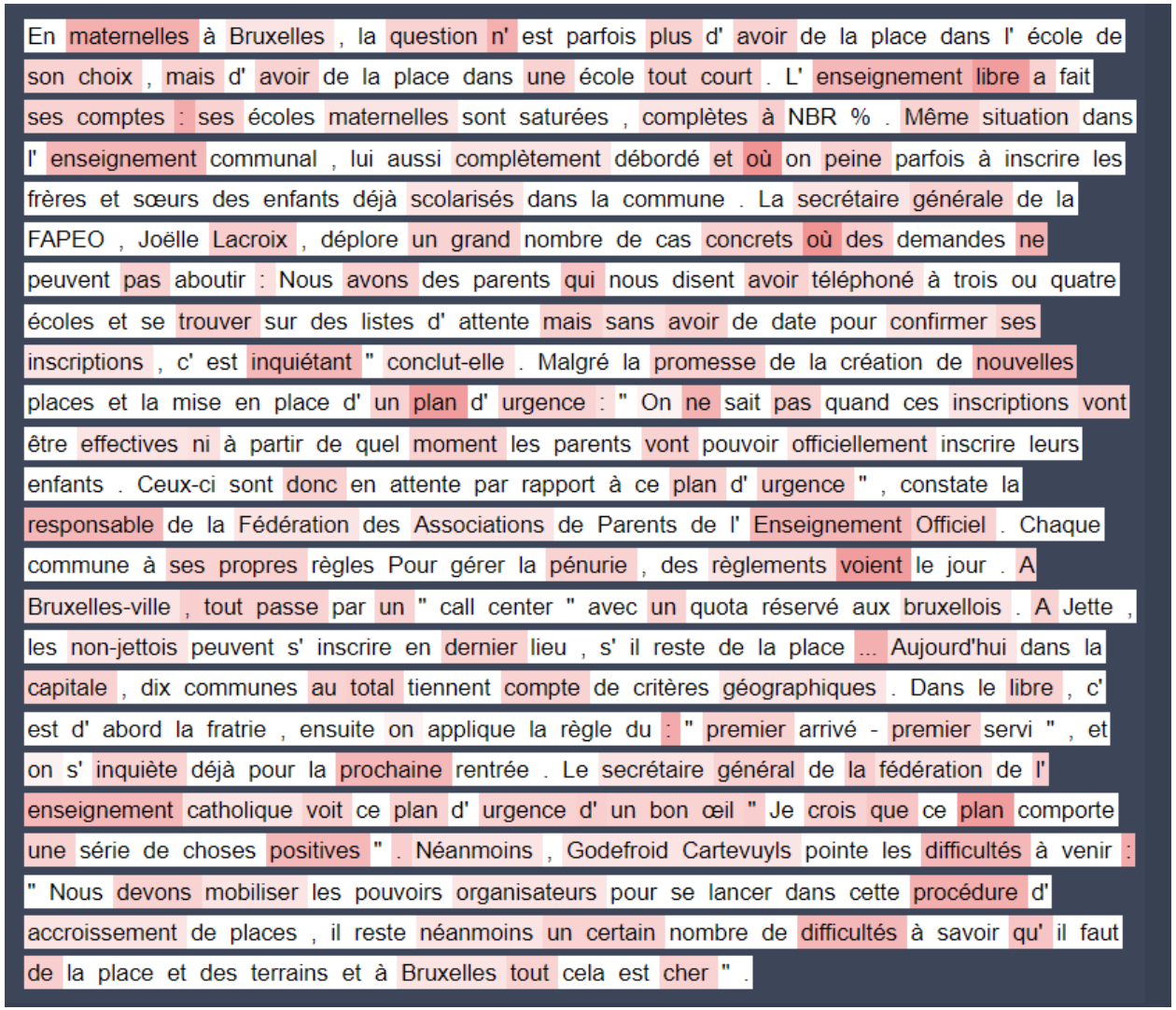
Conclusions and perspectives

- ✓ Our observations **reflect** internal editorial movements, which supports the validity of the score.
- ✓ Recommandation algorithms have **influenced** RTBF's journalistic writing, but **not as expected**.

- ❑ Can this model be used as a **tool** for journalists and readers to improve media literacy?
- ❑ This model was built using linguistic rules. But how and how well do **humans** and **AI** analyze subjectivity?

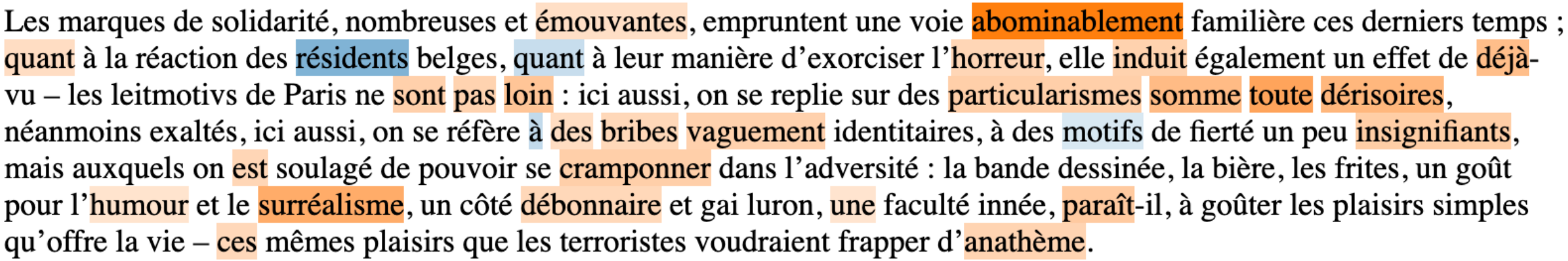
Further research:

- We built an explainer to **visualize** our linguistic model's decisions.



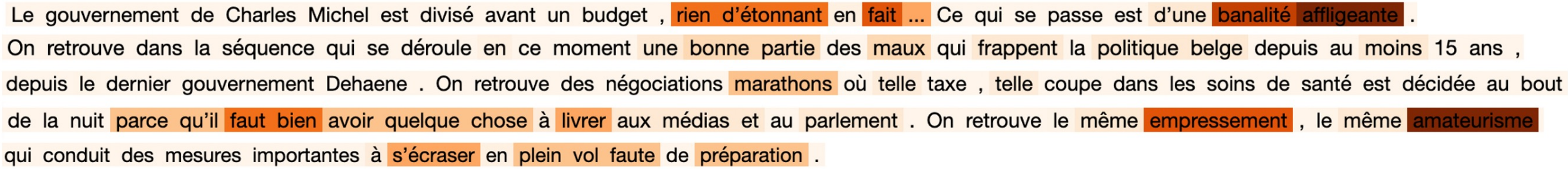
Further research:

- We built an explainer to **visualize** our linguistic model's decisions.
- We trained a ***deep learning* model** (CamemBERT ; Martin et al., 2019) to classify *information* and *opinion* articles (**97% accuracy**).



Further research:

- We built an explainer to **visualize** our linguistic model's decisions.
- We trained a ***deep learning* model** (CamemBERT ; Martin et al., 2019) to classify *information* and *opinion* articles (**97% accuracy**).
- We asked **40 students** in journalism to highlight the **subjective elements** in 150 RTBF articles.



We are **comparing** the 3 approaches in order to build an efficient and explainable model for automated subjectivity analysis.

UCLouvain

OBSERVATOIRE DE RECHERCHE
SUR LES MÉDIAS ET LE JOURNALISME

CENTAL
centre de traitement
automatique du langage

Thank you!

fnrs
LA LIBERTÉ DE CHERCHER

louis.escouflaire@uclouvain.be
antonin.descampe@uclouvain.be
cedrick.fairon@uclouvain.be

Abdaoui A., Aze J., Bringay S., & Poncelet P. (2017). FEEL: A French Expanded Emotion Lexicon. *Language Resources and Evaluation*, 51(3), 833-855.

Alhindi T., Muresan S., & Preotiuc-Pietro D. (2020). Fact vs. Opinion: The Role of Argumentation Features in News Classification. *Proceedings of the 28th International Conference on Computational Linguistics*, 6139-6149. DOI : 10.18653/v1/2020.coling-main.540.

Carlebach, M., Cheruvu, R., Walker, B., Magalhaes, C. I., & Jaume, S. (2020). News Aggregation with Diverse Viewpoint Identification Using Neural Embeddings and Semantic Understanding Models. In *Proceedings of the 7th Workshop on Argument Mining*, 59-66.

Dufrasne, M., & Philippette, T. (2019). Les effets des «bulles de filtres» ou «bulles informationnelles» sur la formation des opinions. *Journée d'étude pour le lancement du projet Alg-opinion*.

Escouflaire, L. (2022). Identification des indicateurs linguistiques de la subjectivité les plus efficaces pour la classification d'articles de presse en français. *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles,* pp. 69-82.

Escouflaire, L., Fairon, C. & Descampe, A. (2022). L'évolution de la subjectivité linguistique dans le journalisme web du XXIe siècle : analyse d'un corpus belge francophone d'articles de 2010 à 2020. *JADT2022*.

Kerbrat-Orecchioni, C. (2009). *L'énonciation : de la subjectivité dans le langage*. Armand Colin.

Keyvanpour, M., Zandian, Z. K., & Heidarypanah, M. (2020). OMLML: a helpful opinion mining method based on lexicon and machine learning in social networks. *Social Network Analysis and Mining, 10(1)*, 1-17.

Koren R. (2004). Argumentation, enjeux et pratique de l'« engagement neutre » : Le cas de l'écriture de presse. *Semen, 17*.

Krüger, K. R., Lukowiak, A., Sonntag, J., Warzecha, S., & Stede, M. (2017). Classifying news versus opinions in newspapers: Linguistic features for domain independence. *Natural Language Engineering, 23(5)*, 687.

Cox D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological), 20(2)*, 215–232.
Diakopoulos, N. (2019). *Automating the news : How algorithms are rewriting the media* (Harvard University Press).
Fisher R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics, 7(2)*, 179-188.
Loria S. (2018). *TextBlob Documentation. Release 0.15, 2.*
Mohammad S. & Turney P. (2013). Crowdsourcing a Word-Emotion Association Lexicon, *Computational Intelligence, 29 (3)*, 436-465.
Matsa, K. E., Silver, L., Shearer, E., & Walker, M. (2018, octobre 30). Western Europeans Under 30 View News Media Less Positively, Rely More on Digital Platforms Than Older Adults. *Pew Research Center's Journalism Project*. https://www.journalism.org/2018/10/30/western-europeans-under-30-view-news-media-less-positively-rely-more-on-digital-platforms-than-older-adults/
Tuchman G. (1972). Objectivity as strategic ritual: An examination of newsmen's notions of objectivity. *American Journal of sociology, 77(4)*, 660-679.
Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, *39*(2-3), 165-210.
Steensen S. (2017). Subjectivity as a Journalistic Ideal. *Putting a Face on it: Individual Exposure and Subjectivity in Journalism.* Cappelen Damm Akademisk, 25-47.

Opinion			
Mesure	**coef.**	**SE**	***p***
nb_on	101.3	7.32	< 0.001
pron_rel	54.34	4.49	< 0.001
nb_neg	49.51	5.31	< 0.001
nb_adj	22.85	1.62	< 0.001
lexique3_sentiment	16.34	5.78	0.005
nrc_sentiment	10.39	1.71	< 0.001
nb_exclam	9.35	0.76	< 0.001
nb_interrog	6.7	0.63	< 0.001
nb_pointvirg	6.22	0.9	< 0.001
nb_susp	3.90	0.79	< 0.001
nb_deuxpoints	3.74	0.48	< 0.001
blob_sent	2.43	0.59	< 0.001
cttr	1.85	0.04	< 0.001

Information			
Mesure	**coef.**	**SE**	***p***
nb_digits	79.26	3.36	< 0.001
pron_1	46.28	4.95	< 0.001
nb_vb	13.2	1.25	< 0.001
length_words	3.72	0.08	< 0.001
nb_citations	0.37	0.01	< 0.001

Most discriminating indicators

- For ***opinion***:

  Pronoun *on* (= *we/one*), relative pronouns, negations, adjectives, sentiment words, expressive punctuation...

- For ***information***:

  Use of numbers, first person pronouns, verbs, long words, citations.

Experiment n°1