Language & Communication 99 (2024) 129–140

Contents lists available at ScienceDirect

Language & Communication

journal homepage: www.elsevier.com/locate/langcom

Automated text classification of *opinion* vs. *news* French press articles. A comparison of transformer and feature-based approaches

Louis Escouflaire*, Antonin Descampe, Cédrick Fairon

UCLouvain, Institute for Language & Communication (IL&C), Place Cardinal Mercier 31 - box L3.03.02, 1348 - Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:

Keywords:
Subjectivity
Transformers
Feature-based model
Text classification
Discourse analysis
Explainability

ABSTRACT

This study explores Natural Language Processing (NLP) methods for distinguishing between press articles belonging to the journalistic genres of 'objective' *news* and 'subjective' *opinion*. Two classification models are compared: CamemBERT, a French transformer model fine-tuned for the task, and a machine learning model using 32 linguistic features. Trained on 8000 Belgian French articles, both models are evaluated on 1000 Canadian French articles. Results show CamemBERT's superiority but highlight potential for hybrid approaches and emphasizes the need for robust and transparent methods in NLP. The research contributes to understanding NLP's role in journalism by addressing challenges of point of view detection in press discourse.

1. Introduction

In the age of information overflow, identifying point of view in press discourse remains a challenge for both readers and researchers. Journalism serves as a vital conduit for information dissemination, yet the infiltration of personal perspectives and biases can blur the line between factual reporting and subjective interpretation containing point of view. Correctly distinguishing opinionated from non-opinionated journalistic texts is a crucial step towards understanding the dynamics of modern journalism. In this context, the need for automated methods for classification of press articles, as robust and as transparent as possible, is more important than ever.

In this study, we explore different Natural Language Processing (NLP) approaches used for automatically identifying a journalistic text as relying on the author's point of view or not. We specifically focus on a binary classification task referred to as "subjective text classification", which consists in predicting whether a given press article belongs to one of the two major categories of journalistic genres: the fact-oriented "objective" *news* genres or the "subjective" *opinion* genres, which overtly present the point of view of the author or the media (Schudson, 2001). We compare the performances of two contemporary classification models based on different approaches and discuss their relevance in the modern field of NLP, considering questions of sustainability and explainability. The first model is a state-of-the-art deep learning model based on the transformer architecture and pre-trained on French data, CamemBERT (Martin et al., 2019), which we fine-tune for our *news* vs.

* Corresponding author. UCLouvain - ILC, Place Cardinal Mercier 31/L3.03.11, 1348 Louvain-la-Neuve, Belgium. Tel.: +32 0 478 71 35 72.
E-mail addresses: louis.escouflaire@uclouvain.be (L. Escouflaire), antonin.descampe@uclouvain.be (A. Descampe), cedrick.fairon@uclouvain.be (C. Fairon).

https://doi.org/10.1016/j.langcom.2024.09.004
0271-5309/

opinion classification task. The second model is a more traditional machine learning model relying on thirty-two linguistic features identified as efficient indicators of point of view in the state-of-the-art on subjectivity in press discourse, and through feature selection experiments involving human annotation and model explainability.

We train both models on the same set of 8000 *news* and *opinion* articles published by four French-speaking Belgian media outlets and evaluate their classification accuracy on a set of 1000 articles published by four Canadian French (Quebec) media. Then, we take a closer look at the coefficients of the feature-based model and carry out a qualitative analysis of some typical errors made by the model. Our observations provide interesting details on the differences between the approaches and pave the multiple paths for bridging the gap between a lighter feature-based model and a more complex and resource-heavy transformer model.

First, we introduce some background on objectivity and subjectivity in journalism and we provide an overview of the state-of-the-art of subjective text classification in NLP, distinguishing between traditional lexicon-based and machine learning approaches, and more recent methods using deep learning. Then, we present our training and test corpus and introduce the two classification models used in this experiment. Section 5 is dedicated to a discussion of the results of the experiment followed by a comparison of the feature-based and transformer models for our *opinion* vs. *news* classification task.

2. Objectivity and subjectivity in journalism

Objectivity has long been a central point of discussion in journalism, while also considered one of its most essential values (Vos, 2012). It encompasses professional criteria such as neutrality, impartiality, factuality, or balance (Schudson, 2001). Although many journalists view objectivity as an ideal to aspire to, most acknowledge that complete journalistic objectivity is unattainable (Post, 2015; Ward, 2019). The inherent subjectivity in journalism is tied to the inevitable processes of selection and decision-making that occur throughout the editorial process, such as choosing stories, deciding their format, and prioritizing certain articles over others (Tong and Zuo, 2021). The presentation of facts is inevitably influenced by the journalist's personal interpretation, which is shaped by their perspective and experiences, making the pursuit of objectivity in news reporting particularly challenging (Muñoz-Torres, 2012). The presence of subjectivity in hybrid genres such as narrative journalism, investigative journalism, or chronicles, also renders the distinction between *opinion* and *news* genres more difficult (Wahl-Jorgensen and Schmidt, 2019). Additionally, objectivity currently faces significant challenges due to the rise of digitalization and social media, which have altered consumption habits (Meier et al., 2022). The introduction of new methods and tools, such as data journalism and artificial intelligence in newsrooms, further complicates the place of objectivity (Henestrosa et al., 2023). As a result, values such as truth, independence, and transparency have been proposed as new criteria for objectivity in modern journalism (McNair, 2017). Given the current decline in public trust in the media, it may be necessary to reassess and redefine the relevance of objectivity (Newman et al., 2021).

Journalists are trained to employ a variety of stylistic techniques to create the appearance of objectivity in their writing, aligning with the concept of the 'strategic ritual of objectivity' proposed by Tuchman (1972). This involves using specific strategies to minimize or mask the journalist's opinions within the article (Koren, 2004). Recommendations corresponding to these practices are often taught in journalism textbooks, enforced in newsrooms, or corrected by editors, and include for example the systematic citation of information sources, the use of impersonal language, neutral vocabulary, and the avoidance of figurative language (Charaudeau, 2006).

Traditionally, the pursuit of textual objectivity is confined to the *news* genres of journalism, such as press agency reports, and does not extend to *opinion* genres, i.e. editorials or columns (Grosse, 2001). While this distinction still holds in most print media nowadays, it may be more difficult for online readers lacking media literacy to differentiate between the two genres when clear labels are not present on social media or on some news websites. Therefore, automated tools for classification of subjective *opinion* and non-subjective *news* articles, at text-level, would be valuable. NLP approaches towards this task are presented in the next section.

3. Subjective text classification in NLP

The field of sentiment analysis encompasses several NLP tasks that involve identifying, analyzing, and extracting point of view, emotion, and opinion from textual data, as well as classifying texts based on the expressions of sentiment they contain (Ravi and Ravi, 2015). A subtask of sentiment analysis is subjective text classification, which involves determining whether a text, regardless of its genre (journalistic or not), presents objective or subjective information (Tang et al., 2009; Krüger et al., 2017), in other words whether it relies on the point of view of its author or of another source reported by the author. Automatically distinguishing subjective text from objective text can also be likened to detecting 'private states', i.e., portions of text that are influenced by the author's personal mental states, which are not objectively verifiable by an external observer (Montoyo et al., 2012). By eliminating objective text portions beforehand to focus on those identified as containing point of view, sentiment analysis models show better results, both in terms of precision and computational cost (Pang and Lee, 2008). Ravi and Ravi (2015) stress that automated classification of subjective texts is one of the most challenging sub-tasks in the field, far overshadowed by polarity analysis and opinion mining. It is also a particularly complex task, as summarized by Mihalcea et al. (2007):

> « The problem of distinguishing subjective versus objective instances has often proved to be more difficult than subsequent polarity classification, so improvements in subjectivity classification promise to positively impact sentiment classification. »

Janyce Wiebe was one of the first to seek to classify texts based on the identification of private states (Riloff and Wiebe, 2003; Wiebe et al., 2004; Wiebe and Riloff, 2005). In her work, she attempted to assess to what extent a sentence should be considered subjective based on the 'potentially subjective elements' (PSE) it contains. According to her, "objective sentences are sentences that do not contain any significant expression of subjectivity"[1] (Wiebe et al., 2004), and the general 'subjectivity degree' of a text can be measured as the ratio between the number of subjective sentences (which contain at least one PSE) and the total number of sentences in the text.

Like other tasks related to sentiment analysis, subjective text classification can be performed at different scales, corresponding to different textual levels: at the word, sentence, paragraph, or entire document level (Marchand, 2012). Classifying subjective words and phrases without the input of larger context often poses problems of ambiguity, while studying subjectivity at the paragraph or document level can be complicated due to the number of sources, opinions, and evaluated entities that may appear within the same text (Boullier and Lohard, 2012; Wankhade et al., 2022). Considering variables such as the presence of negation, irony, metaphorical expressions, or other forms of implicit language also constitutes a major challenge in subjective text classification (Montoyo et al., 2012).

Journalistic corpora are regularly exploited for subjective text classification, since most media outlets themselves indicate whether each article belongs to the journalistic genre of *news*, generally objective, or to that of *opinion*, generally subjective (Wiebe et al., 2004; Krüger et al., 2017). The automated classification of *news* and *opinion* articles (e.g., dispatches vs. editorials) is considered particularly complex because journalistic texts often involve, through quotations, multiple voices that do not always imply the subjectivity of their authors (Pang and Lee, 2008). In this paper, we focus on the task of classifying *opinion* vs. *news* articles as a way of classifying pieces of journalistic discourse containing or not containing some sort of point of view, even though the boundaries between the two journalistic genres may sometimes be relatively porous (Wahl-Jorgensen and Schmidt, 2019).

3.1. Lexicon-based and machine learning approaches

The most direct approach for subjective text classification consists in using a lexicon composed of predetermined subjective words: if a sufficient portion of the lemmatized words that constitute a text are present in the lexicon, the text can be considered subjective (Birjali et al., 2021). For English, some subjectivity lexicons include the MPQA Lexicon (Wilson et al., 2005), SentiWordNet (Baccianella et al., 2010), or NRC EmoLex (Mohammad and Turney, 2013).

Then, the approach which has long shown the best performance in sentiment analysis is the use of machine learning (ML) models. For subjective text classification, it involves training a ML model on a corpus of annotated texts that were previously classified as subjective or objective. During training, the model uses these annotations to learn to recognize texts belonging to each class, based on the patterns it identifies in the corpus. Once the model is trained, it is evaluated on a test corpus and can then be used to classify unseen texts and determine, in turn, whether these texts are more likely to be subjective or objective (Boullier and Lohard, 2012). The most popular ML models for text classification are decision trees, logistic regression, or support vector machines (Birjali et al., 2021). Training can be performed based on different types of features used to represent each text in the model, according to the preferred model type. Firstly, the simple absence or presence of a word in the text (0 or 1) can be used to represent the text in the form of a one-hot vector, which can then be used for n-gram classification. The frequencies of words can also be considered by applying the TF-IDF (*term frequency-inverse document frequency*) technique to measure the importance of words based on their relative frequency in the corpus (Ramos, 2003). Each text can also be reduced to a set of textual features that can be defined by theoretical bases or through preliminary selection steps or experiments, aiming to filter the best combination of features for the model. For example, features may include the frequency of specific morphosyntactic tags (e.g. grammatical class, tense, number, gender). Many types of textual features of subjectivity in news articles in various languages have been used and evaluated to train ML models. For example, Krüger et al. (2017) used a range of 28 different linguistic features to classify news (objective) and opinion (subjective) articles, acknowledging the predictive strength of some syntactic measures (e.g. distribution of verbs and adjectives), sentiment words, and lexicometric variables such as the presence of strong punctuation.

3.2. Deep learning and transformer-based approaches

Nowadays, feature-based machine learning approaches have largely been replaced by deep learning models, which involve the use of neural networks relying on several billion parameters and requiring substantial computational resources.

[1] It should be noted that this definition of subjectivity does not consider the truth value of a piece of text. Subjectivity, in this conceptual sense (Wiebe et al., 2004), concerns the presence or absence of point of view at a purely textual level, but not whether the text presents *true*, *false* or *incomplete* information.

Neural networks have particularly demonstrated their effectiveness for various sentiment analysis tasks, especially in English, including subjective text classification (Luan and Lin, 2019; Joseph et al., 2022).

The emergence of large language models revolutionized the NLP field. Models like BERT (Devlin et al., 2018) and GPT-3 (Brown et al., 2020), based on the *transformer* architecture, significantly advanced the state-of-the-art performance in various tasks and languages previously established by earlier deep learning models (Deepa and Tamilarasi, 2021). Transformer models consist of a series of layers through which the model's weights assigned to each element of the input text are adjusted. These weights are based on the attention mechanism (Vaswani et al., 2017), which allows a model to focus on the most important parts of the text rather than treating it uniformly. Attention is used in conjunction with self-attention, which gives the transformer the ability to consider all input tokens simultaneously and compute a weighted representation of each token based on all other tokens (Otter et al., 2020). Transformers are thus capable of representing complex syntactic and semantic relationships between different elements of the text (Clark et al., 2019). Tokens are represented in the form of contextualized embeddings, which offer the opportunity to capture the different meanings of the same word depending on its different contexts of occurrence (Wiedemann et al., 2019). Transformer-based language models are pre-trained on very large corpora of texts to learn the structure of the language and the corpus in question. Afterwards, they can be fine-tuned, meaning they can be retrained on a specific task (and thus on a more specific, smaller corpus) for which the model will be used. During fine-tuning, the model's lexical embeddings are transformed to represent the new features obtained from the new corpus. Google BERT (Bidirectional Encoder Representations from Transformers) and other similar language models are typically used in their pre-trained form, which can then be fine-tuned on another corpus.

Applied to sentiment analysis of English texts, BERT and its evolution RoBERTa have shown unprecedented results (Batra et al., 2021; Deepa and Tamilarasi, 2021). The French models CamemBERT and FlauBERT also been successfully applied to sentiment analysis tasks, particularly for polarity classification (Le et al., 2019; Essebbar et al., 2021), but have not yet been evaluated for subjective text classification.

Moreover, the high computational costs required by large language models, both in terms of computational time and resources (Cunha et al., 2021; Luccioni et al., 2024), is becoming more and more prominent in the NLP field (Anthony et al., 2020; Bannour et al., 2021). With environmental concerns on the rise in Western societies, the intensive use of transformer models may need to be reconsidered over the next decade. However, although these resource-heavy models may also be more complicated to integrate in practical applications requiring lighter and faster infrastructures, it is likely that they remain in the spotlight for some time. Therefore, our work aims to examine the trade-off between the two approaches (transformer-based and feature-based), relying on the classification task discussed in this paper as a use case.

4. Data

We gathered press articles written in French but published in two different areas of the world: the French-speaking region of Belgium (Brussels and Wallonia), and the French-speaking province of Canada (Quebec). Although they share the same language, Belgian French and Canadian French are two linguistic varieties which show some significant differences, especially in terms of lexicon. More importantly, the media ecosystems and the cultures of the two areas are very distinct, which may lead to interesting results when evaluating on Quebec-based articles models which were trained on Belgian data.

For the sake of reproducibility, the training and test corpora will be made available upon request.

4.1. Training corpus

We assembled a corpus of 40,000 press articles published online between 2014 and 2023 by 4 Belgian media outlets: *RTBF*, *Le Soir*, *La Libre Belgique* and *L'Avenir* (10,000 articles per media). The *RTBF* section of the corpus was used in (Bogaert et al., 2023) and obtained in collaboration with the media outlet, while all other articles were collected using web scraping tools and queries to the Europresse database. For each media outlet, we gathered 5000 articles corresponding to the *opinion* genres of journalism (e.g. op-eds, columns) and 5000 articles belonging to the *news* genres. These two classes correspond to the labels that were assigned to the articles by the media outlets that published them on their websites. Therefore, the training corpus contains a balanced distribution of 20,000 *opinion* articles and 20,000 *news* articles.

To build each sub-corpus corresponding to each media outlet, we used the same approach. We first collected and tokenized 5000 *opinion* articles, then applied latent Dirichlet allocation (LDA), an algorithm used for topic modeling, to identify the different topics addressed in these 5000 articles (Chauhan and Shah, 2021). The granularity of the modeling depends on the number of topics used to statistically explain the data, which we adjusted to minimize the perplexity of the LDA model. A low perplexity indicates that the model is capable of accurately predicting the content of a document based on its thematic structure, thus serving as a good indicator of the quality of the latter (Kobayashi, 2014). We limited the maximum number of identified topics to 10. Next, we built the *news* section of the sub-corpus by harvesting articles from sections corresponding to the different topics identified by LDA (e.g. politics, health), proportional to the representation of each topic in the opinion

articles. This approach allowed us to create a sub-corpus containing 5000 *opinion* articles and 5000 *news* articles, so that the addressed topics were distributed as evenly as possible in both parts of the sub-corpus.

To train and evaluate *opinion* vs. *news* classification models, each sub-corpus is divided into three sets: train (8000 articles), development (1000), and test (1,000), all balanced between the two classes.[2] Standard preprocessing (lemmatization and POS-tagging) was applied using Spacy (Honnibal and Montani, 2017).

4.2. Test corpus

We assembled a corpus of 1000 press articles published online between 2017 and 2023 by four media outlets based in Quebec: *La Presse*, *Le Devoir*, *Le Journal de Montréal*, and *Le Soleil* (250 articles per media). This corpus was collected and built using the same methods as the training corpus. The test corpus is balanced between the two classes, with 1000 *opinion* and 1000 *news* articles. The same preprocessing steps were applied to this corpus.

5. Classification models

We train and evaluate two different models for our subjective text classification task: a transformer model and a feature-based model.

5.1. Transformer model

We use CamemBERT (Martin et al., 2019), a version of BERT's adaptation RoBERTa (Liu et al., 2019) pre-trained on French data. CamemBERT's base version contains 110 million parameters and was pre-trained on a corpus consisting of Wikipedia and the OSCAR corpus (Ortiz Suárez et al., 2019). We fine-tuned CamemBERT on our *opinion* vs. *news* classification task using the train set of 32,000 Belgian press articles of our training corpus. As imposed by RoBERTa's inherent token limit, the fine-tuned model uses the 512 first tokens (following CamemBERT's WordPiece tokenizer) of a text to predict its journalistic genre. The model was fine-tuned for two epochs (with a learning rate of 2×10^{-5} and a batch size of 4).

5.2. Feature-based model

We built a feature-based model which uses a logistic regression classifier to predict their journalistic genre. The model relies on thirty-two textual features which were identified as effective predictors of the *opinion* or *news* genres in multiple experiments that were carried out previously using the *RTBF* portion of the training corpus. First, a feature selection experiment showed that several syntactic and lexical features (previously identified in the state-of-the-art on subjectivity in press discourse in English and other languages) could be reliably used to classify French-written *opinion* vs. *news* articles (Escouflaire, 2022). In terms of syntactic elements, the relative frequency in the text of first and second person pronouns determiners (De Cock, 2016), as well as relative pronouns, are good predictors of the *opinion* class. Regmi and Bal (2015) found that the ratio of some parts-of-speech in the article could help in the classification task between journalistic genres: *news* articles tend to contain more verbs, while *opinion* articles have more adjectives in proportion. Krüger et al. (2017) showed that the ratio of negations in the article could be used as a feature leaning towards *opinion*, while more digits indicated that the text belonged more likely to the *news* category. Those observations were confirmed by the feature selection experiment. It also showed the predictive power of including a sentiment lexicon in the classifier (Vis et al., 2012), using the French Lexique3 emotional valence lexicon (Gobin et al., 2017). At the textual level, a handful of features were evaluated as effective indicators for the *opinion* class: the presence of long words (Alhindi et al., 2020) and lexical complexity using type-token ratio (Krüger et al., 2017), and the use of question marks, semicolons and colons (Todirascu, 2019).

Other features of subjectivity in French press discourse were uncovered through an annotation experiment carried out on a sample of 150 excerpts of *opinion* and *news* RTBF articles (Escouflaire et al., 2024). A group of 36 master's students in journalism were asked to rank the overall subjectivity of the excerpts and to highlight the text spans that they considered to be "indicators of the article's subjectivity". Among the tokens that were most consistently highlighted as markers of *opinion* was the French pronoun *on*, commonly used in journalism as an example of "enunciative erasure" (Aouda, 2019). Expressive punctuation marks, including ellipses (" …") and exclamation marks ("!") were identified multiple times as subjective tokens, reflecting the results of Todirascu (2019). The annotators also highlighted many modal adverbs (e.g. "évidemment", "peut-être") and discourse markers (e.g. "en fait", "alors"), as in Küppers (2013), which are two features that were included in the regression model used in the present study.

Then, some more features were identified through an analysis using explainability methods to uncover patterns from explanations made by a CamemBERT model fine-tuned on 8000 *opinion* and *news* articles from the RTBF corpus (Bogaert et al., 2023). The Layer-wise Relevance Propagation (LRP) method developed by (Chefer et al., 2021) was applied to produce token-level explanations from predictions made by CamemBERT on 1000 articles. The LRP method explains predictions

[2] The full training corpus containing the articles of the four media outlets is therefore similarly divided into three sets: train (32,000 articles), development (4,000), and test (4,000).

made by deep learning models by evaluating the attention given to each token by the model. It measures a token's attention by backtracking layer by layer (from the output to the input) its contribution to the model's final prediction (Bach et al., 2015). Once 1000 token-level explanations of articles classified as either *opinion* or *news* by CamemBERT were obtained, the tokens which were given the most attention overall throughout both classes were examined. The tokens were ranked based on their average attention score. We qualitatively analyzed the 100 tokens with the highest average attention for each class.[3] Through an inductive process, we grouped tokens into linguistic categories corresponding to textual features, which led to the identification of several features for classifying *opinion* and *news* that were not found the two previous experiments: ratio of deictic and non-deictic temporal markers (two features), thinking verbs, citation verbs, passive verbs, average concreteness, imageability, and subjective frequency of the words in the text. In this paper, concreteness is measured using the lexicon produced by (Bonin et al., 2018), while imageability and subjective frequency are measured using the lexicons of (Desrochers and Thompson, 2009).

The logistic regression model is trained on the same 8000 articles used for fine-tuning the CamemBERT model. To replicate CamemBERT's token limitation, we only allow this feature-based model access to the first 512 tokens (using CamemBERT's tokenizer). We apply a grid search optimization to find the best combination of hyperparameters for the model. Except for global features such as type-token ratio and average word length, all features used in the model are measured as relative frequencies of selected tokens in the article. All thirty-two measures are normalized between 0 and 1.

6. Results and discussion

In this section, we compare the two models presented in section 5 by assessing their classification accuracy on the test set. While accuracy is a straightforward and widely accepted metric for evaluating classification models, it is crucial to recognize its limitations, especially in the context of tasks with potential real-world applications (Hossin and Sulaiman, 2015), like classifying journalistic texts. Accuracy, as a measurement, provides a quantitative assessment of how well a model performs on a given dataset. However, as an empirical indicator, accuracy on its own may not consider important dimensions such as fairness or transparency of a model (Palumbo et al., 2024), and it does not capture the nuances and complexities inherent in journalistic content, as elaborated in section 2. While we acknowledge other metrics such as computational cost and model explainability, which are also tackled in this section, we remain conscious that the simple measure of accuracy may overlook the subtleties of editorial choices and the multifaceted nature of journalistic discourse. We qualitatively analyze some errors in section 6.3 to better understand how our feature-based model lacks journalistic comprehension, but further research might be needed to explore potential combination of computational methods with insights from journalism studies and ethics.

6.1. Classification accuracy

First, we evaluate the accuracy of both classification models on the test set included in the training corpus. This test set contains 1000 articles published by four Belgian media, the same four media from which the model's training data was extracted. We refer to this test set as the *Belgium* test set. We also evaluate the models' classification accuracies on the test corpus containing 1000 articles published by four Quebec-based media. We refer to this external test set as the *Quebec* test set.

The accuracy results of the *opinion* vs. *news* classification task, presented in Table 1, show that the transformer model is 9.6% more accurate than the feature-based model on the 1000 articles of the *Belgium* test set. When evaluating both models on the *Quebec* corpus, which contains articles published in a different country and by other media outlets than those on which the models were trained, this gap tightens slightly: the transformer model is 7.6% more accurate than the feature-based model on the 1000 articles of the *Quebec* corpus. Looking more closely at the results obtained by the two models, it is interesting to notice that the drops in accuracies between the *Belgium* and *Quebec* evaluations are different: the feature-based model loses 3.9% accuracy between the two test sets, while the transformer model undergoes a drop in accuracy of 5.9% from the *Belgium* to the *Quebec* test set. While the fine-tuned transformer model is overall more efficient for this classification task, as could be expected, the feature-based model appears to be slightly more transferable to data external to the media on which

Table 1
Classification accuracy (%) of the transformer and feature-based models on the *Belgium* and *Quebec* test sets (N = number of articles in the test set).

	Transformer model	Feature-based model
Belgium (N = 1000)	93.9	84.3
Quebec (N = 1000)	88.0	80.4

[3] To neutralize the inherent variability of the explanations produced by the LRP method depending on the random seed on which the corpus was fine-tuned, we fine-tune 10 near-equivalent versions of CamemBERT on the same dataset of *RTBF* articles, based on 10 different random seeds. We selected only the tokens which were present in the top-100 tokens of at least 5 of the 10 near-equivalent CamemBERT models (Bogaert et al., 2023).

it was trained. This may be due to the fact that the transformer model is likely to overfit on elements present in the corpus on which it was fine-tuned, which may not be as relevant in different data (Gururangan et al., 2018), while the feature-based model relies on more universal theoretical features used for identifying point of view in press discourse and for classifying *opinion* and *news* articles.

These results also raise other important questions raised by the ubiquity of state-of-the-art transformer models, such as CamemBERT, in the current NLP field. First, large language models require heavy computational resources (Cunha et al., 2021) at several stages of the pipeline in which they are used (especially for pre-training and fine-tuning). With environmental concerns rising in different domains, including machine learning and NLP, the use of large resource-consuming models, which tend to become bigger and more complex, is being questioned (Anthony et al., 2020; Bannour et al., 2021). In this context, it is interesting to consider the trade-off between a more accurate but less 'resource-friendly' transformer model, and a traditional, less accurate but lighter, feature-based model.

In addition, an important difference between transformer and feature-based models lies in their potential for explainability. In the context of ML and artificial intelligence, explainability refers to the ability to understand and interpret the decisions made by a given model (Lyu et al., 2024). It entails transparency regarding how the model arrives at its predictions or classifications, allowing users to understand the underlying factors leading to the model's outputs. Explainability is crucial for various reasons, including ensuring the model's trustworthiness and identifying biases induced by the training data or the model's architecture, especially in contexts where the model's decisions may have important real-life implications, for example in the legal domain (Zini and Awad, 2023). We argue that identifying point of view in press discourse and classifying *opinion* and *news* articles are tasks for which the ability to explain the outputs made by automated models is important. It is essential however to note that explainability does not necessarily entail accuracy and reliability, especially in the field of journalism. Ethical considerations and the context in which decisions are made play a significant role in evaluating the trust that we can place in 'black-box' deep learning models (Loi and Ferrario, 2019). The challenge of ensuring that these models not only provide understandable outputs but also maintain high standards of accuracy and ethical integrity is still ongoing (Rai, 2020; Lyu et al., 2024).

The architectural complexity of multi-layered transformer models and their increasing number of parameters makes the understanding of their decision-making process very opaque (Danilevsky et al., 2020). Although various approaches towards interpreting outputs made by transformer models have been proposed alongside their rise in popularity, from model perturbation (Ribeiro et al., 2016) to attention probing through backpropagation (Bach et al., 2015), none of these methods can yet be considered sufficiently reliable (Lyu et al., 2024). In contrast, traditional feature-based models, while usually less accurate, can be considered explainable by nature, because their decision-making process relies on interpretable features.

In this context, we choose to take a deeper look at the faster and more transparent feature-based model, and at how it is possible to explain its decisions, while keeping in mind that the transformer model remains the most accurate model for this classification task.

6.2. Analysis of model coefficients

The thirty-two features used in the model and the regression coefficients associated to each of them after training are presented in Table 2. Features with a positive coefficient contribute to the *opinion* class, while features with a negative coefficient contribute to the *news* class.

Table 2
Regression coefficients (w) with standard error (S) for each of the 32 features in the logistic regression model trained on the training corpus.

opinion			*news*		
feature	w	S	feature	w	S
nb. of question marks (?)	0.737	0.004	avg. imageability of lexicon	−0.811	0.004
nb. of exclamation marks (!)	0.731	0.002	nb. of absolute time markers	−0.771	0.009
corrected type-token ratio	0.648	0.015	nb. of verbs of citation	−0.696	0.007
nb. of relative pronouns	0.437	0.011	nb. of quotation marks (")	−0.646	0.016
nb. of relative time markers	0.432	0.002	nb. of digits	−0.437	0.014
nb. of semicolons (;)	0.341	0.003	avg. length of words	−0.370	0.002
nb. of negation words	0.318	0.008	avg. subjective frequency of lexicon	−0.215	0.002
nb. of ellipses (…)	0.264	0.003	avg. concreteness of lexicon	−0.113	0.001
nb. of 2nd p. pronouns and determiners	0.246	0.005	nb. of words longer than 8 characters	−0.082	0.033
nb. of sentiment words (Lexique3)	0.224	0.007	nb. of verbs	−0.077	0.019
nb. of adjectives	0.214	0.019	nb. of 1st p. pronouns and determiners	−0.048	0.012
nb. of colons (:)	0.184	0.004	nb. of passive verbal forms	−0.036	0.001
nb. of verbs of thought	0.182	0.007	nb. of complex conjunctions	−0.024	0.003
nb. of modal adverbs	0.157	0.001			
nb. of discourse markers	0.131	0.007			
nb. of commas (,)	0.091	0.017			
nb. of emotion words (NRC)	0.065	0.035			
nb. of hapaxes	0.039	0.046			
nb. of indefinite pronoun *on*	0.001	0.005			

We observe that many features which push the model towards predicting the *opinion* class for an article are related to quite simple lexical and syntactical measures: strong markers of *opinion* ($w > 0.2$) include several expressive punctuation marks ("?", "!", "; ", " …"), as well as adjectives and negation words which can be used as means of expressing sentiment in discourse. Deictics, in particular relative time markers (e.g. *yesterday*, *now*) and second person pronouns and determiners, also appear to play an important part. We also find that indicators of lexical complexity, such as the type-token ratio and the presence of relative pronouns (which tend to increase the level of embedding of a sentence), are used by the model as *opinion* features. Interestingly, another indicator of lexical complexity, the average length of words, is considered by the model as a feature predicting the *news* class. The longer the words in an article, the more likely the feature-based model is to classify it as belonging to the *news* genre. Regarding other *news* features with a substantially low negative coefficient ($w < -0.2$), we find measures of the lexicon's imageability (the ease of generating a mental image for a word) and familiarity (frequency of exposure to a word in daily life, also called "subjective frequency"), and features related to the article's reliance on facts, such as verbs of citation, quotation marks, digits, and absolute time markers (e.g. *Tuesday*, *November*). Overall, the features on which the model appears to rely most include features that were identified throughout the experiments previously carried out on *opinion* vs. *news* classification in French press discourse (presented in Section 3.2.2): features identified in the state-of-the-art and by feature selection (type-token ratio, average word length), features identified through manual annotation by 36 participants (ellipses, quotation marks), and features derived from token-level explanations extracted from CamemBERT predictions (imageability, time markers).

6.3. Error analysis

In this section, we take a closer look at some qualitative examples. In an attempt to better understand what makes the feature-based model less accurate than the transformer model, we first focus on articles from both test sets that were misclassified by the feature-based model but for which the transformer predicted the correct class.

Example [1] is an excerpt of a Belgian article labeled as an *opinion* piece in the media by which it was published. Our feature-based model predicted it as belonging to the *news* class, which is due to the high frequency in the article of tokens related to features with a negative coefficient in the model, among which absolute time markers (e.g. *August*, *September*, *December*) and digits. This excerpt also contains quotation marks, which are used by the model as strong indicators of the *news* class. However, in the article, they are not used by the journalist to report direct speech, as they are habitually used in informative press articles, but rather to express distance from a particular word or expression (this use is sometimes referred to as "scare quotes"). While the feature-based model does not have access to the contextual information needed to distinguish between these two usages of quotation marks, it is likely that the transformer model is able to capture this kind of nuance.

[1] Si une rentrée des classes au mois d'août en Belgique francophone est évidemment un événement, ce n'est en rien une révolution. C'est dans l'ordre des choses. Avec une coupure estivale de 7 semaines, nous sommes dans la moyenne des pays européens. Si vos enfants trouvent que c'est "dégueulasse" de reprendre le cartable avant le 1er septembre, vous pouvez leur expliquer que chez nos voisins allemands ou anglais, les élèves ont droit à 6 semaines de vacances et qu'au Japon, c'est 5 semaines en août (en plein milieu de l'année scolaire), 2 semaines en décembre et 3 semaines en mars, avant la rentrée. Soit 10 semaines seulement pour toute l'année.

While the start of the school year in August in French-speaking Belgium is obviously an event, it is by no means a revolution. It's just the way things are. With a 7-week summer break, we're in line with the European average. If your children think it's "so unfair" to go back to school before the 1st *of September, you can explain to them that in Germany and England pupils are entitled to 6 weeks' holiday, and that in Japan it's 5 weeks in August (in the middle of the school year), 2 weeks in December and 3 weeks in March before the new school year starts. That's only 10 weeks in the whole year.*

(example 1: excerpt of an *opinion* article from the *Belgium* test set)

The excerpt presented in example [2] is one of the articles from the *Quebec* test set with the highest average imageability. It also contains a lot of words with a high subjective frequency. Many tokens in the excerpt are indeed words which represent everyday life items (e.g. *snow*, *street*, *truck*). These two features are among the strongest markers of *news* in the feature-based model. In this article, they appear to push the model towards the wrong prediction (*news*), outdoing the high frequency in the text of exclamation marks (which are *opinion* features). On the contrary, the transformer model, because it does not rely on a limited number of token-based features, is able to correctly evaluate the importance of any token towards a given class, depending on the context of the article in which it is used. Transformer models such as CamemBERT use millions of parameters which allow them to apprehend textual dimensions beyond the token level, such as long-term dependencies, which our simpler feature-based model is not able to manage.

[2] Avec le gel, les grilles de rues sont remplies de neige. Résultat: cinq charrues sont nécessaires pour déneiger une petite rue. Donc, si une seule charrue passe des deux côtés et ramène la neige au centre et par la suite le souffleur et les camions ramassent le tout, il me semble que ce serait plus efficace, mieux nettoyé et prendrait moins de temps. Si quelqu'un d' autre propose quelque chose de meilleur, tant mieux ! Alors, bon déneigement lors de la prochaine tempête !

With the frost, the street grates are filled with snow. As a result, five ploughs are needed to clear a small street of snow. So, if a single plough went over both sides and brought the snow to the center and then the blower and lorries picked it all up, it seems to me that it would be more efficient, cleaner and would take less time. If someone else comes up with something better, so much the better! So, good luck clearing the snow next time there's a storm!

(example 1: excerpt of an *opinion* article from the *Quebec* test set)

Then, we analyze a *news* article which was misclassified as an *opinion* by the feature-based model. In example [3], the elements which make the model lean towards *opinion* are the high number of sentiment words (e.g. *preferred*, *assaults*, *elite*) and adjectives (e.g. *sexual*, *minor*, *psychological*). However, while several of these words and adjectives can be used to express some kind of point of view, they do not systematically carry an evaluative value (Wiebe et al., 2004). This example clearly shows that the feature-based model, which represents the text as a simple bag-of-words reduced to a vector of features, lacks the capacity to disambiguate between the subjective and the non-subjective uses of these words. On the contrary, the transformer model represents tokens under the form of contextualized embeddings, allowing it to properly determine whether an ambiguous token is used to express point of view or not (Wiedemann et al., 2019).

[3] Trois des victimes d'agressions sexuelles de l'ex-entraîneur de ski Bertrand Charest poursuivent en dommages son employeur, Canada Alpin, parce qu'elles soutiennent que la fédération sportive avait amplement d'informations pour savoir qu'il y avait des problèmes, mais a préféré fermer les yeux. Les trois femmes sont Geneviève Simard, Gail Kelly et Anna Prchal, d'anciennes skieuses d'élite. Elles étaient mineures à l'époque des agressions alléguées. Elles réclament de la fédération 300 000 $ chacune en dommages pour les abus psychologiques, physiques et sexuels qu'elles ont subis.

Three of the victims of sexual assault by former ski coach Bertrand Charest are suing his employer, Canada Alpin, for damages because they claim that the sports federation had ample information to know that there were problems but preferred to turn a blind eye. The three women are Geneviève Simard, Gail Kelly and Anna Prchal, former elite skiers. They were minor at the time of the alleged assaults. They are claiming $300,000 each in damages from the federation for the psychological, physical and sexual abuse they suffered.

(example 3: excerpt of a *news* article from the *Quebec* test set)

These qualitative observations lead to interesting perspectives towards potential improvements for our feature-based model. For example, replacing its bag-of-words representation with a bigram or trigram representation may improve its disambiguation capacity. Building press-specific lexicons for sentiment words or imageability (among other lexicon-based features) could also increase the model's accuracy for this *opinion* vs. *news* classification task. Another area for improvement concerns the handling of quotes: considering or neutralizing the content of the text inside or outside quotations would likely be beneficial to the feature-based model. These prospects will be explored in further research.

Finally, we consider a fourth text: an excerpt of an *opinion* article which was incorrectly classified as a *news* article by both the feature-based and the transformer models. While we are not able to explain the prediction process of the transformer model faithfully nor transparently (Lyu et al., 2024), we may look at the features that influenced the feature-based model towards predicting the same class for example [4]. We find that the article contains a high frequency of question marks and interrogation marks, which are the two strongest indicators of *opinion* in our regression model, and that it also shows a high frequency of adjectives. This last example is interesting because it shows that the inherent explainability of a feature-based model may offer (to a certain extent) some insights into the reasons behind errors made by a transformer model: here, it is likely that CamemBERT predicted example [4] as an *opinion* in part because it was influenced by the same elements that led the feature-based model to the wrong prediction (such as the presence of expressive punctuation). In addition, the 'wrong' consensus between the two models on this article's class also raises an important question regarding the labels associated to the articles in our corpus: can a media be trusted regarding the journalistic genres it attributes to its own articles?

[4] Dans quelles conditions peut-on encore franchir une frontière de la Belgique ? Dans un arrêté ministériel paru le 3 avril, le Conseil National de Sécurité a défini de manière stricte les trois raisons qui restent valables: si l'on travaille de l'autre côté de la frontière (ou que l'on doit ponctuellement le faire dans le cadre de ses activités professionnelles), pour des raisons médicales ou encore pour un regroupement familial ! Et puis c'est tout, plus question, même si on habite à quelques kilomètres à peine de la frontière, d'aller faire ses courses ou le plein de carburant !

Under what conditions is it still possible to cross a border into Belgium? In a ministerial decree published on 3 April, the National Security Council strictly defined the three reasons that remain valid: if you work on the other side of the border (or have to do so from time to time as part of your professional activity), for medical reasons or for family reunification! And that's all there is to it, even if you live just a few kilometers from the border, so you can't go shopping or fill up with petrol!

(example 4: excerpt of a *news* article from the *Belgium* test set)

7. Conclusions

Some inherent limitations to our study should be taken into account when considering our results. First, the assumptions made in Section 5.3 about the prediction mechanisms of the transformer model are partly hypothetical and should be

thoroughly tested using explainability methods, although it is still unclear whether reliable explanation approaches for transformer models have yet been introduced (Lyu et al., 2024). Then, the *opinion* and *news* labels on which the models were trained and evaluated are labels which were assigned to the articles exclusively by the media in which they were published. This approach implies that these labels indeed match with the actual journalistic genres of the articles, which is an assumption that should be tested in a different study. A large-scale annotation experiment with trained annotators (journalists and researchers) is underway to verify whether the genre labels assigned by the media outlets correspond to human interpretations of the presence of point of view in these articles.

Our study showed that, as could be expected, a transformer-based large language model (CamemBERT) performs significantly better than a traditional logistic regression based on textual features for automatically classifying French press articles belonging to the *opinion* or *news* genres of journalism. However, we observed that the transformer model is very resource-costly and suffers from a crucial lack of explainability, while the feature-based model is fast and relies on thirty-two linguistic features, which makes its decisions completely transparent. Therefore, we decided to examine the features of the regression model more closely and to qualitatively analyze a sample of articles which it classified into the wrong class. These investigations helped us to better understand the gap in accuracy between the two models and to identify paths for improvement for the feature-based model. Further research could focus on enhancing this model with mechanisms inspired by the transformer model's representations and inner workings, while keeping the computational cost of the model low and maintaining it as explainable as possible, assuming a technically inevitable but necessary drop in accuracy. Building hybrid models that balance accuracy with explainability, could be viable alternatives which would help address the ethical concerns associated with the deployment of transformer models in journalistic contexts, ensuring that the models used are transparent, reliable and aligned with journalistic standards.

Declaration of competing interest

None.

CRediT authorship contribution statement

Louis Escouflaire: Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation. **Antonin Descampe:** Project administration. **Cédrick Fairon:** Project administration.

Funding

This research is part of a PhD thesis project entirely funded by the FNRS (Belgian National Fund for Scientific Research).

Declaration of competing interest

None.

References

Alhindi, T., Muresan, S., Preotiuc-Pietro, D., 2020. Fact vs. Opinion: the role of argumentation features in news classification. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 6139–6149.
Anthony, L.F.W., Kanding, B., Selvan, R., 2020. Carbontracker: tracking and predicting the carbon footprint of training deep learning models. ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems.
Baccianella, S., Esuli, A., Sebastiani, F., 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. LREC 10 (No. 2010), 2200–2204.
Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10 (7), e0130140.
Bannour, N., Ghannay, S., Névéol, A., Ligozat, A.L., 2021. Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools. In: Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing, pp. 11–21.
Batra, H., Punn, N.S., Sonbhadra, S.K., Agarwal, S., 2021. BERT-based sentiment analysis: a software engineering perspective. In: Database and Expert Systems Applications: 32nd International Conference (DEXA 2021), Proceedings Part I. Springer International Publishing, pp. 38–148, 32.
Birjali, M., Kasri, M., Beni-Hssane, A., 2021. A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl. Base Syst. 226, 107134.
Bogaert, J., Escouflaire, L., de Marneffe, M.C., Descampe, A., Standaert, F.X., Fairon, C., 2023. TIPECS: a corpus cleaning method using machine learning and qualitative analysis. International Conference on Corpus Linguistics (JLC).
Bogaert, J., Escouflaire, L., de Marneffe, M.C., Descampe, A., Standaert, F.X., Fairon, C., 2023. Sensibilité des explications à l'aléa des grands modèles de langage : le cas de la classification de textes journalistiques. Trait. Autom. Des. Langues 64 (3).
Boullier, D., Lohard, A., 2012. Chapitre 5. Détecter les tonalités : opinion mining et sentiment analysis. Opinion mining et Sentiment analysis : Méthodes et outils. OpenEdition Press, Marseille.
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.
Charaudeau, P., 2006. Discours journalistique et positionnements énonciatifs. Frontières et dérives. Semen. Revue de sémio-linguistique des textes et discours (22).
Chauhan, U., Shah, A., 2021. Topic modeling using latent Dirichlet allocation: a survey. ACM Comput. Surv. 54 (7), 1–35.

Chefer, H., Gur, S., Wolf, L., 2021. Transformer interpretability beyond attention visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 782–791.
Clark, K., Khandelwal, U., Levy, O., Manning, C.D., 2019. What does BERT look at? an analysis of BERT's attention. arXiv preprint arXiv:1906.04341.
Cunha, W., Mangaravite, V., Gomes, C., Canuto, S.D., Resende, E., Nascimento, C., Viegas, F., França, C., Martins, W.S., Almeida, J.M., Rosa, T., Rocha, L., Gonçalves, M.A., 2021. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: a comprehensive comparative study. Inf. Process. Manag. 58 (3), 102481.
Danilevsky, M., Kun, Q., Ranit, A., Yannis, K., Ban, K., Prithviraj, S., 2020. A survey of the state of explainable AI for natural language processing. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. Association for Computational Linguistics, Suzhou, China, pp. 447–459.
De Cock, B., 2016. Register, genre and referential ambiguity of personal pronouns: a cross-linguistic analysis. Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA) 26 (3), 361–378.
Deepa, D., Tamilarasi, A., 2021. Bidirectional encoder representations from transformers (BERT) language model for sentiment analysis task. Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12 (7), 1708–1721.
Desrochers, A., Thompson, G.L., 2009. Subjective frequency and imageability ratings for 3,600 French nouns. Behav. Res. Methods 41 (2), 546–557.
Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
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. Volume 2: 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL), 69–82.
Escouflaire, L., Descampe, A., Fairon, C., 2024. Unveiling Subjectivity in Press Discourse: a Statistical and Qualitative Study of Manually Annotated Articles, p. 34. *Discourse*.
Essebbar, A., Kane, B., Guinaudeau, O., Chiesa, V., Quénel, I., Chau, S., 2021. Aspect based sentiment analysis using French pre-trained models. Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) 1, 519–525.
Gobin, P., Camblats, A.M., Faurous, W., Mathey, S., 2017. Une base de l'émotionalité (valence, arousal, catégories) de 1286 mots français selon l'âge (EMA). Eur. Rev. Appl. Psychol. 67 (1), 25–42.
Grosse, E.U., 2001. Évolution et typologie des genres journalistiques. Essai d'une vue d'ensemble. Semen. Revue de sémio-linguistique des textes et discours (13).
Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A., 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.
Henestrosa, A.L., Greving, H., Kimmerle, J., 2023. Automated journalism: the effects of AI authorship and evaluative information on the perception of a science journalism article. Comput. Hum. Behav. 138, 107445.
Honnibal, M., Montani, I., 2017. spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
Hossin, M., Sulaiman, M.N., 2015. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process 5 (2), 1.
Joseph, J., Vineetha, S., Sobhana, N.V., 2022. A survey on deep learning based sentiment analysis. Mater. Today: Proc. 58, 456–460.
Kobayashi, H., 2014. Perplexity on reduced corpora. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 797–806.
Koren, R., 2004. Argumentation, enjeux et pratique de l'«engagement neutre»: le cas de l'écriture de presse. Semen. Revue de sémio-linguistique des textes et discours, 17.
Krüger, K.R., Lukowiak, A., Sonntag, J., Warzecha, S., Stede, M., 2017. Classifying news versus opinions in newspapers: linguistic features for domain independence. Nat. Lang. Eng. 23 (5), 687–707.
Küppers, A., 2013. Private State in Public Media: Potential Subjective Elements in French-speaking (Online) News. Doctoral dissertation, UCLouvain.
Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D., 2019. FlauBERT: unsupervised language model pre-training for French. arXiv:1912.05372.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
Luan, Y., Lin, S., 2019. Research on text classification based on CNN and LSTM. In: 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), pp. 352–355.
Luccioni, A.S., Viguier, S., Ligozat, A.-L., 2024. Estimating the carbon footprint of BLOOM, a 176B parameter language model. J. Mach. Learn. Res. 24 (1), 253: 11990-253:12004.
Lyu, Q., Apidianaki, M., Callison-Burch, C., 2024. Towards faithful model explanation in NLP: a survey. Comput. Ling. 70.
Marchand, M., 2012. État de l'art: L'influence du domaine sur la classification de l'opinion. Actes de la Conférence JEP-TALN-RECITAL 2012, 177–190.
Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B., 2019. CamemBERT: a tasty French language model. arXiv:1911 03894.
McNair, B., 2017. After objectivity? Schudson's sociology of journalism in the era of post-factuality. Journal. Stud. 18 (10), 1318–1333.
Meier, K., Schützeneder, J., García Avilés, J.A., Valero-Pastor, J.M., Kaltenbrunner, A., Lugschitz, R., Porlezza, C., Ferri, G., Wyss, V., Saner, M., 2022. Examining the most relevant journalism innovations: a comparative analysis of five European countries from 2010 to 2020. Journalism and media 3 (4), 698–714.
Mihalcea, R., Banea, C., Wiebe, J., 2007. Learning multilingual subjective language via cross-lingual projections. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 976–983.
Mohammad, S.M., Turney, P.D., 2013. Crowdsourcing a word–emotion association lexicon. Comput. Intell. 29 (3), 436–465.
Montoyo, A., Martínez-Barco, P., Balahur, A., 2012. Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decis. Support Syst. 53 (4), 675–679.
Muñoz-Torres, J.R., 2012. Truth and objectivity in journalism: anatomy of an endless misunderstanding. Journal. Stud. 13 (4), 566–582.
Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C.T., Nielsen, R.K., 2021. Reuters Institute Digital News Report 2021. Reuters Institute for the Study of Journalism. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3873260. (Accessed 21 August 2024).
Ortiz Suárez, P.J., Sagot, B., Romary, L., 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In: 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7). Leibniz-Institut für Deutsche Sprache.
Otter, D.W., Medina, J.R., Kalita, J.K., 2020. A survey of the usages of deep learning for natural language processing. IEEE Transact. Neural Networks Learn. Syst. 32 (2), 604–624.
Palumbo, G., Carneiro, D., Alves, V., 2024. Objective metrics for ethical AI: a systematic literature review. International Journal of Data Science and Analytics, 1–21.
Pang, B., Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and Trends in information retrieval 2 (1–2), 1–135.
Post, S., 2015. Scientific objectivity in journalism? How journalists and academics define objectivity, assess its attainability, and rate its desirability. Journalism 16 (6), 730–749.
Rai, A., 2020. Explainable AI: from black box to glass box. J. Acad. Market. Sci. 48, 137–141.
Ramos, J., 2003. Using TF-IDF to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning 242 (No. 1), 29–48.
Ravi, K., Ravi, V., 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl. Base Syst. 89, 14–46.

Regmi, S., Bal, B.K., 2015. What make facts stand out from opinions? Distinguishing facts from opinions in news media. Creativity in Intelligent, Technologies and Data Science, pp. 655–662.
Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.
Riloff, E., Wiebe, J., 2003. Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, pp. 105–112.
Schudson, M., 2001. The objectivity norm in American journalism. Journalism 2 (2), 149–170.
Tang, H., Tan, S., Cheng, X., 2009. A survey on sentiment detection of reviews. Expert Syst. Appl. 36 (7), 10760–10773.
Todirascu, A., 2019. Genre et classification automatique en TAL: le cas de genres journalistiques. Linx. Revue des linguistes de l'université Paris X Nanterre 78.
Tong, J., Zuo, L., 2021. The inapplicability of objectivity: understanding the work of data journalism. Journal. Pract. 15 (2), 153–169.
Tuchman, G., 1972. Objectivity as strategic ritual: an examination of newsmen's notions of objectivity. Am. J. Sociol. 77 (4), 660–679.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Vis, K., Sanders, J., Spooren, W., 2012. Diachronic changes in subjectivity and stance – a corpus linguistic study of Dutch news texts. Discourse, Context and Media 1 (2–3), 95–102.
Vos, T.P., 2012. 'Homo journalisticus': journalism education's role in articulating the objectivity norm. Journalism 13 (4), 435–449.
Wahl-Jorgensen, K., Schmidt, T.R., 2019. News and storytelling. In: The Handbook of Journalism Studies. Routledge, pp. 261–276.
Wankhade, M., Rao, A.C.S., Kulkarni, C., 2022. A survey on sentiment analysis methods, applications, and challenges. Artif. Intell. Rev. 55 (7), 5731–5780.
Ward, S.J., 2019. Journalism ethics. In: The Handbook of Journalism Studies. Routledge, pp. 307–323.
Wiebe, J., Wilson, T., Bruce, R., Bell, M., Martin, M., 2004. Learning subjective language. Comput. Ling. 30 (3), 277–308.
Wiebe, J., Riloff, E., 2005. Creating subjective and objective sentence classifiers from unannotated texts. In: *Computational Linguistics and Intelligent Text Processing: 6th International Conference*, Mexico City, Mexico. Springer Berlin Heidelberg, pp. 486–497.
Wiedemann, G., Remus, S., Chawla, A., Biemann, C., 2019. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. arXiv preprint arXiv:1909.10430.
Wilson, T., Wiebe, J., Hoffmann, P., 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pp. 347–354.
Zini, J.E., Awad, M., 2023. On the explainability of Natural Language Processing deep models. ACM Comput. Surv. 55 (5), 103.