Explanation Sensitivity to the Training Randomness in Text Classification Transformers

Jérémie Bogaert

École polytechnique de Louvain
Université catholique de Louvain
Louvain-la-Neuve
Belgium

Thesis Committee:

Pr. Laurent Francis	*UCLouvain*
Pr. Cédrick Fairon	*UCLouvain*
Pr. Benoît Frénay	*UNamur*
Dr. Loic Masure	*CNRS*
Pr. Antonin Descampe	*UCLouvain*
Pr. François-Xavier Standaert	*UCLouvain*

November 2025

Acknowledgments

Before starting this thesis, I would first like to address a word to all the individuals and entities without whom it would not have existed. I would like to thank:

The TRusted Artificial Intelligence Lab (TRAIL) and in particular ARIAC for financing this thesis.

My supervisors, Pr. François-Xavier Standaert & Antonin Descampe for their invaluable insights and dedicated supervision.

My teammate (and often co-author) now Dr., Louis Escouflaire for his productivity, linguistic knowledge and his very appreciated serenity when I was freaking out.

My colleagues and members of the Crypto Group and EJL for creating such a peaceful environment, leading to both insightful discussions and very enjoyable coffee breaks.

My family & friends, for being in my life and for their support and kindness about the countless meeting skips, rescheduling and canceling due to my terrible work schedule.

And last but certainly not least, my soon to be fiance, Mathilde, for her love and belief in me throughout the entirety of the thesis. She gave me the strength to continue when I was about to stop, and inspires me to be a better person everyday.

In general, I dedicate this work to all the people that I cannot extensively list but who's presence helped me going through this difficult but enriching process that is a Ph.D. thesis.

Introduction

Introducing a work that lasted four entire years is not an easy task, in particular when doing research in a field changing as fast as Natural Language Processing. When starting, the initial title of this thesis was *IA-SPAN: Improving accountability and security of possibly automated news*. But this thesis, as its author and the domain it tackles, evolved during this period. To grasp how its final title changed to become *Explanation Sensitivity to the Training Randomness in Text Classification Transformers*, one first has to walk through the different steps that were followed, going from one research question to another.

To first provide a bit of context, the initial main motivation was that the spreading of misinformation, disinformation and fake news on social media platforms had been (and still is) a topic of increasing interest (Lazer et al. 2018). Back in 2020, advances in Natural Language Processing (NLP) and Natural Language Generation (NLG) based on neural machine learning already made it easier to generate massive amounts of seemingly consistent texts that subsequently contributed to increase information disorder. As a natural reaction, research topics focusing on detecting misleading content (sometimes under the term of "neural fake news detection") also rose in popularity.

This general research topic is composed of various approaches which can lead to different challenges depending on the definition of the task. See (Conroy et al. 2015; Sharma et al. 2019; X. Zhou and Zafarani 2020) for surveys and (Hassan et al. 2017; Ruchansky et al. 2017; Shu et al. 2017; Zellers et al. 2019) for a few technical case studies. For example, if the goal is to gauge the veracity of a news item, one faces the (non-technical) challenge of defining what a fake news is (Tandoc et al. 2018). Besides, many different elements can be taken into account to detect whether news are genuine or fake: the source of the information, people or platforms relaying them, the author's

reputation, a fact-checking process confronting the content with an external knowledge base, the style and the lexical field found in the text, etc. Among these elements, some are subjective (e.g., the author's or publishing sites' reputation) and most are based on contextual information not found in the news content itself. As a result, identifying labeled data sets for supervised machine learning in the context of fake news detection is inherently error-prone, which potentially implies label noise (Frénay and Verleysen 2014). In order to circumvent these difficulties, fake news detection is sometimes restricted to the (better defined) problem of distinguishing machine-generated news from human-generated ones (Zellers et al. 2019). Such a step back is interesting since it provides a basis for the detection of neural potential fake news, or more generally for the detection of mass disinformation relying on automatic text generation. The initial idea behind this thesis was to focus on this line of work. The first goal was to enhance the models' detection capabilities. The second one was to explore the possibility of generating better "neural news" (fake or not) in order to try to discriminate between "neural fake news" and "neural non-fake news" afterwards. Finally, since no good French text generator was available at the time, we wanted to focus on this language in particular.

Our first idea was to study the news generated by a language generation model called `Grover` (Zellers et al. 2019). Even if it was miles away of the currently world spread ChatGPT (which only appeared 2 years after the start of this work) in terms of generation performances, both these models share the same architecture (i.e., the transformer architecture). `Grover` (Zellers et al. 2019) was considered to be among the best tools to generate news at that time. We observed that news in general, and fake news in particular, are usually targeting recent and fast evolving topics. We therefore built an experiment where classifiers are required to detect whether submitted news are human- vs. machine-generated, for texts that were more and more remote from `Grover`'s training database. For this purpose, we considered different classification methods: logistic regression (McCullagh and Nelder 1989), support vector machines (Cortes and Vapnik 1995), naive Bayes (see Zhang 2004 for motivations to use this model), and long short-term memory networks (Hochreiter and Schmidhuber 1997). We also evaluated a manual detection experiment performed over four weeks, with 30 participants, each of them asked to evaluate 10 human- or machine-generated pieces of news per week. Unsurprisingly, our results confirmed that domain adaptation[1] was challenging

[1]Domain adaptation refers to the model being trained on a different data distribution than

for language (as for other) models (Xu et al. 2020): texts generated by `Grover` for unseen topics were identified as machine-generated both by simple automatic tools and by manual detection. These results thus rose the question of how fast such language generation models could be adapted to newsfeeds, e.g., thanks to emerging plug-and-play language models (Dathathri et al. 2020), which was another work direction that we considered. We then qualitatively investigated the decisions made by the readers, and the reasons for which they labeled news as human- or machine-generated in our manual-detection experiment. To do so, we leveraged the open answers provided by the annotators to explain their label choice (machine- or human-generated) and grouped them in four categories (syntax, semantic, background, and other) (see Appendix for more informations).

We gauged that these justifications were a nice starting point in order to both try to detect generated texts more easily and to strengthen existing language models capabilities. However, at this point, one of our concerns was that we had no way to know why the models were predicting a given label. It means that it would have been possible for the models to already leverage the same clues as the human readers without us noticing it. We therefore quite naturally turned to explainable AI (XAI) methods in order to explain our classifiers in the first place.

To explore these methods, as it was a very broad and new topic, we decided to have a research partnership with Louis Escouflaire, a Ph.D. student of Antonin Descampe (one of my supervisors) doing research about the detection of subjectivity in French press articles. This scientific collaboration turned out to be fruitful for our respective interests. It allowed us to explore various tools and phenomenons, which was my goal, on a specific subjectivity detection use case, which was his goal.

To run our studies, we needed a dataset composed of French news and opinions. For this purpose, we built the RTBF and InfOpinion datasets, that we used during all our next experiments. The RTBF dataset is a large diachronic corpus of 767 204 Belgian French press articles published between 2008 and 2021 by the Belgian public service media RTBF. The InfOpinion dataset is a subset of this corpus, that was obtained by selecting and cleaning some press articles with extra steps. It contains 10 000 news that were specifically labeled

---

the one it has to infer on

as pieces of information or opinion by their authors. This dataset was used to investigate to what extent explainability techniques were able to give insights about model behaviors.

The first thing we wanted to explore was the capacity of the still quite recent transformer architecture (Vaswani et al. 2017) to perform subjectivity detection. This architecture already had been of interest for the whole NLP community. Models like BERT (Devlin et al. 2019) and its variants such as RoBERTa (Y. Liu et al. 2019) or AlBERT (Lan et al. 2020) achieved state-of-the-art results on numerous classification tasks and datasets. It was however unclear to us what was happening inside these models during their training phase. More specifically, these models are typically first pre-trained on big amounts of data in order to learn how to transform texts into embeddings, and then adapted to an end task. To do so, Devlin et al. (2019) presented two different approaches: the **fine-tuning** approach, *where a simple classification layer is added to the pre-trained model and all parameters are jointly fine-tuned on a down stream task*, and the feature-based approach, often referred to as the **frozen** one, *where fixed features are extracted from the pre-trained model and fed to another model*. While the former is usually preferred, and may marginally improve the model performances (Yichu Zhou and Srikumar 2022), both models reach similar results on many classification tasks (Peters et al. 2019; N. F. Liu et al. 2019). It therefore rose the question of the fine-tuning impact. To evaluate it, we assessed different properties of fine-tuned vs frozen models based on two French text classification tasks (including subjectivity detection) using CamemBERT (Martin et al. 2020). We first observed quantitatively that, for both tasks, both approaches reached similar performances. We visualized the impact of fine-tuning on the data representation through nonlinear dimension reduction methods (de Bodt et al. 2020), observing that the embedding structure of the fine-tuned models led to much more separated clusters and concentrated information across fewer dimensions.

Finally, we wanted to see whether both of these training methods would produce similar explanations. For our explanation method, as the model was not intrinsically interpretable, we chose to go with post-hoc approaches such as model perturbation (Cavalcanti et al. 2024) or Layer-wise Relevance Propagation (LRP) (Bach et al. 2015). These methods generate token-level attribution maps that provide insights about which parts of the input contributed most to a prediction. The quality of such explanations is often discussed in terms of two key criteria: faithfulness (i.e., whether the explanation reflects the model's ac-

tual decision-making process (Ribeiro et al. 2016; Jacovi and Goldberg 2020)) and plausibility (i.e., whether the explanation is understandable and convincing to a human reader (Herman 2017)). Yet, these criteria remain difficult to evaluate in practice, and their interplay is still poorly understood. We decided to opt for the implementation of the Layer-wise Relevance Propagation (LRP) method made by Chefer et al. (2021) since it had already shown promising results and was considered as a good trade-off (despite not being perfect) with respect to both aspects.

While performing our study, we discovered something that was quite surprisingly poorly documented. When re-running our experiments with different randomness (in the form of a seed), we observed similar performances but when explaining the model's predictions with LRP, we obtained different list of most relevant words from one run to another. This phenomenon immediately rose our attention. Funnily enough, this discovery was done around the time when ChatGPT was publicly released. Its capabilities were far superior than those of any text generator released until then and its generated texts were much harder to detect. This was (and still is) even worse when working with short texts and/or allowing slight adversarial changes (Antoun et al. 2023; Sadasivan et al. 2025). From this point on, it quickly became a game of cat and mouse where the limiting factor would be the computational power of the generator against the one of the detector (Bhattacharjee and H. Liu 2023). We inferred that this research area was a dead end for us to follow as all the insights and ideas we had from prior works were shattered by this release. We therefore decided to focus only on the explanation sensitivity to randomness from that point on, as there seemed to be more room for fruitful research. I dedicated the last two years of my PhD thesis to it, which is what this document will cover in depth. It follows that all the prior works (automated news detection, datasets and transformers' fine-tuning study) which led to this question of explanation sensitivity to randomness will not be described more than already done but their respective published papers are available in the Appendix.

To be a bit more specific about this sensitivity to randomness, modern neural architectures are typically trained with stochastic optimization methods that vary depending on multiple sources of randomness such as parameter initialization, data shuffling, and dropout. This randomness can be controlled by a seed parameter that usually only serves as a mean to ensure the reproducibility of the training process. As a result, two models trained on the same data with different random seeds may achieve indistinguishable accuracy, while still pro-

ducing different explanations for their predictions. This phenomenon resonates with Breiman's Rashomon effect (Breiman 2001), which highlights the existence of many functionally equivalent models for a given task. The problem is that, during an experiment, the seed parameter is typically chosen uniformly at random, as there is no way to know *a priori* the performances a seed will lead to. It follows that, in the extreme case where explanations would vary arbitrarily across equivalent models, selecting a single explanation would be meaningless. This raises fundamental questions about the stability of classification transformers' explanations. Among them :

1. Can sensitivity to randomness be significant enough to affect the explainability of transformer models ?
2. Can we design meaningful metrics to quantify sensitivity to randomness ?
3. How does the sensitivity to randomness of model-based explanations compares to the variability observed among human annotators ?
4. What factors can influence the sensitivity to randomness ?

This thesis addresses them through a series of contributions. Chapter 1 will first briefly introduce all the prior knowledge to have before jumping into the different research questions.

In Chapter 2, we demonstrate that explanation sensitivity to randomness in transformers is not anecdotal but systematic: equivalent models trained with different seeds can produce substantially different attribution patterns, despite having statistically indistinguishable accuracy. This highlights the need to treat randomness as a critical factor in transformer explainability analysis.

In Chapter 3, we propose a framework to move beyond purely qualitative visualizations (e.g., box plots) by introducing quantitative stability metrics. We first formally define the explanation format that we use. We then evaluate information theoretic metric in the name of signal, noise, and signal-to-noise ratio (SNR) as candidate measures of explanation stability. We show the partial unsuitability of the signal and the SNR metrics. On top of it, while noise captures absolute fluctuations, we show that it fails to fully reflect relative differences between words in a text, which are often crucial for interpretation. To address this limitation, we introduce a correlation-based metric that complements the noise by capturing relative attribution differences, which leads to

a more robust methodology to characterize explanation sensitivity to randomness.

In Chapter 4, we extend the analysis beyond models by comparing the variability of explanations across equivalent LLMs with the variability across human annotators. While most prior work has compared single explanations from humans and models (Sen et al. 2020), we instead focus on the distribution of explanations in each group. This perspective reveals systematic divergences between model and human explanation variability, but also shows that simple post-processing strategies inspired by human annotation patterns (e.g., discretization) can significantly reduce this gap.

Finally, in Chapter 5, we investigate different variations of the setup to characterize explanation sensitivity to randomness. While Chapter 4 focuses on post-processing the explanations themselves, this last chapter evaluates the impact of changes on the data and the model side. First, we study the impact of the context (in the sense of the text's structure) and of the difficulty of the task. We show that both these factors can have a significant impact on explanation sensitivity to randomness. Then, on the model side, we study whether the fine-tuning surface (i.e., the amount of bits effectively modified when fine-tuning a transformer model) serves as a proxy for explanation sensitivity to randomness. We explore this question through two interventions: weight quantization (Dettmers et al. 2023; Jacob et al. 2018) and parameter freezing. Our findings reveal that these strategies affect explanation variability in markedly different ways.

Contents

Chapter 1

Background

1.1 Transformer architecture

The Transformer architecture (Vaswani et al. 2017) represents a fundamental shift in the design of natural language processing (NLP) models. Unlike recurrent neural networks (RNNs) or convolutional neural networks (CNNs), which rely on sequential or local processing, the Transformer is based entirely on attention mechanisms, allowing it to model global dependencies. This design led to significant improvements in terms of both efficiency and performance across a wide range of NLP tasks, first for text classification (Devlin et al. 2019) and then for text generation (Brown et al. 2020; Achiam et al. 2023). At a high level, transformers follow an encoder-decoder structure as shown in Figure 1.1

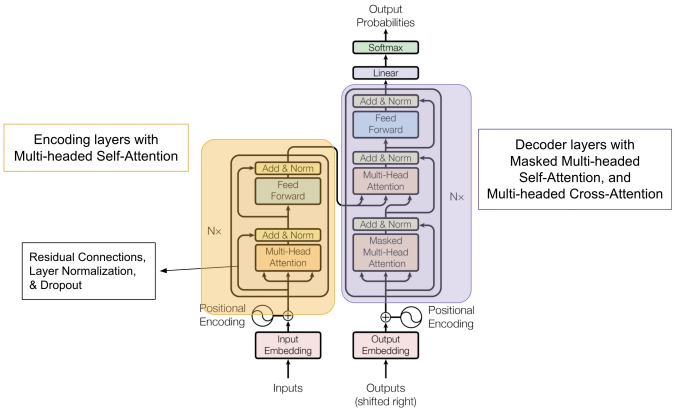


Figure 1.1: Scheme of the transformer architecture as presented by Vaswani et al. (2017).

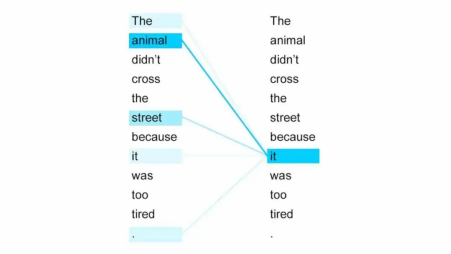


Figure 1.2: An example of the attention mechanism.[1] The word "It" has a high attention toward the word "Animal" which it refers to in this sentence. Its internal representation inside the model, or embedding, will therefore reflect that information.

The encoder maps an input sequence of tokens into a set of continuous representations, that we refer to as "embeddings". It is mainly based on the attention mechanism that enables the model to attend to information from different positions in the input simultaneously. This mechanism, schematized in figure 1.2 allows to take context into account to adapt each word representation. It is usually implemented in the form of multi-head self-attention, in order to capture multiple word dependencies in the input text at the same time.

The decoder generates a prediction (e.g., for BERT-like models) or an output sequence step by step (e.g., for GPT-like models), attending both to previously generated tokens and to the encoder's representations. For generative models, it is usually at least of the same size as the encoder, and it uses both self-attention (to attend to previously generated tokens) and cross-attention (to attend to the encoder's output). For classification models, which is the only type of model we cover in this thesis, the decoder is usually a simple fully connected neural network only using linear combinations of the encoder outputs to perform its task.

The introduction of the Transformer architecture had a huge impact. Its parallelizable structure makes it more computationally efficient than RNNs, and its ability to model long-range dependencies surpasses CNN-based approaches. This architecture forms the foundation of modern large language models (LLMs) but has since then also been extended beyond NLP into domains such as vision (Han et al. 2023) or speech (Latif et al. 2023).

[1] Image from https://spotintelligence.com/2024/02/29/bert-nlp/

1.2 Transformer training

Transformers are trained in two different steps. The first one consists of a self-supervised pre-training where the model learns to fill blanks created by randomly hiding some words in the texts. It allows to learn a good language representation in the form of embeddings. Its duration depends on the model's number of parameters and the machine it runs on, but it usually requires a high amount of computational resources to be achieved. This operation is therefore usually only done once, by the entity releasing the model.

The second phase consists in adapting the model to a smaller dataset in order to perform a given task (e.g., classify texts between two sources). This step is usually much faster than the first one. To give an order of magnitude, pre-training the models we used would take around 15 days on the server we had acess to, while the adaptation to a specific task based on a corpus of 8000 texts lasted approximately 10 minutes.

To adapt a pre-trained transformer to an end task, Devlin et al. (2019) presented two different methods. In the first one, that we denote as `fine-tuned`, all the weights of the encoder blocks and the classification head are jointly trained. In the second one, that we denote as `frozen`, only the classification head, that acts as the decoder part, is trained while the encoder blocks are frozen. It results in the model only learning to use the embeddings without modifying them.

Regardless of the chosen alternative, this adaptation to an end task is not a deterministic process. It relies on random initialization as well as stochastic heuristics in order to converge to a local optimum. It is governed by a random seed used for optimization, which influences three key aspects: (*i*) the initialization of the weights of the classification head, (*ii*) the order of the texts in the training set, and (*iii*) the neurons targeted by the drop-out technique, that aims at limiting overfitting.[2] Aside from the random seed, there are other hyper-parameters that influence the final model's performances. Among them, we can cite the learning rate (i.e., how impactful are each of the training inputs), the batch size (i.e., how many examples are considered simultaneously) and the number of epochs (i.e., how many times the model trains over all the examples).

[2]We did not modify this third parameter during our work: it remained at its default value of 10 %

The particular transformer model we used during most of this thesis is called CamemBERT (Martin et al. 2020). It is a French language model of 110M parameters based on the same architecture as RoBERTa (Y. Liu et al. 2019), specialized for text classification. It is mainly composed of an encoder bloc on top of which a classification head composed of two connected layers acts as a decoder. It was pre-trained on the French part of the OSCAR dataset (138 GB of text) by its creators, and we adapted it during our experiments. A series of combinations of values for each hyper-parameter were empirically evaluated, with the optimal values selected according to the accuracy obtained by the model on the validation set and for the random seed 0. For the learning rate, we tested values ranging from 1×10^{-6} to 1×10^{-4}. For the batch size, we tested values ranging from 1 to 64. Finally, for the number of epochs, we evaluated the accuracy obtained by training the model during 1 to 4 epochs. We will specify the hyper-parameters and the dataset that we used for each of our experiments when presenting the different setups.

1.3 Model Explainability

One of the main effect of Transformers high parallelizability is that it led to surprisingly high scaling capacities (Dehghani et al. 2023; Kaplan et al. 2020) which in turn pushed toward the adoption of bigger and bigger models. While transformers initially had around 10^8 parameters (Devlin et al. 2019), modern models can have up to 1000 times more. Considering the amount of internal connections it leads to, the first concern arising when training and using these models is their inherent lack of explainability. While the high-level behavior of a transformer can be explained easily, the specific reasons why they perform so well at a given tasks are intrinsically much harder to grasp. The lack of understanding of these complex, often described as "black-box", models decisions is a major concern in many contexts in which they are used. This is increasingly concerning when their decisions have important implications, as in the legal domain, among others (Zini and Awad 2023). It led to a growing body of research trying to dissect and put into light what was used by these models when performing their task. This field is usually presented under the name of Explainability or Explainable Artificial Intelligence (XAI).

The definition of the necessary conditions for model explainability is not yet the subject of a broad consensus (Murdoch et al. 2019). As detailed by Lyu et al. (2022), various supposedly desirable criteria for the explainability of a

model have been introduced in the literature, but the relationship between these criteria is not formally established and their rigorous evaluation is often difficult. For example, the two criteria most commonly cited as fundamental to the explainability of a model are **faithfulness** and **plausibility**. Faithfulness is defined as the ability of an explanation to accurately reflect the (algorithmic) reasoning process that led to a prediction (Ribeiro et al. 2016; Jacovi and Goldberg 2020). Plausibility is defined as the ability of an explanation to be understandable and convincing to a human reader (Herman 2017; Jacovi and Goldberg 2020). The minimality of explanations is sometimes presented as an additional desirable criterion (Miller 2019). Following Occam's razor, this criterion suggests that among different explanations, the simplest one is usually the best. Note that all these criteria are inherently hard to evaluate formally.

A more technical and therefore easier to quantify criterion is the explanation **sensitivity** to different types of variation (Adebayo et al. 2018; Yeh et al. 2019; Manna and Sett 2024; Sundararajan et al. 2017). For example, sensitivity to input data implies that an explanation must (resp. must not) depend on variations in the texts being processed if these variations change (resp. do not change) the model's prediction. More directly related to our concerns, it has also been proposed that an explanation should be model-sensitive, i.e., that it should (resp. should not) depend on changes in the model that affect (resp. do not affect) predictions. More specifically, the concept of implementation invariance formalizes the fact that functionally equivalent models (i.e., models that have identical predictions for any input) should have identical explanations (Sundararajan et al. 2017). However, this need is mitigated by some authors, since there is nothing to prevent two different algorithmic methods from deterministically reaching the same solution (Lyu et al. 2022).

1.4 Explanation methods

For classification transformers, various explanation methods have been proposed in the literature to combine these criteria. They fall into two categories, based on their scope (Danilevsky et al. 2020). **Global explanation** methods aim to explain the reasoning of the model for classifying any document. On the other hand, **local explanation** methods focus on the reasoning of the model for a given prediction. Our work is centered around the explanation methods falling into that second category. These methods are further divided into several subcategories (Lyu et al. 2022), depending on whether they are for exam-

ple based on similarity with other inputs (Hanawa et al. 2021), input perturbation (Ribeiro et al. 2016), back-propagation mechanisms (Wu and Ong 2021; Chefer et al. 2021), or counterfactual analysis (McAleese and Keane 2024).

We have interests in methods based on gradients and back-propagation mechanisms that attempt to interpret the attention layers used by transformer models (Kovaleva et al. 2019; Clark et al. 2019). Although the debate about whether attention alone can be used as a valid source of explanation remains open (Bibal et al. 2022), previous works showed that combining multiple layers can generate convincing explanations (Srinivas and Fleuret 2019; Abnar and Zuidema 2020). A crucial feature of these methods for our study is that it produces deterministic explanations for a given model instance: the same input and the same model parameters always yield the same explanation, on the contrary of other post-hoc local methods, such as LIME (Ribeiro et al. 2016) which relies on stochastic input perturbation. This determinism is crucial, as it allows us to isolate the effect of randomness introduced during model training.

Among these gradient-based methods, we chose to use Layer-wise Relevance Propagation (LRP) (Bach et al. 2015) as it appears to be one of the most faithful methods (Arras et al. 2017; Lyu et al. 2022) while providing more plausible explanations than other gradient based methods (Wu and Ong 2021; Chefer et al. 2021). It therefore represents a good starting point to analyze the sensitivity of the explanations of Transformers to their training randomness. The LRP method can explain the predictions made by any deep learning model. It assigns a relevance score to each token, which is measured by tracking, layer by layer, the contribution of the token to the prediction outputted by the model. It starts from the output value and back-propagates the relevance of a neuron in layer $l + 1$ to its contributing neurons in layer l using conservation constraints. It follows this principle, layer by layer, all the way up to the input.[3]

Different rules define how the relevance of a neuron at a given level of the model should be propagated to the previous level, with the constraint that the relevance scores at each level must sum up to 1. This propagation constraint is more difficult to satisfy for certain layers and this method is continuously being improved (Binder et al. 2016; Voita et al. 2019). In this thesis, we use (Chefer

[3]To obtain a value per token, readable in natural language, and not per BERT 'WordPiece' (as originally encoded by the CamemBERT and RoBERTa architectures), we concatenate the different parts and compute the average of their attention values.

et al. 2021)'s implementation of the LRP method in all our experiments. Figure 1.3, taken from Montavon et al. (2019) shows an illustration of the method.

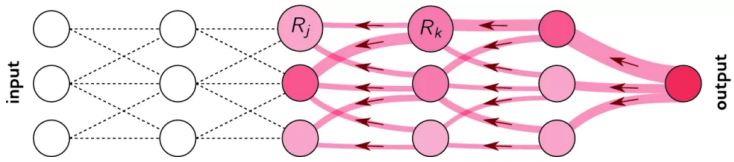


Figure 1.3: Illustration of the LRP method

1.5 Visualization of explanations

The way in which explanations are visualized may influence their plausibility to human readers (Reif, Yuan, Wattenberg, Viégas, et al. 2019). One of the most prevalent approaches to visualize the explanations of predictions made by a text classification model is the use of saliency maps (J. Li et al. 2016), also interchangeably referred as "attention" maps in the literature, which are assumed to be easily understandable by a human reader (Sen et al. 2020). These maps comprise the highlighting of text elements which were identified as the most influential in the model's decision-making process, employing varying shades of color, allowing the importance of each token to be visualized separately. Consequently, attention maps are constrained to providing localized explanations at token-level. They are thus unable to highlight the influence of other types of features that may be decisive in the model's prediction, such as long-term dependencies and syntactic or semantic relationships between different explanatory elements. The attention map format nevertheless has the advantage of being very comprehensible a priori to human readers (Sen et al. 2020), which contributes to the plausibility of the explanations.

This format also has the advantage of being adaptable to other models, even if they are intrinsically explainable. For example, in Chapter 2, we compare the discriminant features used by a logistic regression model (presented in the next section) with the ones found in transformer explanations. To do so, we generate word-level explanations using so called "linguistic attention maps". In this case, a word is highlighted in a darker color if it is associated with one or more linguistic features and if the weights of these features in the model are high. Thus, the linguistic attention attributed to a word i for a feature j can be written as:

$$A_{ij} = \begin{cases} \frac{w_{ij}}{\sum_i w_{ij}} \times T_j & \text{if sign}(T_j) = \text{sign(prediction)}, \\ 0 & \text{in the opposite case,} \end{cases} \tag{1.1}$$

where the first term represents the relative importance of the word *i* for the feature *j* and the second represents the coefficient associated with the feature *j* in the regression.[4] Finally, the relevance of a word *i* for all the combined features is averaged as $A_i = \sum_j A_{ij}$.

Note that, on the contrary of using the regression's coefficients, this approach is restricted to the explanation of word-level features and cannot reflect the impact of features related to the entire text. Yet, since the feature-based model only contains two of them (the type token ratio and the average word length), we consider that this explanation method is sufficiently faithful to allow for comparisons with attention maps obtained thanks to LRP. Figure 1.4 shows an example of two attention maps.

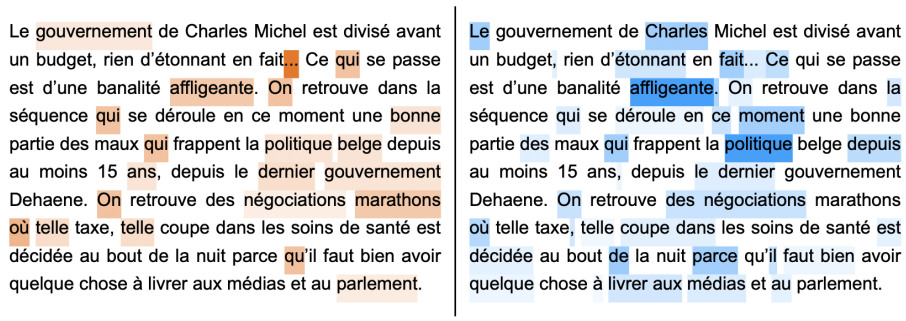


Figure 1.4: Attention maps derived from the logistic regression model (left, in orange) and from CamemBERT (right, in blue) for the same article.

1.6 Linguistic model

To have a deterministic baseline to compare the explanations of the transformer models to, we use a logistic regression classifier. Such models are recognized as being easier to explain (Gémes et al. 2021). However, they are limited

[4]For categorical traits (adjectives, verbs...), this term is 0 or $\frac{1}{n}$, but it may differ from these limiting cases when considering continuous variables (imageability, concreteness...).

by achieving lower accuracy for many applications and are therefore often neglected in favor of deep learning approaches (Q. Li et al. 2020). They represent a very different trade-off between accuracy and explainability.

Since most of our experiments revolve around the task of classifying between pieces of *news* and *opinion*, the classifier utilizes nineteen textual features derived from the state of the art on linguistic subjectivity and identified as effective predictors of opinion in French-language press discourse (Escouflaire 2022). Most of these features are based on the presence or proportion of specific tokens or types of tokens in the text to be classified. The following linguistic features are considered: adjectives, verbs, first person pronouns and determiners, relative pronouns, the indefinite pronoun 'on' (*we*/*one*), expressive punctuation marks (semi-colons, exclamation marks and question marks), inverted commas, numbers, negation words, words longer than seven characters, words appearing in the sentiment lexicon of (New et al. 2004) or the NRC lexicon (Mohammad and Turney 2013). Two features are not related to the tokens but apply to the entire article: Carroll's corrected type-token ratio (Carroll 1964) and the average word length.

We used a grid search to identify the optimal combination of hyper-parameters for the regression. The resulting model will be referred to as LING-LR, an abbreviation derived from the linguistic traits upon which it is based. In this work, it will serve as an example of a deterministic model always providing the same explanations. While simply giving the regression coefficient would lead to perfectly faithful explanations, we rather use linguistic attention maps to visualize the explanations of a feature-based model in a format similar to that of LRP.

1.7 News vs. Opinion Classification

Unless mentioned otherwise, the task we train our models on is to determine whether a text is categorized as `news` or `opinion`. Even if it is not directly related to our topic of study which is the explanation sensitivity to randomness, we next present the history and reasons that pushed us to consider that specific use-case during most of this work.

The journalistic world traditionally separates its production in two main classes: *opinion* articles, such as columns and editorials, which are overtly subjective, and *news*, which includes all press content intended to follow the stan-

dards of journalistic objectivity (Schudson 2001; Esser and Umbricht 2014). In today's media ecosystem, where information is disseminated via the (mostly) uncontrolled and unprofessionalized channels of digital social networks and where users are placing less and less trust in traditional media (Latkin et al. 2023; Newman et al. 2024), it is increasingly important to be able to clearly distinguish between *opinion* and *news* content. To this end, automated text classification can improve citizens' ability to navigate online information, improving their understanding of societal issues, and helping them form their own opinions.

In the NLP field, classifying press articles according to their genre is a popular automation task, in English as well as in less-resourced languages such as French (Katari and Myneni 2020; Vernier et al. 2009; Todirascu 2019). It can be accomplished to some degree with traditional feature-based or dictionary-based bag-of-words approaches (Krüger et al. 2017), or through more complex dynamic systems also capturing discursive information such as argumentation or figurative language (Benamara et al. 2017). In recent years, studies have demonstrated that transformer models are able to predict text genre, in particular distinguishing between journalistic genres, with higher accuracy than previous state-of-the-art NLP models (Alhindi et al. 2020; Singh et al. 2020). We however wanted to evaluate to what extent it was possible to explain their predictions to end users thanks to explainability methods.

To train our models, we use the InfOpinion dataset (Bogaert, Escouflaire, et al. 2023b) that we designed during the first part of this thesis. The paper introducing the dataset is available in appendix. It is based on the RTBF dataset (Escouflaire, Bogaert, et al. 2023), whose paper is also available in the appendix, and is composed of 5,000 texts for the `news` class (editorials, chronicles, ...) and 5,000 texts of the `information` class (reporting, press dispatch, ...), tackling similar topics. This binary categorization relies solely on the articles' annotation by their authors as either *news* or *opinion*. The dataset is split in train, validation and test sets with an 80%/10%/10% ratio.

In order to assess the robustness of the models to potential changes in the data distribution, a second corpus was created. It consists of press articles published by another media, *Le Soir* (www.lesoir.be), which is the most popular daily newspaper in French-speaking Belgium. In this work, the LeSoir-InfOpinion corpus serves only as a test set for the classification models. It was built following the same methodology and preprocessing steps presented

by Bogaert, Escouflaire, et al. (2023b). This corpus contains 1,000 articles published online between 2015 and 2021. As with the RTBF-InfOpinion dataset, the LeSoir-InfOpinion corpus is divided equally between opinion and news articles, following the same editorial categories as those chosen for the RTBF corpus. The corpus comprises a total of 669 154 tokens, with an average of 859 tokens for opinion articles and 480 for news articles.

Chapter 2

Can Model Sensitivity to Randomness impact Explainability ?

The content of this chapter has been published in French in (Bogaert, Escouflaire, et al. 2023a) and translated in English in (Bogaert, Escouflaire, et al. 2024).

2.1 Motivations

This chapter focuses on whether the sensitivity of large language models to the random elements of their training can be significant enough to affect their explainability. More specifically, training a learning method usually requires the selection of a number of hyper-parameters, both for the model itself (e.g., network size and topology) and for the optimization algorithm used (e.g., learning speed and batch size). In addition, stochastic optimization methods exploit a degree of randomness that is often generated deterministically from an additional parameter, usually called "seed". The seed is used to initialize a pseudo-random number generator to produce the amount of randomness required by the optimization algorithm. This parameter is usually made public for the purpose of reproducibility of results. Considering the seed as a hyper-parameter in training is rarely recommended in practice, as it does not allow any strategy other than an exhaustive search (as nothing distinguishes the randomness generated with one seed from that generated with another). However, there is nothing preventing it from being used as an hyper-parameter nor preventing the

comparison of the quality of the models obtained with different seeds. It is of interest to us in this context because it constitutes a pristine example of a completely random element used during model training. In this work, we examine the effects of random seeds used as hyper-parameters in the training phase. Note that we could also extend this consideration to other hyper-parameters which, even if they are not selected as randomly, involve some random elements in their selection.

Since a training process with several randomly chosen hyper-parameters can lead to different models, these models can naturally have different classification accuracy. The hyper-parameter values that lead to the best performing model are generally selected. Our study therefore requires a first restriction: we are interested in the subsets of hyper-parameters values (i.e., only random seeds in our case) that lead to a subset of models with sufficiently close accuracy. This subset of models, refered as a Rashomon set (Breiman 2001) or (statistically) equivalent models are models whose differences in performance are not statistically significant. Moreover, even statistically equivalent models are not necessarily functionally equivalent: we are interested in the subset of data inputs for which these models output the same prediction. We refer to these equivalent model inputs giving the same prediction as "compatible".

Informally, these definitions allow us to restrict our study to subsets of inputs for which there is no way to determine whether one model is preferable to another. In this context, we claim that if a learning method leads to a non-negligible set of equivalent models whose explanations differ, then limiting ourselves to the explanation of a single model obtained with that learning method is insufficient. In particular, if it is true that different algorithmic methods can lead to the same solution (in which case the explanation of a single method may be sufficient), the presence of randomness in the process of training equivalent models forces us to ensure that the statistical distribution of the explanations from equivalent models is sufficiently different from the uniform distribution. Such a distribution would imply that all possible explanations are equiprobable, and the choice of an explanation over another would then be completely arbitrary.

Complementary work investigating the sensitivity of explanations to the hyper-parameters (chosen randomly or heuristically) used in the explanation methods themselves have already been done (Bansal et al. 2020). This sensitivity is generally presented as detrimental because it implies unpredictability

of explanations for a given model and prediction. However, this study differs from ours, which focuses on the randomness of the model's hyper-parameters used during training rather than on the explanation methods. We also mention (Bethard 2022), which highlights different types of use (justified or hazardous) of randomness in training. However, this general discussion is not directly related to the issue of explainability. Finally, while the topic of explanation sensitivity to randomness recently gained interest, mostly under the name Rashomon effect (Müller et al. 2023), and was studied to a limited extend in medical imaging (Watson et al. 2022), this is, to the best of our knowledge, the first in depth analysis of this phenomenon for transformers used for text classification.

This chapter has the following structure: we first show that for a given reasonable combination of a CamemBERT model, fully fine-tuned according to (Devlin et al. 2019) and the Chefer's Layer-wise Relevance Propagation (LRP) implementation (Chefer et al. 2021), such non-trivial sets of equivalent models can be observed in practice. This combination of a language model, a fine-tuning method and an explanation tool is applied to a specific text classification task that consists in predicting whether French press articles belong to the journalistic genre of opinion (editorials, chronicles) or news (dispatches, newswire). We then examine the correlation between explanations obtained thanks to models differing only in terms of the randomness used to optimize them in section 2.3. We continue with a visual characterization of the effect of this randomness on explanations in section 2.4. We offer a qualitative analysis of specific attention maps to illustrate the previous assessments in section 2.5. Given the sensitivity of the explanations provided using CamemBERT, we finally compare these results with a more traditional NLP method based on a logistic regression model applied to textual features, referred as LING-LR (see Section 1.6 for further details), in order to illustrate the key explainability differences between non-deterministically and deterministically trained models.

2.2 Generation of statistically equivalent models

To generate equivalent models, we first repeat the fine-tuning of the CamemBERT model 200 times (with 200 different random seeds), using the same set of 8 000 texts during each training. We train our models during 2 epochs using a learning rate of 2×10^{-5} and a batch size of 4. All the other hyper-parameters remain at their default value. To modify the randomness used during the train-

ing, we make use of the seed parameter. The seed controls the random initialization of the classification head's layers, the droped-out neurons during parts of the training and the order in which the training database is shuffled (as we have *a priori* no reason to use a pre-determined order). We then compute the accuracy on the test set for all models and for subsets of models, selecting the models that produced the closest or highest accuracies. We then evaluate if the difference between the most accurate model (a) and the least accurate model (b) in the subsets is significant or not. To do this, we estimate the Z-statistic (Lehmann and Romano 2008), which is used to determine whether two proportions (in this case, accuracies) differ:

$$z = \left| \frac{a - b}{\sqrt{\frac{\frac{a+b}{2} * (1 - \frac{a+b}{2})}{n}}} \right|. \tag{2.1}$$

We assume that the differences in accuracy are significant for z greater than 1.96, which corresponds to a p-value of less than 0.025. For lower z (and higher p values), we conclude that the accuracy of the models does not differ significantly. We therefore define the models in the corresponding subsets as (statistically) equivalent, since their accuracy does not allow to prefer one of them over any of the others.

	Min. Acc.	Max. Acc.	p-value
CamemBERT(200 models)	93,1	96,6	$2,8 \times 10^{-7}$
CamemBERT(150 closest)	94,5	95,9	0,0191
CamemBERT(150 most accurate)	95,0	96,6	0,0860
CamemBERT(100 closest)	95,0	95,7	0,1466
CamemBERT(100 most accurate)	95,4	96,6	0,3227
CamemBERT(50 closest)	95,3	95,6	0,3245
CamemBERT(50 most accurate)	95,7	96,6	0,5616
LING-LR	88,9	88,9	/

Table 2.1: Minimum and maximum accuracy, ϵ value and p-value of sets of models.

The results of these estimates are presented in Table 2.1. It also provides the accuracy, estimated on the same texts, of the LING-LR model which converges to a unique solution. These results show that from the 200 trained models, we are able to select a subset of 100 equivalent models. It should be

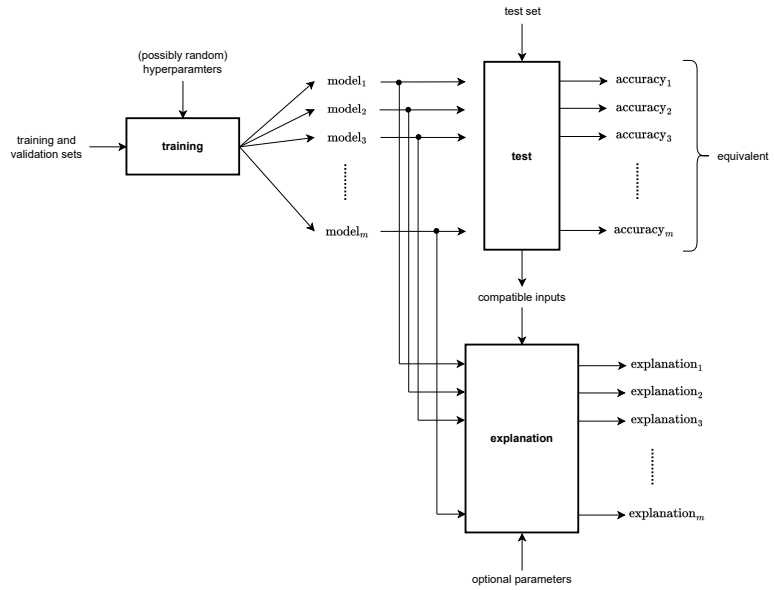


Figure 2.1: Generation of equivalent models and compatible inputs.

noted that this definition of equivalent models depends on the size of the test set: as its size increases, the accuracies of the models are estimated more precisely and smaller differences are considered significant. This means that more models have to be trained to select a subset of equivalent ones. However, as the amount of randomness used to optimize a model is virtually unlimited, this observation does not fundamentally change the explanation sensitivity to randomness phenomenon that we are going to study: it only increases the computational cost of highlighting it. Finally, we explain our subset of 100 equivalent models thanks to the LRP explanation method (Chefer et al. 2021). Figure 2.1 shows a scheme of the entire process.

In additon, Table 2.2 provides the computational time needed to train these models (on 8000 texts) and to explain their predictions (on 1000 texts) in our setup. For fair comparison, while explaining the LING-LR model could be done instantly by providing the regression coefficients, we take into account the time to translate them in a format similar to LRP explanations. To do so, we make use of Linguistic Attention Maps (LAM) as presented in Section 1.5. Even when using LAM, it takes much longer to both train and explain

a CamemBERT model. While this is not surprising, it follows that having to generate bigger sets of equivalent models' explanations can quickly become a concrete problem.

Models	LING-LR	CamemBERT
Pre-training	15 min.	13 days*
Training	0.5 sec./fit	4 min./epoch
Inference	1.3 sec.	9.6 sec.
Method	LAM	*LRP*
Explanation	4 sec.	4 min.

Table 2.2: Computation time of LING-LR and CamemBERT models, estimated on the 1 000 articles from the RTBF test set, using a NVIDIA RTX A6000 GPU.

2.3 Explanation correlation

Having identified sets of equivalent models, we now assess the extent to which the explanations provided by these models differ. We selected two pieces of compatible text of similar length (49 tokens for text 1, 51 tokens for text 2), classified by all the models as *news* articles. Note that the study of nearly compatible inputs would also be possible, and would simply require examining the correlations for the *news* and *opinion* decisions separately. As a first informal analysis to highlight differences, we estimate the correlation between two vectors of explanations. Based on the 100 equivalent explanations for each text, we construct these vectors by concatenating 50 explanations. We then compute the correlation between these two vectors. In the case of identical (resp. random) explanations, this correlation would be 1 (resp. 0). Finally, we repeat this operation for 10000 different concatenation vectors to compute a *bootstrap* confidence interval (Efron and Tibshirani 1993). The results of these estimates are reported in Table 2.3.

We observe that the explanations differ significantly, regardless of whether the subsets of models are selected according to the value or the proximity of their accuracies. Interestingly, we also observe that these correlations vary depending on the text. This suggests that the impact training randomness has on

	Model	Pearson correlation	
		Estimate	Confidence interval
Text 1	CamemBERT (200 models)	0,152	± 0.058
	CamemBERT (150 closest)	0,145	± 0,072
	CamemBERT (150 most accurate)	0,140	± 0.066
	CamemBERT (100 closest)	0,126	± 0.086
	CamemBERT (100 most accurate)	0,149	± 0.078
	CamemBERT (50 closest)	0,118	± 0.131
	CamemBERT (50 most accurate)	0,184	± 0.093
Text 2	CamemBERT (200 models)	0,395	± 0.053
	CamemBERT (150 closest)	0,398	± 0.062
	CamemBERT (150 most accurate)	0,409	± 0.058
	CamemBERT (100 closest)	0,380	± 0.074
	CamemBERT (100 most accurate)	0,423	± 0.072
	CamemBERT (50 closest)	0,344	± 0.107
	CamemBERT (50 most accurate)	0,466	± 0.088

Table 2.3: Pearson correlation (with bootstrap confidence interval) between two vectors of equivalent models' explanations for compatible inputs.

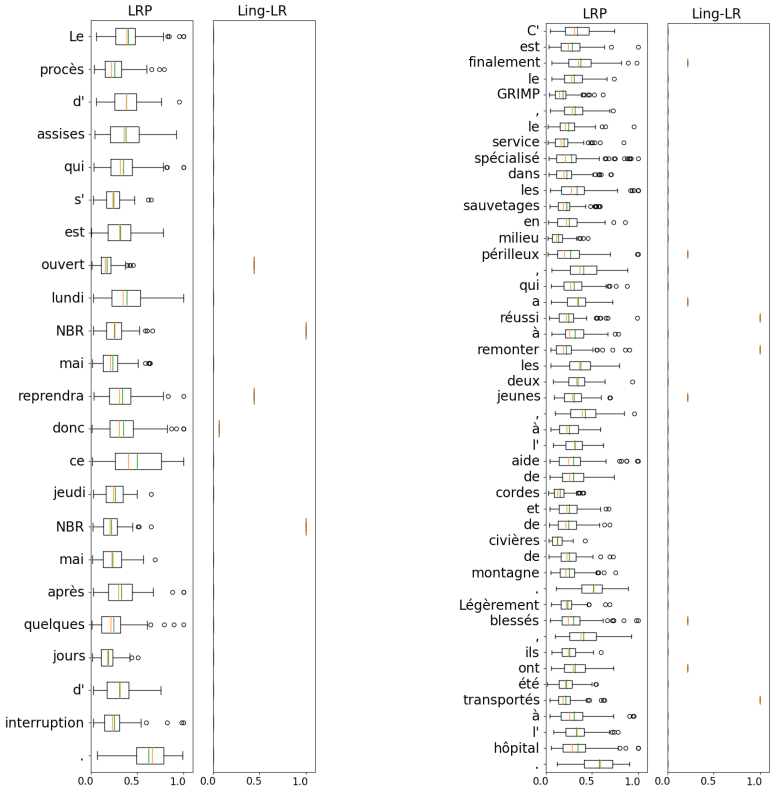
the explanations is text specific, which further confirms the interest in characterizing explanation sensitivity to randomness.

2.4 Visual characterization of the explanations sensitivity to randomness

To illustrate the variability in the CamemBERT models' explanations, we generate box plots corresponding to the explanations of two short texts from our test set. These give an intuition on how often the explanations of equivalent models give importance to each token.

We emphasize again that, to the best of our knowledge, the faithfulness (i.e., the extent to which an explanation accurately reflects the algorithmic reasoning process that led to a prediction) and plausibility (i.e., the extent to which an explanation is understandable and convincing to a human being) of a model's explanations have so far only been studied for fixed models, resulting from a given execution of their optimization. But because of the sensitivity to randomness, a human reader has to consider a more complex and possibly

richer explanation distribution, rather than a unique explanation as for the deterministic model. On top of it, each of these explanations could be more or less faithful to its corresponding model, and more or less plausible for a human reader.



(a) Explanations for text 1.

(b) Explanations for text 2.

Figure 2.2: Visual characterization of the explanations of 100 equivalent models. The X axis corresponds to the distribution of token relevance.

The results in Figure 2.2 show that when considering all models, all tokens in the text end up having a non-zero relevance. Although the distribution of independently considered tokens is not uniform [1], these results at least question the plausibility of the obtained explanation distribution. In addition, and rather

[1] and a uniform distribution of independently considered tokens does not imply a uniform distribution of all possible explanations

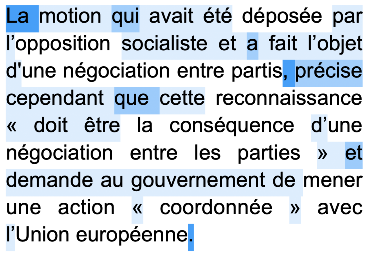
trivially, they also confirm a reduction in the minimality of the explanations compared to the (deterministic) explanations of the LING-LR model. Such results ask for a more global characterization of model explainability in the future, taking into account how they can vary depending on random elements used during model's training.

2.5 Qualitative analysis of attention maps

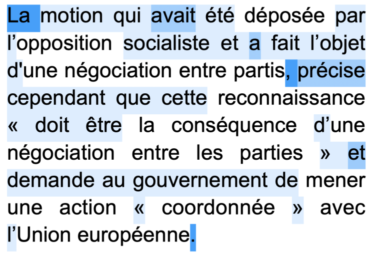
Finally, to confirm that the differences in explanations quantified in section 2.3 and visually characterized in section 2.4 are not simply a combination of similar explanations and easily filtered outliers, we conclude this chapter with some examples of attention maps in Figure 2.3. We first observe that the explanations in Figures 2.3a and 2.3b are visually very similar for a human reader. Despite minor differences in the relevance certain tokens, both models seem to focus on the same elements, which are mainly structural elements in the text's discourse such as strong punctuation marks (".", ","), conjunctions (*que*/*that*, *et*/*and*), inverted commas, and the first words of propositions. In contrast, Figure 2.3c, which was created from a third seed, shows a very different attention map from the first two. Here, the emphasis lies more on semantically charged tokens, in particular nouns representing the political actors referred to in the text (*partis*/*parties*, *gouvernement*/*government*, *l'Union européenne*/*the European Union*). These attention maps suggest that very different explanations, which are therefore difficult to filter out without further analysis, can be extracted from equivalent models. The map in Figure 2.3d finally shows that the tokens which most influence the LING-LR model are the presence of quotation marks and verbs. It also differs from the explanations of the CamemBERT models.

2.6 Discussion

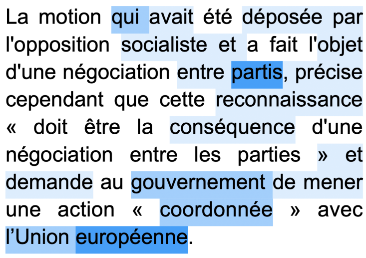
The previous results clearly show that the explanations of large language models can be sensitive to the randomness in their training. On the one hand, we conclude that it is necessary to characterize this sensitivity, if only to be satisfied that the distribution of explanations is sufficiently different from the uniform one. Otherwise, the choice of one explanation over another would be completely arbitrary. To the best of our knowledge, the characterization of this sensitivity has not yet been the subject of systematic studies in the literature.



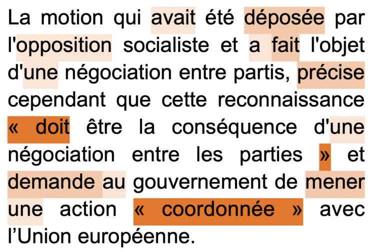
(a) CamemBERT (seed = 1).



(b) CamemBERT (seed = 2).



(c) CamemBERT (seed = 3).



(d) LING-LR.

Figure 2.3: Attention maps of four different models for a *news* article (correctly classified by all models).

On the other hand, these results raise the essential question of the extent to which this sensitivity is a problem. Taking the example of legal applications, a situation might occur where an automated judgment offers to a person under trial a large number of explanations for his sentence. These explanations are quite different from the point of view of human understanding, but indistinguishable from the point of view of the algorithm. This potential situation appears incompatible with the requirement that a legal judgment should be comprehensible. On top of it, presenting the litigant with one single explanation, even if it happens to be plausible, would be at least partly arbitrary. While we could also find a situation where a model produces several clusters of explanations corresponding to different but plausible interpretations of legal texts, which would be less pathological, it would nevertheless be difficult to come up with a single explanation in the end.

Furthermore, the combination of this need for characterization with the computational complexity of large language models training (cf. Table 2.2) suggests that simpler models may become increasingly interesting alternatives

when explainability and computational cost are considered important for a given application. In this context, we also note that the cost of characterizing sensitivity to randomness depends on the distribution of explanations. For example, as the variance of the attention given to a word in Figure 2.2 increases, more explanations have to be generated to accurately estimate its average attention.

Based on this observation, and although several recent studies have proposed the insertion of theoretical features into transformer models to improve their potential for explainability (Koufakou et al. 2020; Polignano et al. 2022), we propose, conversely, to enrich our model based on linguistic features with a set of new features extracted from explanations derived from a transformer model. At the very least the success of this hybrid approach would suggests that large language models can be valuable (i.e., they provide non uniformly random explanations) in exploratory tasks such as identifying working hypotheses to be confirmed by inductive analysis. On the other hand, it leaves open the fundamental problem of the explainability of large language models.

2.7 Enriching the feature-based model

We now add to the feature-based model LING-LR new features found through the explanations of the fine-tuned CamemBERT models. This hybrid model will be referred to as the enriched linguistic model LING-LR-E. To illustrate the potential for enrichment with reasonable computational time, we limit the number of equivalent models used to 10.

The first step consists in measuring the average relevance attributed to each token occurring at least ten times in our 1,000 texts test set. This restriction helps to avoid considering too specific words. We then rank all tokens based on their average relevance (across all their occurrences for a given seed), in descending order. This operation is repeated for the 10 models studied. We keep only the tokens appearing among the first hundred tokens with the highest relevance for at least five equivalent models out of ten. Then, we qualitatively analyze the final list of most discriminant tokens in order to identify linguistic patterns which can be converted into features. This method is similar to the approach presented by (Yilun Zhou et al. 2022), which consists in deriving information about the internal reasoning of a complex model from local explanations of its predictions. We restrict our analysis to the fifty tokens which are

the most relevant for each class (*opinion* and *news*). The two resulting lists are presented in table 2.4 and 2.5.

News	
précise (precises)	*indiqué* (indicated)
jeudi (Thursday)	*mardi* (Tuesday)
indique (indicates)	*expliqué* (explained)
mercredi (Wednesday)	*explique* (explains)
lundi (Monday)	*vendredi* (Friday)
précisé (specified)	*a-t-elle* (did she)
poursuit (continues)	*souligne* (underlines)
adaptation	*souligné* (underlined)
ajouté (added)	*dimanche* (Sunday)
parking	*poursuivi* (pursued)
conclu (concluded)	*AFP*
suspect	*correctionnel* (correctional)
ajoute (adds)	*aéroport* (airport)
samedi (Saturday)	*affirmé* (affirmed)
trafic (traffic)	*assuré* (assured)
locales (local)	*affirme* (affirms)
chanteuse (singer)	*selon* (according to)
priorités (priorities)	*températures*
déclaré (declared)	*a-t-il* (did he)
disponible (available)	*rappelé* (recalled)
février (February)	*ordinateur* (computer)
communiqué (statement)	*organisateurs* (organizers)
blessé (injured)	*km*
GMT	*pourront* (will be able)
dépistage (screening)	*visiteurs* (visitors)

Table 2.4: List of the 50 tokens (from left to right then top to bottom) with the highest average relevance in the explanation maps derived from the **News** predictions made by the fine-tuned CamemBERT models for the *RTBF-InfOpinion* test set.

Our qualitative examination of the fifty tokens in the *opinion* list reveals several recurring patterns: axiological words (*disaster*, *poor*), expressive punctuation marks (“. ..”, “!”), verbs of thought (*imagine*, *forget*), words referring to abstract concepts (*ideology*, *humor*), or discourse markers (*in short*, *of course*). Regarding *news*, we can identify words referring to non-deictic temporal entities (*lundi*, *GMT*), verbs of quotation (*precise*, *affirmed*), words with a high subjective frequency (*computer*, *airport*), i.e., words perceived as frequent in

everyday language (Balota et al. 2001), and words referring to sources of information (*according to*, *AFP*). Some of these patterns overlap with features already present in the LING-LR model, such as the number of adjectives and expressive punctuation marks (for the *opinion* class), while others represent original findings, such as the presence of discourse markers and the average concreteness of words in the text (Bonin et al. 2018). Finally, we note that some tokens can be considered as artifacts (Gururangan et al. 2018) related to the data used (*parking*, *Flemish*).

Opinion	
révélations	*fed*
chômeurs (unemployed)	*imaginez* (imagine)
bref (in short)	*désigne* (designates)
ressemble (resembles)	*pension*
fout (does)	*extension*
...	*latin*
formateur (instructor)	*aiment* (like)
Hitler	*inverse* (reverse)
tort (harm)	*disons* ([we] say)
aujourd'hui (today)	*ombre* (shadow)
accords (agreements)	*bulle* (bubble)
démontrer (demonstrate)	*mélange* (mixture)
politiquement (politically)	*libéral*
désastre (disaster)	*dépit* (spite)
flamandes (Flemish)	*chômage* (unemployment)
suffisamment (sufficiently)	*correction*
pire (worse)	*illustre* (illustrates)
retraite (retirement)	*idéeologie* (ideology)
médiatique (media-related)	*immobilier* (real estate)
voyons ([we] see)	*;*
!	*oublier* (forget)
pauvre (poor)	*monétaire* (monetary)
certes (of course)	*utilise* (uses)
mauvais (bad)	*impôt* (tax)
calcul (calculation)	*inutile* (useless)

Table 2.5: List of the 50 tokens (from left to right then top to bottom) with the highest average relevance in the explanation maps derived from the **Opinion** predictions made by the fine-tuned CamemBERT models for the *RTBF-InfOpinion* test set.

This analysis allowed us to extract 9 new linguistic features from the patterns identified in the attention lists of Table 2.4 and 2.5: the ratio of deictic temporal markers, non-deictic temporal markers, thinking verbs, quoting verbs, passive verbs, and discourse markers, as well as concreteness, imageability and the average subjective frequency of words in the text. Concreteness is measured using Bonin et al. (2018)'s lexicon, while imageability and subjective frequency are measured using the lexicons of Desrochers and Thompson (2009). The enrichment with these nine new features (added to the nineteen features of the original LING-LR model) gives LING-LR-E an accuracy of 89.6 % on the RTBF-InfOpinion test set, an increase of 0.8 % compared to LING-LR (which is not statistically significant). On the LeSoir-InfOpinion test set[2], LING-LR-E achieved an accuracy of 80.6 % compared to 76.8 % for LING-LR, an increase of 3.8 % (significant according to the same Z statistic and the same p-value of 1.96 as in the previous sections). It shows that the elements extracted from the explanations of the CamemBERT models have contributed to making the LING-LR-E model more generalizable. For comparison, the best accuracy obtained by a CamemBERT model on the LeSoir-InfOpinion test set is 90.5 % vs. 96.6 % on the RTBF test set.

2.8 Chapter's conclusion

In this first chapter, we highlighted and investigated a question that appears to have been under-studied at this stage (*Is the sensitivity of explanations to the randomness of large language models significant ?*), even though it crystallizes the essential difficulty of explaining the predictions of large language models. In particular, we have shown that transformers fine-tuned on the same training set but with different randomness in the form of a seed, although they yield similar accuracies, can provide different explanations for texts on which they give the same prediction. In other words, we have observed explanations that depend on the structure of the models generated from different random parameters, rather than on the outcome of their predictions. Since there is no reason to prefer one model over another in this context, we conclude that restricting ourselves to the explanation of a single model is insufficient. We assert that explaining the decisions of this type of model requires a characterization of their explanation distribution as well. However, while we claim that characterizing

[2]This dataset was built using the same step as for the first one, but using texts from a different media as described in Section 1.7

the explanation sensitivity to randomness is a necessary step when performing explainability evaluations, we do not claim that the observed variability is forcibly detrimental for other desirable explanations criteria such as their plausibility, only that it should at the very least be taken into consideration.

Having found that an initial visual characterization of explanation sensitivity to randomness using box plots reduces their minimality (Miller 2019), we then evaluated the extent to which a simpler model allows for more compact and more stable explanations. First, we trivially showed that a model based on linguistic features combined with logistic regression produces explanations that do not exhibit sensitivity to training randomness, but that it was at the cost of having a reduced accuracy. We observed, thanks to the attention maps derived from the two models, that they do not rely on the same tokens for their predictions. We therefore tried to enrich the linguistic model by extracting new features from the attention maps computed on the fine-tuned CamemBERT model (the most accurate model) and integrating them into the logistic regression model based on linguistic features (the most explainable model). This approach improved the generalization of the model without reducing the stability of its explanations. It follows that despite this sensitivity to randomness, CamemBERT explanations still contain useful pieces of information. Our claim is therefore not that these explanations should not be used at all, but rather that their level of randomness should be quantified. The higher it is, the more models should be trained in order to get a representative distribution of the explanations they can produce. We will investigate this sensitivity to randomness quantification question in the next chapter.

Since the enriched linguistic model LING-LR-E still shows a lower accuracy than the fine-tuned CamemBERT one, we also conclude that large language models characterize textual features, useful for classification, that are currently not (and probably cannot be) integrated into the feature-based model. The fundamental problem of the explainability of the decisions made by large language models thus remains open, magnified by the randomness dependency we just highlighted. Our results therefore suggest new avenues of research aimed at identifying the origin of this sensitivity to randomness and reducing it. One could assess the extent to which this sensitivity to randomness is due to a lack of faithfulness of the explanations. In particular, we could hypothesize that methods which provide simple explanations, for example based on tokens (as those used in this paper), are not adapted to the multiplicity of features exploited by large language models, and that this mismatch between a complex

model and simple explanation formats increases the sensitivity of the outputted explanations to randomness. To test this hypothesis, it would be necessary to assess the extent to which a model simpler than CamemBERT (e.g., with fewer parameters) appears less sensitive to randomness. Chapter 5 partially tackles this question.

Another avenue of research would consist in assessing the impact of the highlighted explanations sensitivity to randomness on plausibility. Indeed, asserting that the dependence of a model's explanations on randomness must be characterized does not necessarily imply a reduction in plausibility. In particular for the task studied, it could be observed that the variability observed reflects the variability of the explanations given by human annotators. It would therefore be interesting to set up such a human annotation experiment. Chapter 4 focuses on this aspect. Investigating the extent to which explanations of equivalent models can be grouped into clusters would also be particularly relevant in this perspective.

Finally, one could consider to improve the faithfulness of the explanations by adapting the methods and formats (see Chapter 3 for a discussion about explanation formats) used for model explanation. However, the potentially higher complexity of such formats would quite logically come as a trade-off with respect to their plausibility. Reducing the impact of training randomness could simplify this problem. However, this direction was not followed during this thesis and further research would be needed to study these ideas.

Chapter 3

How to Quantify Explanation Sensitivity to Randomness ?

The content of this chapter was published in (Bogaert and Standaert 2024a) and (Bogaert, Descampe, et al. 2025)

3.1 Motivations

The main observation of the last chapter is that it is sometimes possible to produce many models of which the training only differs by the (indistinguishable) random seeds they use, that are "equivalent" from the accuracy viewpoint and nevertheless lead to different explanations.[1] We argued that this sensitivity to the training randomness must at least be characterized, since in the extreme case where the explanations would be uniformly distributed, any selection of explanation would be completely arbitrary.

Evaluating the explanation sensitivity to the training randomness of LLMs can be done qualitatively. For example, visualization tools like box plots, as used during the last chapter, provide a good intuitive understanding of single texts.[2] Yet, more quantitative tools become useful to explore the explanations' design space. For example, one could be interested to compare the explanation sensitivity to randomness of different texts, and for explanations assigning

[1] Equivalent meaning that there is no statistically significant difference in their accuracies, implying that there is no "better" model from the accuracy viewpoint.

[2] Note that they are increasingly impractical to use when working with long texts

relevance scores to various numbers of words. One could also be interested to compare bigger vs. smaller models, for various tasks, datasets, languages or explanation methods.

As a natural next step, this chapter therefore aims to try pushing our initial investigations towards a more quantitative view. For this purpose, we first introduce a taxonomy of explanations in Section 3.2. It illustrates that the explanations' sensitivity to randomness has for now been exhibited in the case of "simple" explanations, which are of course not expected to be perfectly faithful, although we assume they reflect the models' reasoning to a sufficient extent. They are not expected to be the only plausible ones either. Yet, they provide a useful theoretical framework to answer the question: *how stable can the simple explanations of complex models be?* Note that, in our case, since all models use the same parameters and data aside from the randomness, and since the LRP explanation method is deterministic, we can directly map the explanations' instability with the explanations' sensitivity to the training randomness.

The quantitative evaluation of explanation stability is quite related to the problem of inter-annotator agreement – see for example (Hayes and Krippendorff 2007; Artstein 2017). One key difference is that LRP explanations provide continuous relevance scores where human annotations are discrete, invalidating some of the potential metrics from this field such as Cohen's and Fleiss' Kappas (McHugh 2012). While some measures such as Krippendorff's alpha (Hayes and Krippendorff 2007) can tackle continuous relevance scores, they rely on pairwise comparisons to provide a final metric. We however would like to have a metric that scales linearly (instead of quadratically) with the amount of explanations, as we could have to compute it based on a lot of equivalent models explanations, for a lot of texts and for multiple combinations of other parameters of interest (e.g., with different levels of explanation post-processing), which could end up being somewhat computationally intensive. We note that our study is also partially related to (Wu and Ong 2021) which, among others, performed an experiment to test whether the words' relevance obtained thanks to four different types of explanations were impacted by the random seeds used for model initialization. They used Pearson's correlation for this purpose, but only considered two random seeds where we can potentially have a very high number of them. On top of it, using pairwise Pearson correlations would lead to the same complexity issue as for Krippendorff's alpha.

To reach our goal, we first try to exploit information theoretic metrics in the name of the Signal, Noise and Signal-to-Noise Ratio (SNR). We then make use of these candidate metrics to illustrate what a good explanation stability metric should focus on. Following this principle, in section 3.6, we highlight the limited ability of the SNR to reflect the relative differences of relevance (between words) in a set of explanations corresponding to different (random) training seeds. We additionally show that a correlation metric mitigates this issue, at the cost of being unable to reflect absolute differences of relevance (for single words), which are better captured by the SNR. We continue by putting forward additional desirable features of the correlation metric such as easier interpretation, unbiased estimation thanks to cross-validation and simple confidence intervals thanks to a well-known statistical distribution.

Finally, we discuss the consequence of these observations and argue that combining a design space exploration with the correlation metric and a more qualitative approach thanks to word level noise analysis (e.g., thanks to box-plots) appears as a good trade-off. The first one better captures relative differences within explanations, whereas the second one reflects absolute differences at the individual word level, exhibiting possibly interesting intuitions that a quantitative metric only may hide.

3.2 Explanation taxonomy and plausibility assumption

We next define the explanation format we are using and give a brief overview of possible alternatives. First, explanations for the classification of n-word texts can display a relevance for every word independently or for t-tuples of (ordered or un-ordered) words (i.e., display the importance of combination of words). Then, they can be univariate (i.e., provide one relevance value per word or tuple of words) or d-variate (i.e., provide d such values per word or tuple of words). Finally, in case these explanations show a sensitivity to the training randomness, one can characterize this sensitivity with first-order methods (i.e., consider only the means of the explanations' distribution as informative) or higher-order methods (i.e., use higher-order moments of the explanations' distribution).

As a working hypothesis, we preliminarily assume that in case of variability, only the mean of word-level univariate explanations or $(1, 1, 1)$ explanations can be understood by a target audience, which we next denote as the

word-level univariate first-order plausibility assumption, or $(1, 1, 1)$ plausibility assumption. Word-level means that all m explanations of an n-word text display a weight for each word independently. Univariate means that each of these weights is a single value. First-order means that variable explanations are summarized by their mean (i.e., a first-order statistical moment). An example of an explanation format that does not respect this assumption and is therefore not considered as understandable by a target audience under it would be displaying the attention between two words in each of the h attention heads of the model. It corresponds to $(2, h, o)$ explanations where o is the order of the moment representing the explanation distribution (one for the mean, two for the variance, ...). It is important to note that this working hypothesis, while intuitive, is not based on any human experiment nor literature. We will come back in conclusion on possible further work that could back up this assumption.

Based on this taxonomy, it appears that the explanations obtained in the form of saliency maps are $(1, 1, o)$ explanations. To characterize $(1, 1, o)$ explanation sensitivity to randomness, one can construct an explanation matrix A of m rows (corresponding to different random seeds) and n columns (corresponding to different words), where each $a_{s,w}$ corresponds to the relevance value assigned by the s-th model (seed) to the word at the w-th position, as showed in Figure 3.1. The left part of the figure additionally shows the average curve which corresponds to the only way to represent variable explanations for it to be understandable by human beings under the $(1, 1, 1)$ plausibility assumption.

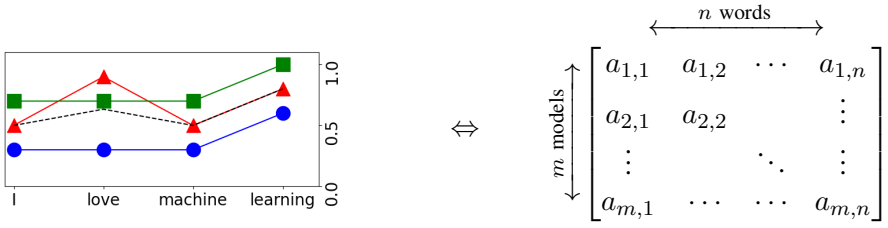


Figure 3.1: Explanations of a 4-word texts for 3 seeds with mean explanation (dotted).

Finally, to capture the possibility that shorter explanations are more plausible, one can also evaluate so-called k-word explanations, obtained by keeping only the $0 \leq k \leq n$ highest relevance values of each explanation. To

further simplify the individual explanations, one can even use k-word binary explanations, where only the k top words are considered relevant, without any distinction among them:

$$a_{s,w} = \begin{cases} a_{s,w} & \text{if } a_{s,w} \in TOP_k(a_{s,:}), \\ 0 & \text{oth.} \end{cases} \quad a_{s,w} = \begin{cases} 1 & \text{if } a_{s,w} \in TOP_k(a_{s,:}), \\ 0 & \text{oth.} \end{cases}$$

3.3 Information Theoretic Metrics

Inspired by information theory (Cover and Thomas 2006), we suggest a first quantification of explanation stability based on Signal-to-Noise Ratio (SNR). Such a metric allows to evaluate the ratio of relevant and useful information (the signal) over irrelevant information due to variability (the noise). It therefore helps to grasp to what extent the explanations' variability erodes their statistical informativeness.

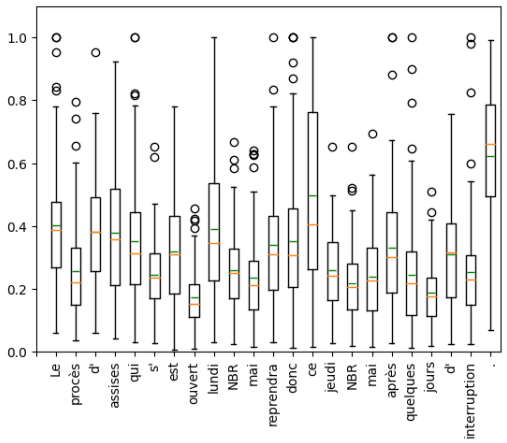


Figure 3.2: Explanations' box-plot for a transformer model. The green and orange dashes respectively show the mean and the median of the relevance distribution for each word. Words with a mean relevance value far (resp., close) from 0 are attributed more (resp., less) relevance in the explanations in general. Tight boxplots (e.g., "ouvert") show a low variability in the relevance attributed by the explanations of equivalent models. On the contrary, wide boxplots (e.g., "ce") show a high variability.

Our limitation to word-level, univariate and first-order explanations, formalized as the $(1, 1, 1)$ plausibility assumption, implies that the only parts of Figure 3.2 that is assumed to be understandable by a human reader are the green dashes that represent the mean attention values. Those values are given by $\hat{\mathsf{E}}\big(a_{s,w}\big)$, with $\hat{\mathsf{E}}$ the sample mean computed over the 100 models so that it is a n-element vector. Under this restriction, we assume that explanations are statistically more informative when words relevance across a text are distributed unevenly, leading to a definition of signal as the variance of the words' relevance means:

$$S = \underset{n \text{ words}}{\hat{\mathsf{Var}}} \left(\underset{m \text{ seeds}}{\hat{\mathsf{E}}} \big(a_{s,w}\big) \right), \tag{3.1}$$

with $\hat{\mathsf{Var}}$ the sample variance computed over the n words of the text, and a definition of noise as the mean of the words' relevance variances:

$$N = \underset{n \text{ words}}{\hat{\mathsf{E}}} \left(\underset{m \text{ seeds}}{\hat{\mathsf{Var}}} \big(a_{s,w}\big) \right), \tag{3.2}$$

Intuitively, the signal reflects the flatness of the average explanation and the noise reflects the averaged variation of the relevance scores. The SNR is simply defined as the ratio between both.

$$SNR = \frac{S}{N}\cdot \tag{3.3}$$

In case of perfectly stable explanations, the noise is null and only the signal can be computed as $\hat{\mathsf{Var}}(A)$. Since neither the noise nor the signal can be negative, it results in the SNR being in the range [0;∞].

All these quantities being computed from a limited number of words and models, we additionally evaluate the quality of our estimates. For the signal of the feature-based model we use a standard 95% confidence interval for the variance. For the signal, noise and SNR of the transformer-based model, we use a 95% bootstrap confidence interval (Efron and Tibshirani 1993), which allows dealing with the fact that these quantities are means or variances of estimated vectors.

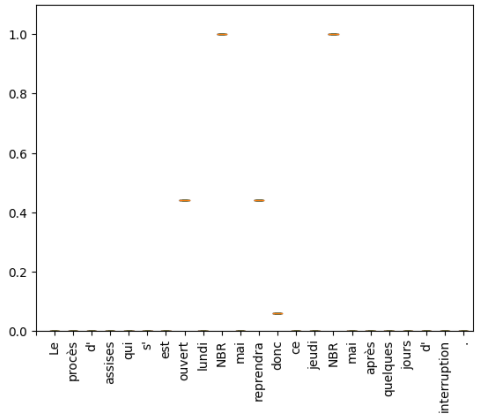


Figure 3.3: Explanation's box-plot for a feature-based model. As there is no variability in the relevance attributed by the feature based model, each boxplot is a simple line. Words that are not used in any feature have an attention score of 0.

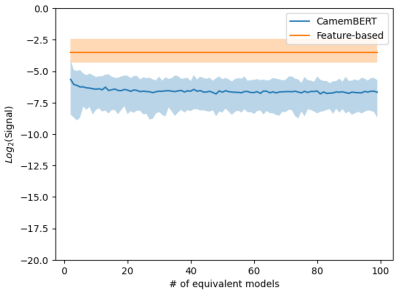
3.4 Experimental setting

We compare the CamemBERT explanations' signal and SNR with the one of our enriched feature-based model (cf. Section 2.7) for which we used the aforementioned linguistic attention maps (cf. Section 1.6) as our explanation method. We explained 20 different texts from our test set: 10 short ones (with less than 50 words) and 10 long ones (with more than 400 words). Each of these texts are compatible (i.e., they are assigned the same class by all the models). As a result, we gather 20 $m \times n$ matrices of explanations $\boldsymbol{A}$ corresponding to the $m = 100$ equivalent CamemBERT models explanations for each of the 20 n-word texts (where n can differ from one text to another). We also obtain 20 n-element vectors based on the linguistic attention maps explanations of the feature-based model on each of the 20 texts.

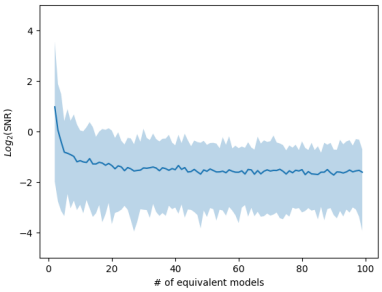
3.5 Experimental results

We now discuss the main observations that can be extracted from the estimation of the signal and SNR given in Figure 3.4 for exemplary texts.

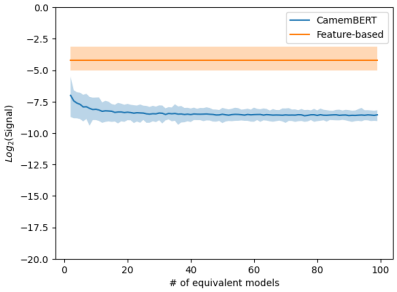
Starting with Figures 3.4a and 3.4b vs. 3.4c and 3.4d which represent the signal of both the CamemBERT and feature-based models and the SNR of the



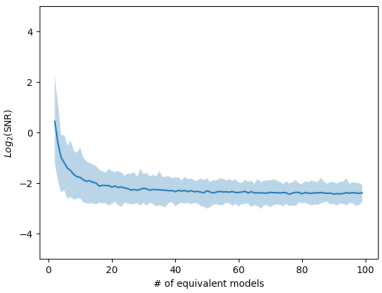
(a) Short text, raw explanations, signal.



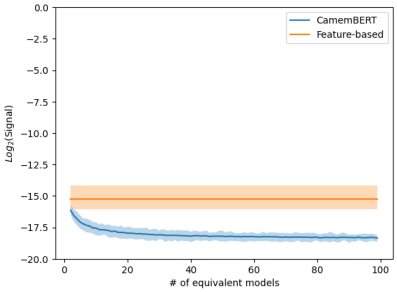
(b) Short text, raw explanations, SNR.



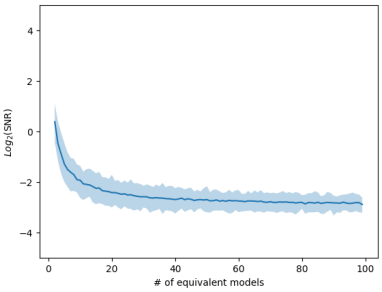
(c) Long text, raw explanations, signal.



(d) Long text, raw explanations, SNR.



(e) Long text, normalized explanations, signal.



(f) Long text, normalized explanations, SNR.

Figure 3.4: Metrics estimation for illustrative texts.

CamemBERT models for a short and long text, a preliminary observation is that the confidence intervals naturally get tighter with longer texts and more equivalent models, allowing to improve the estimation of the signal, the noise and the SNR.

Following with the signal estimations in Figures 3.4a and 3.4c, we can observe two main trends. First, the signal of the CamemBERT models decreases when the number of equivalent models used in its estimation increases. This is because as this number of equivalent models increases, the aggregated explanations tend to cover more words, as observed in Figure 3.2, and their average tends to be flattened. Second, the signal of the CamemBERT models is always lower than the one of the feature-based one. This essentially suggests that when limited to a simple (first-order) analysis of simple (word-level, univariate) explanations, as formalized by the $(1, 1, 1)$ plausibility assumptions, the simpler feature-based model tends to provide more informative explanations than the CamemBERT ones, in the statistical sense captured by our definitions of signal and noise.

In order to try reducing the impact of the more disparate assignment of weights in the explanations of the CamemBERT models with the LRP method, we additionally tried to post-process them. For example, Figure 3.4e shows the signal of normalized explanations, where we restrict CamemBERT's explanations to k-words ones, with k being the number of words with non-zero weights in the feature-based model explanations. We also ensure these weights sum to one. Both of these processes reduces the faithfulness of the CamemBERT explanations. Compared to Figure 3.4c, this naturally reduces the explanations' signal for all models and makes them closer. Yet, the average signal of the CamemBERT models remain substantially lower.

Eventually, the easiest to interpret observations are obtained from the SNR in Figures 3.4b and 3.4d, which are both reaching values below one. It follows that the statistical informativeness of the explanations is lower than their noise. The concrete value (≈ 0.25) also gives an indication of how many equivalent models must be produced in order to reach a good estimation of the signal. This averaging requirement is increasingly undesirable for models of which the optimization is computationally-intensive.

3.6 Discussing our candidate metrics

To perform our study, we used a statistical definition of the signal metric. This definition could possibly not align with more semantic definitions, making the results harder to interpret for human beings. We therefore next discuss the adequacy of the SNR metric to reflect the explanations stability. For this purpose, we use illustrative hand-made examples and compare how the stability of

some explanations is captured by the SNR and by an alternative simple metric, namely the **M**ean **C**orrelation **W**ith a **M**ean **E**xplanation (MCWME). We are in particular interested in the ability of these metrics to reflect the relative differences of relevance between words and the absolute differences of relevance for single words in a set of variable explanations.

3.6.1 Relative differences (between words)

Figure 3.5 illustrates two pairs of explanations such that the relevance of some words are swapped when going from the left to the right plots. As a result, these left and right plots show quite disparate relative differences of relevance between words. Interestingly, the SNR metric is unable to reflect these relative differences. This is because the swaps do not affect the mean explanations (which are the same on the left and right plots, leading to the same signal) nor the absolute difference between words (hence the noise). By contrast, the correlation metric captures these relative differences: the explanations of the left plot are highly correlated with the mean explanation; the ones of the right plot are not.

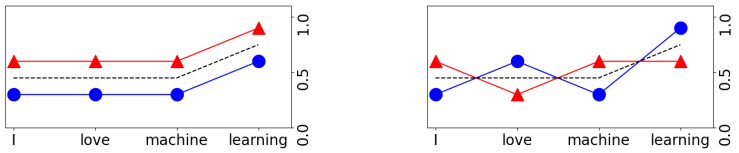


Figure 3.5: Two pairs of explanations (△ and ◯), with the same absolute differences and different relative differences, with the mean explanation in dotted line.

3.6.2 Absolute differences (for single words)

A complementary situation is illustrated in Figure 3.6, in which an offset δ was added to all the relevance values of one explanation and subtracted for the other. As a result, the left and right plots show disparate absolute differences. This time, the correlation metric is unable to reflect the discrepancy between the left and right plots (because the correlation is invariant to the δ offset). By contrast, the SNR reflects it because the noise of the left and right plots differs.

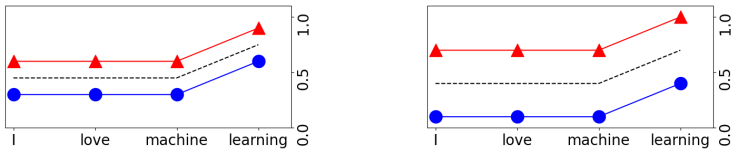


Figure 3.6: Two pairs of explanations (△ and ◯), with the same relative differences and different absolute differences, with the mean explanation in dotted line.

3.7 Additional limitations of the information theoretic metrics

The two examples above suggest a quite natural tradeoff between the SNR and correlation metrics: the first one better captures absolute differences, the second one better captures relative differences. While this may encourage using both metrics in parallel, we next show additional limitations of these information theoretic metrics:

3.7.1 Implicit assumption of the signal metric

As illustrated in Figure 3.7, it is possible to obtain explanations such that their absolute and relative differences are identical, but their signal differs, because the signal is focused on the flatness of the mean explanation. It implicitly suggests that more relative differences within this mean explanation lead to better explainability, which may not be connected to a definition of plausibility.

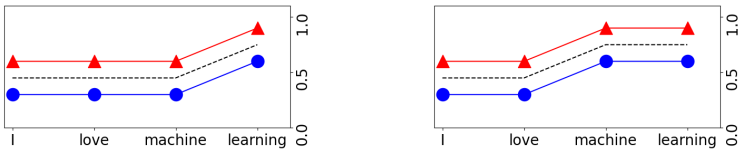


Figure 3.7: Two pairs of explanations (△ and ◯), with the same absolute and relative differences but different signal, with the mean explanation in dotted line.

3.7.2 Noise aggregation

Figure 3.8 highlights additional limitations of the noise component of the SNR metric. Namely, it illustrates that the noise is averaged over (possibly dependent) words, which may hide important intuition regarding which word is causing the noise. By contrast, the correlation can be averaged over independent seeds.

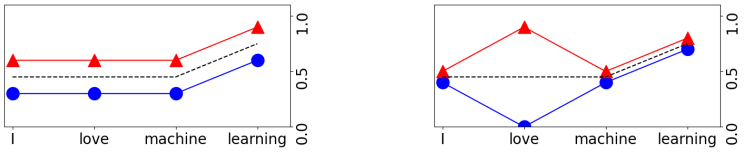


Figure 3.8: Two pairs of explanations (△ and ◯), with the same (average) absolute differences but distributing these differences differently among the words of a sentence.

3.7.3 Suitability for design space exploration

As illustrated in Figure 3.9, we cannot use the signal or the noise alone to explore our design space, as their range is directly impacted by the amount of top words k. This is the case even for deterministic/random explanations, which leads to an hypothesis that selecting a certain ratio of word leads to more stable explanations, even if all the models perfectly agree on the relevance of every token. This is not the case for the correlation metric that is always at its maximum for deterministic explanations, and close to 0 for fully random ones.

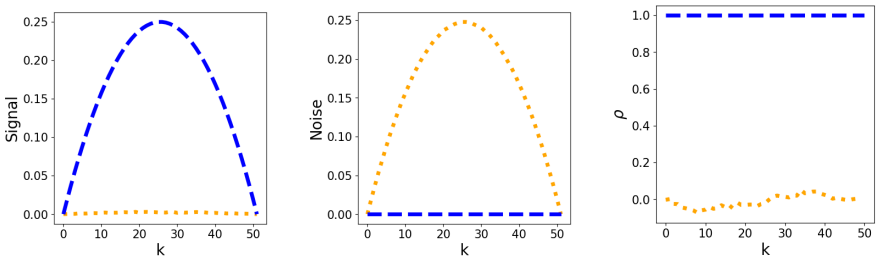


Figure 3.9: Signal (left), noise (middle) and correlation (right) of deterministic (-) and random (.) explanations, for the binary variant of k-word explanations.

3.8 Estimation and confidence interval

The SNR is also slightly less convenient to manipulate from the statistical viewpoint. First, it is a biased metric since small estimation errors in the mean explanations are considered as signal by definition. Second, its interpretation in case of small noise levels is not always intuitive as it tends to infinity when the noise tends to zero. Despite these drawbacks do not lead to fundamental issues (i.e., the SNR bias decreases with the amount of seeds and can be corrected, intuition is just less direct), we show that the MCWME also comes with advantages in this respect. It can be estimated without bias thanks to cross-validation, benefits from a well-known sample distribution leading to easy-to-obtain confidence intervals and its interpretation is direct.

To compute it, one first builds a mean explanation by averaging all but one of the explanations. The correlation between this average explanation and the remaining one is then computed to obtain a correlation with the mean explanation. The process is then repeated m times, leaving out a different explanation each time. Figure 3.10 shows a scheme of this process.

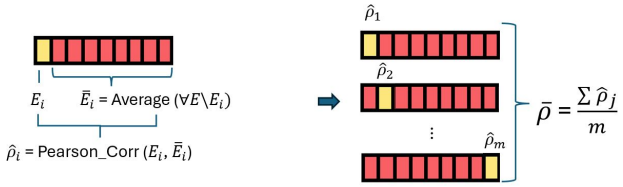


Figure 3.10: Scheme of the MCWME computation

Different quantities of the correlation distribution could then be considered to summarize the explanations' stability. Since simple (word-level, univariate and first-order) explanations naturally suggest simple quantities to capture their sensitivity to randomness, we suggest to use the average correlation. The Mean Correlation With Mean Explanation (MCWME) can therefore be interpreted as a measure of the approximation quality of only showing the average explanation, which, under the (1,1,1) plausibility assumption, is the only one understandable by human beings. We note that while it is in general better to estimate the correlation between two variables based on a large set of samples than averaging correlations estimated from several smaller sets of samples, this approach can serve as a useful heuristic in our context, if interpreted carefully. Namely, as a way to capture a global tendency for many explanations, possibly

leading to different correlation values. In order to provide a confidence interval for the metric, we follow (Silver and Dunlap 1987) and first use the following "Fisher Z transformation":

$$F(\hat{\rho}) = \frac{1}{2}\ln(\frac{1+\hat{\rho}}{1-\hat{\rho}}) = \operatorname{arctanh}(\hat{\rho}),$$

which projects the correlation samples in a space where they are normally distributed. We can then compute the average Fisher value $\bar{F}$, as well as its sample variance $\hat{\sigma}^2$. A 95% confidence interval on the estimation of $\bar{F}$ is obtained by adding or removing $2\frac{\hat{\sigma}}{\sqrt{m}}$ to $\bar{F}$. Applying the inverse function $\rho = \tanh(F(\rho))$ finally leads to 95% confidence interval for the average correlation:

$$\left[\tanh\left(\bar{F} - \frac{2\hat{\sigma}}{\sqrt{m}}\right); \tanh\left(\bar{F} + \frac{2\hat{\sigma}}{\sqrt{m}}\right)\right].$$

This interval indicates that the average correlation is better estimated with more models (i.e., large m values). By contrast, longer texts (i.e., large n values) lead to better estimated correlation samples, but do not necessarily decrease the variance $\hat{\sigma}$, since the correlations of different explanations may differ.

3.9 Application to case studies

The examples of Section 3.6 suggest using the correlation of different explanations with their mean as a good way to quantify explanation sensitivity to randomness. Based on these previous observations, we suggest using it for design space exploration and adding box plots for a qualitative analysis of the noise. As will be experimented next, this appears as a relevant combination to characterize the explanations' sensitivity to the training randomness, capturing both the absolute and relative differences within these explanations. Since our feature-based model is fully stable, it would lead to a MCWME value of 1 all the time, not adding much to the comparison. We therefore compare the fine-tuned and frozen (c.f. Section 1.2) versions of CamemBERT instead as they can both show sensitivity to randomness.

3.9.1 Quantitative analysis

Figure 3.12 shows the evolution of MCWME for k-word explanations. Positing that shorter and more aligned explanations are more plausible, such an exploration can lead to identify relevant parameters to investigate more qualitatively. For example, we can see on the left plot that the MCWME increases up to $k = 7$ and then reaches a plateau. Hence, larger values of k (i.e., explanations with more words being assigned a non-zero relevance) may not lead to a reduced sensitivity to the training randomness. We next complete this observation with a qualitative analysis for the explanations obtained for $k = 7$ and $k = 51$ (i.e., the maximum value for k). Figure 3.11 shows a scatter plot of all the correlations to the mean (i.e., one per trained model, so 100 in our case study), highlighting the high disparity of the results depending of the training randomness.

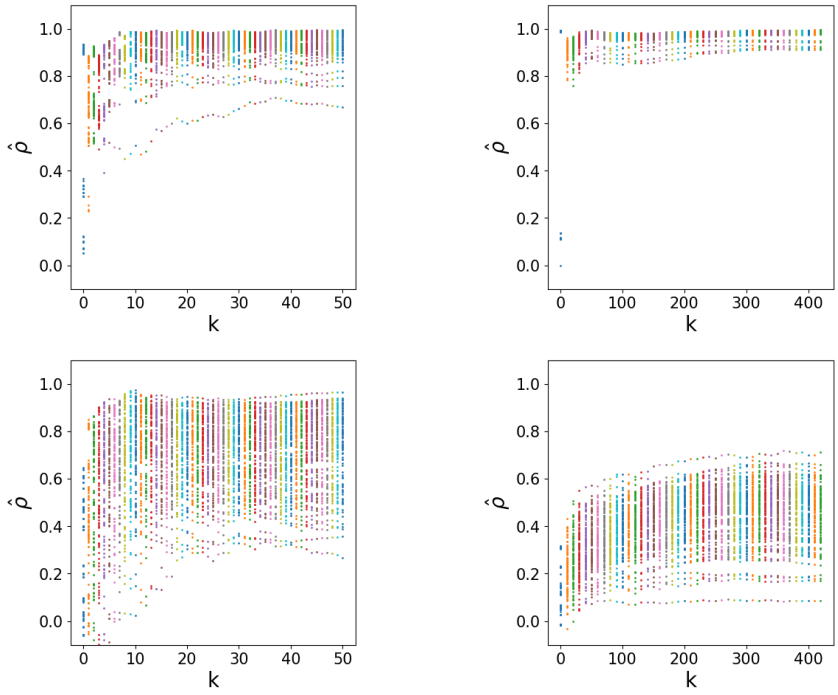


Figure 3.11: Scatter plot of the correlation to the mean for LRP explanations corresponding to the *frozen* (top) and *fine-tuned* (bottom) models w.r.t. the number of top-words considered in each explanation, illustrated for a short (left) and a long (right) text. For readability, only the values $k = 10, 20, 30, ...$ are displayed for the long text.

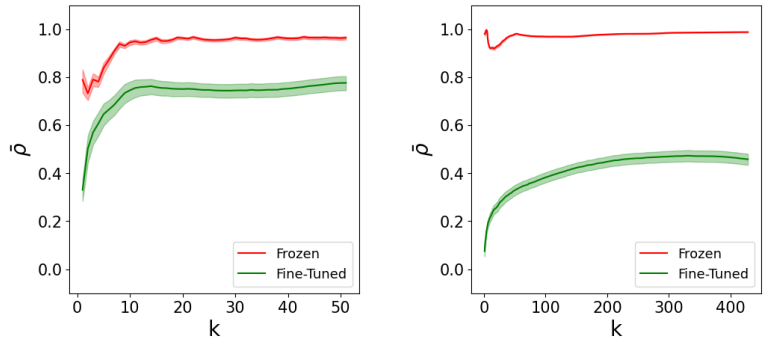


Figure 3.12: MCWME (with confidence intervals) for LRP explanations corresponding to the *frozen* and *fine-tuned* models w.r.t. the number of top-words considered in each explanation, illustrated for a short (left) and a long (right) text.

3.9.2 Qualitative analysis

Starting with the box plot for $k = 7$ displayed on Figure 3.13, we can observe that, qualitatively as well, the LRP explanations of the frozen model are significantly less sensitive to the training randomness than the ones of the fine-tuned model. What is maybe less expected is that the variability per word is also distributed very differently for both models. Namely, 7-word explanations across the 100 seeds only consider 10 different words in the frozen case, while most words are considered relevant by the fine-tuned models. This tends to justify our proposed methodology, where we do not analyze the absolute difference with the noise metric (which is averaged over the words).

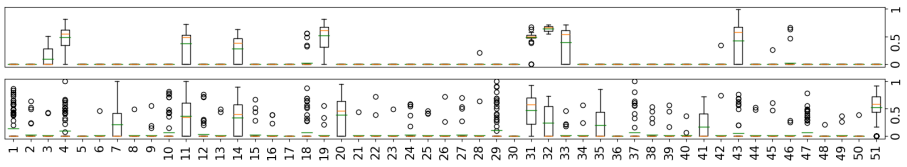


Figure 3.13: Box plot for $k = 7$ and the *frozen* (top) and *fine-tuned* (bottom) models.

More interestingly, Figure 3.14 shows the box plots obtained for the same models and $k = 51$. Its upper part is particularly relevant: it confirms that

increasing the explanations' length beyond $k = 7$ is not only discouraged by the correlation metric, it actually also leads to harder to interpret first-order explanations assigning non-zero relevance scores to most words, as the fine-tuned model.

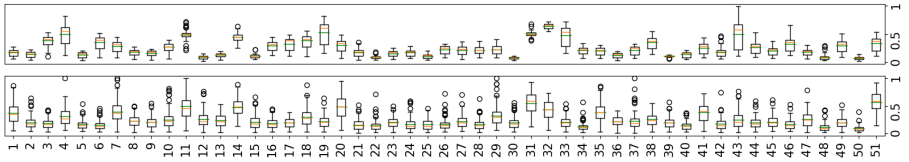


Figure 3.14: Box plot for $k = 51$ and the *frozen* (top) and *fine-tuned* (bottom) models.

3.10 Chapter's conclusions

In the first part of this chapter, we highlighted that, if limited to simple - formalized as $(1, 1, 1)$ - explanations, simple models may carry more statistical signal (i.e., produce an average explanation assigning relevance to fewer words with more contrast) and less noise than transformer-based models like CamemBERT. Unless other simple explanation tools lead to different conclusions, it means that combining such models with as good explanation informativeness as simple ones may require more complex explanations.

We then challenged these findings by discussing the suitability of these information theoretic metrics. First, we showed that the noise and the SNR are more suited to detect absolute than relative differences, but that we had more interest in the latter. We then showed that the aggregation of the noise across different words could lead to harder to interpret results, and that the statistical definition of the signal may not be correlated with what human readers perceive as useful information. We concluded by showing that the Mean Correlation With Mean Explanation (MCWME) metric helped alleviate these flaws, in particular when used in combination with noise visualization in the form of boxplots.

We emphasize that our results are based on the $(1, 1, 1)$ plausibility assumption, as defined formally in this chapter. This assumption could be challenged to investigate whether human readers can understand more complex (t, d, o) explanations considering t-tuples of words, d attention values per word

and an higher-order statistics, which could all contribute to improve the explainability of LLMs. Explanations based on t-tuples of (ordered or un-ordered) words could improve the faithfulness of the explanations and make them more informative since it is likely that the improved accuracy of LLMs is taking advantage of such combinations of words. Explanations based on d attention values per word could have a similar impact, since they would avoid aggregating the weights that different explanation features add to each word. Finally, moving to higher-order evaluations could lead to transform a part of the variance that we currently capture as noise into useful signal. This would for example take place if the variance of the explanations observed was corresponding to a variance of viewpoints, similar to the one that could be observed for human annotators. Investigating this question would be an interesting follow up work, given that our subjectivity detection task would likely lead to different readers having different opinions on which words are important to explain a classification. We study this last question in the next Chapter.

Following that reasoning, it is also possible that even if explanations appear unstable when considering their average correlation to a mean explanation, some clusters exist within these explanations. This would mimic a situation where a few groups of human annotators share very similar explanations within the groups and have very different ones between the groups.

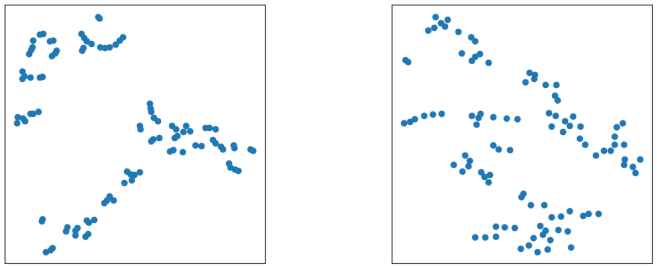


Figure 3.15: TSNE for $k = 51$ and the *frozen* (left) and *fine-tuned* (right) models.

We performed a very little analysis of a TSNE visualization of 100 explanations used in our experiments in Figure 3.15. Despite not seeing clear clusters, we still believe that it would be interesting to investigate whether they could be extracted or created based on the proposed metric. One could then qualitatively evaluate whether explanations in the same clusters are in-

deed more similar from a human point of view than the ones in different clusters. Despite being an interesting research direction, we did not extensively followed it during this thesis. We therefore leave it open for further works, as will be presented in our final conclusion. We also note that our observations may differ for other modalities than texts. For example, a stability metric for image explanations may leverage absolute differences instead of relative ones, due to the more intrinsically correlated nature of adjacent pixels explanations (compared to consecutive words in a text). While we evaluate explanation stability for different tasks in Chapter 5, other modalities than text are also left as a potential further work.

Chapter 4

Comparison with Human Annotators

4.1 Motivations

For now, we have put into light that transformer models trained with different random seeds may produce different explanations. We then discussed the best way to characterize this explanation sensitivity to randomness. In this chapter, using the tools we designed, we aim at comparing this sensitivity with the variation that can be observed across different individuals' explanations. We examines the variability of human and Layer-wise Relevance Propagation (LRP) explanations of the CamemBERT model on a use case of French journalistic text classification between *news* and *opinion* articles. Using a low amount of texts annotated by a high number of annotators and classified by multiple equivalent models, we analyze qualitative patterns and quantitative trends in terms of MCWME.

Some prior works have already compared these two types of explanations. For example, Sen et al. (2020) compared transformer-derived token attention maps with so-called "human attention maps" of the same predicted texts, measuring the degree of overlap between the sets of highlighted tokens. While informative on the textual information captured by humans and models, such approaches overlook our factor of interest: the *variability* of explanations across multiple annotators or across multiple functionally equivalent models on compatible inputs. Specifically, we explore whether the variability observed in the explanations of equivalent models' decisions is to some extent similar to

the inter-annotator disagreement. We treat the raw distribution of explanations (across both multiple humans and models) as the central object of study. This allows us to directly measure and compare the intrinsic variability of explanations in each group.

This chapter's first introduces a systematic framework for measuring the variability of explanations both within and between pools of human annotators and equivalent models. We then use it to quantify the differences in explanation stability across classes and explanation formats, revealing consistent divergence patterns between human and model explanations on the same predictions. We finally show that simple explanation post-processing techniques,i.e., discretization strategies informed by human annotation patterns, can substantially reduce the gap in stability between model and human explanations of the same texts.

4.2 Prior works about Human Annotations' Variability

Many experiments in formal linguistics or NLP rely on human-annotated texts that serve as training data or as a gold standard for evaluating models. Tasks such as sentiment analysis or text classification often depend on labeled datasets, manually annotated by human participants or experts (Wankhade et al. 2022). However, human annotation of linguistic data presents several challenges (Mieleszczenko-Kowszewicz et al. 2023).

Crowd-sourced data may suffer from inconsistent quality due to annotators' varying levels of expertise or dedication to the task (Basile et al. 2021). Moreover, representations, knowledge, and subjectivity of annotators always exert an influence on the results of their annotations (Sap et al. 2021), which may have an impact on the reliability of annotated datasets. Even expert annotators frequently disagree, especially on tasks involving subjective or ambiguous categories (Jiang and Marneffe 2022). This phenomenon is often discussed in terms of inter-annotator agreement (IAA), where low IAA can limit the reliability of the annotated data. More recently, this variability in labels has been formally conceptualized as Human Label Variation, a term popularized by Barbara Plank (Plank 2022). This reframing highlights that the variability of labels assigned by different annotators to the same item is not merely "noise" or a failure of annotation, but rather a valuable piece of linguistic information.

Human Label Variation may reflect the inherent nuance or complexity of a task (Wiebe et al. 2004; Weber-Genzel et al. 2024), or genuine differences in human interpretation that models should potentially be designed to capture.

4.3 Methodology

4.3.1 Data

The sample of texts analyzed in this chapter contains 60 texts selected from the test set (1,000 texts) of the Belgian journalistic texts corpus introduced by (Escouflaire, Descampe, Venant, et al. 2024). All texts are excerpts consisting of the first paragraph or first lines of press articles written in French, published between 2014 and 2024 by four different Belgian media: public service news website *RTBF Actus*, regional daily newspaper *L'Avenir*, and national daily newspapers *Le Soir* and *La Libre*. The texts contain between 69 and 129 tokens and were selected following several criteria: 30 are excerpts of texts considered by the media in which they were published as belonging to the *opinion* class, 30 to the *news* class. Of those 60 excerpts, 40 were predicted by our linguistic feature-based classification model (LING-LR-E) as very close to their corresponding class, having regression scores below 0.2 or above 0.8. The twenty others texts were selected because they appeared to be less trivial to classify using the same model, obtaining a regression score between 0.3 and 0.7.

4.3.2 Human Annotations

We collected manual annotations through an experiment involving 42 human annotators, all native French speakers. They were tasked with reading the journalistic texts, classifying them as either *news* or *opinion* and explaining their decisions by highlighting relevant segments of the text. Of the 42 participants, 26 are master's students studying journalism in Belgium, aged between 21 and 24 (except two students aged 29 and 43). All students willingly participated in the experiment, in the context of a course on information literacy. They were given the choice between participating in this experiment or writing a short essay on journalistic objectivity. Respect of deadlines and serious participation in the experiment were the only explicit criteria for the evaluation of the course. In addition to the 26 students, 16 additional annotators were recruited through social networks. They are 27 years old on average, occupy a variety

of professions (e.g., lawyers, IT specialists, researchers), and have also agreed to take part in our experiment willingly. All 42 participants signed a consent form for their participation in the experiment. The design of the experiment was approved by the ethical board of the Language & Communication institute of UCLouvain.

Annotations were collected using a custom made platform inspired by PADDLe (Pirali et al. 2022). The platform was designed to enable participants to create an account, log in and out, and annotate excerpts at their own pace. For each extract, participants completed three annotation steps, similarly to what was done by (Escouflaire, Descampe, and Fairon 2024): (1) reading the extract and selecting the journalistic genre it belongs to (*opinion* or *news*); (2) indicating their level of confidence in this choice on a scale from 1 (*not confident at all*) to 5 (*absolutely confident*); and (3) highlighting the textual elements that led them to choose one class over the other. In the third step, annotators could use three colors to indicate the importance of each highlighted element: yellow (*slightly important*), orange (*important*), and red (*very important*). Additionally, an optional free-text field allowed annotators to provide further comments or considerations. Figure 4.1 provides an example annotation made by a student for a given text.

The first step of the annotation campaign, involving the 26 student participants, lasted four weeks (from October 21st to November 17th, 2024). Each Monday, students were assigned 15 random texts from the corpus, which they had to annotate by the following Sunday. By the end of the campaign, all 26 students had annotated each of the 60 texts. Only 16 texts were unanimously categorized under the same class by all 26 annotators (6 as *opinion*, 10 as *news*). From these 16 texts, 10 (5 per class) were randomly selected for further annotation. The second step of the experiment, involving the 16 additional participants, lasted between December 10th, 2024, to January 31st, 2025. They were given the 10 selected texts in random order and as much time as they needed to read and annotate them. Most participants annotated all the texts in a single session, a few spread them out over several days. One of the 16 annotators completed only 6 out of 10 annotations, while all others annotated all 10 excerpts. In the end, 6 texts were annotated by 42 different annotators, and 4 texts by 41.

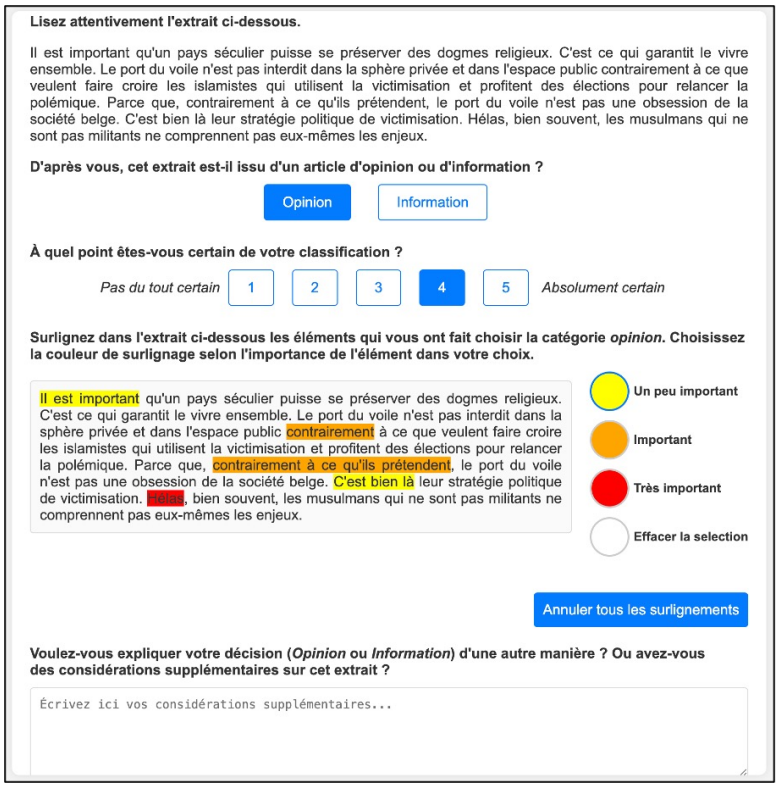


Figure 4.1: Screenshot of an annotation made using the custom interface used for the experiment.English translation: 1) Read the excerpt below carefully. 2) [Excerpt] 3) According to you, is this excerpt from an opinion article or a news article? 4) How confident are you with your classification? [Not confident at all ... Absolutely confident] 5) Highlight in the excerpt below the elements that led you to choose the opinion category. Choose the highlighting color according to the importance of the element in your choice. Keys: Slightly important (yellow), Important (orange), Very important (red), Erase selection (white), Cancel all highlights. 6) Would you like to explain your decision (opinion or news) in another way? Or do you have additional considerations about this excerpt?

4.3.3 Attention-based Explanations

To obtain the model's explanations, we apply the same procedure we previously described. We fine-tune a CamemBERT model on the *opinion* vs.

news classification task using a training set of 8,000 texts, perfectly balanced between the two classes (4,000 *opinion* and 4,000 *news* texts). The hyper-parameters remained at their default values, aside from the learning rate (2×10^{-5}), the batch size (4) and the number of epoch (2). To generate 100 equivalent versions of the fine-tuned model, we replicate the fine-tuning step on the same corpus separately, using 100 different random seeds (0 to 99).

Then, we have all 100 equivalent models predict the class of the 10 texts which constitute our target sample (i.e., the same text that were annotated by the human readers). Finally, using the LRP method, we generate an explanation of each prediction made by each model for each text, resulting in a set of 1,000 attention-based explanations.

4.3.4 Comparable Explanation Formats

Our goal is to produce explanation formats that are as comparable as possible between human annotations and model-based attention scores, while being explicit about the limitations of this approach. We therefore represent both human and model-generated explanations using attention maps, formatted as vectors associating each token with an importance value, and allowing for visualizations which are easily interpretable and enhance the plausibility of explanations to human readers (Reif, Yuan, Wattenberg, Viegas, et al. 2019). Although attention maps primarily offer localized, token-level insights and may not capture higher-level features such as long-range dependencies or syntactic relations, their clarity makes them well-suited for comparing explanations across different sources (e.g., human vs. LRP explanations).

For human explanations, attention maps are computed by mapping the highlights to numerical values for all annotators who selected the same class (*opinion* or *news*) for the same text. Each token in the text is assigned a score depending on whether it was not highlighted (0), slightly important (0.33), important (0.66), or very important (1). We acknowledge that these values are chosen arbitrarily and may not perfectly reflect the relative importance perceived by annotators (e.g., an "important" highlight may not actually be twice as relevant as a "slightly important" one in the eyes of the annotator), but they provide a consistent and reproducible way to map human explanations to numerical vectors. To obtain a global idea of what words were important for the human annotators in general, we can then visualize the average explanation across all annotators. For example, the human attention map presented

in Figure 4.2 (News 2) represents a vector combining the annotations of 34 participants for a given text.

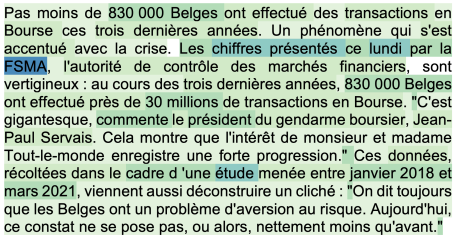


Figure 4.2: Example of a human attention map obtained thanks to the combination of all annotations.

Both human and LRP explanations can be visualized using color-coded attention maps. For *opinion* predictions, tokens are highlighted in shades of red (most important), orange, and yellow (least important); for *news* predictions, shades of blue, teal, and green are used. This formatting is inspired by (Sen et al. 2020), and adapted by (Escouflaire, Descampe, and Fairon 2024) to aggregate multiple explanations of a given text, making average explanations visually salient.

4.3.5 Explanation stability and Measures of Similarity

Finally, to measure the similarity between the different explanations, we use the **M**ean **C**orrelation **W**ith **M**ean **E**xplanation (MCWME) metric introduced in Chapter 3.

As long as we can sort the words in an explanation by their importance (which is the case for explanations displaying continuous values for each word), we can compute a plot of the MCWME with respect to k (i.e., the amount of top words who's relevance is not assigned to 0) and evaluate if the explanations differ or not on the most relevant words. Figure 4.3 shows such a plot.

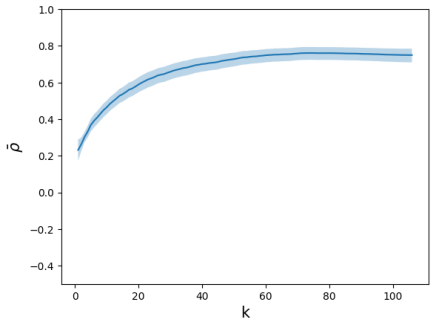


Figure 4.3: An example of a plot showing the evolution of the MCWME w.r.t the amount of selected top words.

For human explanations, which are discrete by design (only 4 different categories or colors can be used to identify relevant words), we can not sort the words among each of the colors. It means that it is not possible to get a unique explanation containing the k most relevant values for all possible values of k. The only easy way to obtain such explanations is the trivial *top-N* one, were we keep all the relevance values given by all the annotators. In order to have a similarity value for shorter explanation as well, we also compute the *top-1* MCWME by applying the procedure illustrated in Figure 4.4: We first create all variations of an explanation, that we denote as a block, by considering that each of the words having the highest relevance values is the *top-1* word. For example, for an explanation where x words are highlighted in red a human annotators we generate x variations, each of them having one of these x words assigned with the highest relevance value and all the others being assigned a relevance of 0.[1] We then obtain the MCWME by first computing the Pearson correlation between each explanation and the average of the explanations in the other blocks. We finally compute the average correlation by block before averaging the values obtained for all the blocks. It follows that for human explanations, we only compute the *top-1* and *top-N* MCWME values (the first and last points of the plots in Figure 4.3).

[1]This process scales very poorly when considering more than 1 top word, and is the reason why we did not consider more than *top-1* explanations as it would quickly become too computationally intensive.

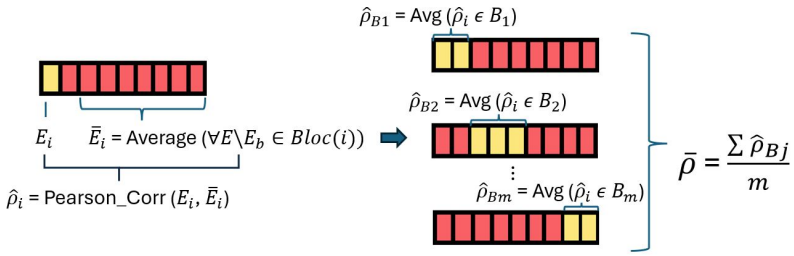


Figure 4.4: MCWME computation for *top-1* explanations extracted from discrete explanations containing respectively 2, 3, ... and N most relevant words.

4.3.6 Explanation discretization

Model-based explanations are inherently continuous: each token is associated with a value between 0 and 1. While it allows to evaluate the influence of the k top words on explanation stability, it makes the comparison between LRP and human explanations somewhat impractical. To allow direct comparison with human explanations, we can discretize these continuous values into four categories to match those used for human annotations. Note that the ranges used to discretize the LRP explanations, while ensuring a consistency across formats, are again completely arbitrary and may be replaced by any alternative thresholds. We considered two such sets of thresholds.

The first one, referred to as **linear** discretization, is a natural way to map continuous values between 0 and 1 to 4 values using equals spans: values between [0 - 0.25] are assigned to white (*not highlighted*), [0.26 - 0.50] to yellow (*slightly important*), [0.51 - 0.75] to orange (*important*) and [0.75 - 1.00] to red (*very important*).

The second one, referred to as **human-aligned**, aims at mimicking the same color distribution as the one used by the human annotators. Since this distribution varies from one text to another, the sizes of our 4 spans vary as well. By applying this process, we ensure that, for each color, the same amount of words are highlighted in the LRP and human explanations. To obtain the different color breakpoints, we first compute the amount of words highlighted in each of the colors by the human annotators for a given text, denoted as c_{red}, c_{orange}, c_{yellow} and c_{white}. We then sort all relevance values from the LRP explanations in descending order and take as breakpoints the values at the c_{red}, $c_{red} + c_{orange}$ and $c_{red} + c_{orange} + c_{yellow}$ positions.

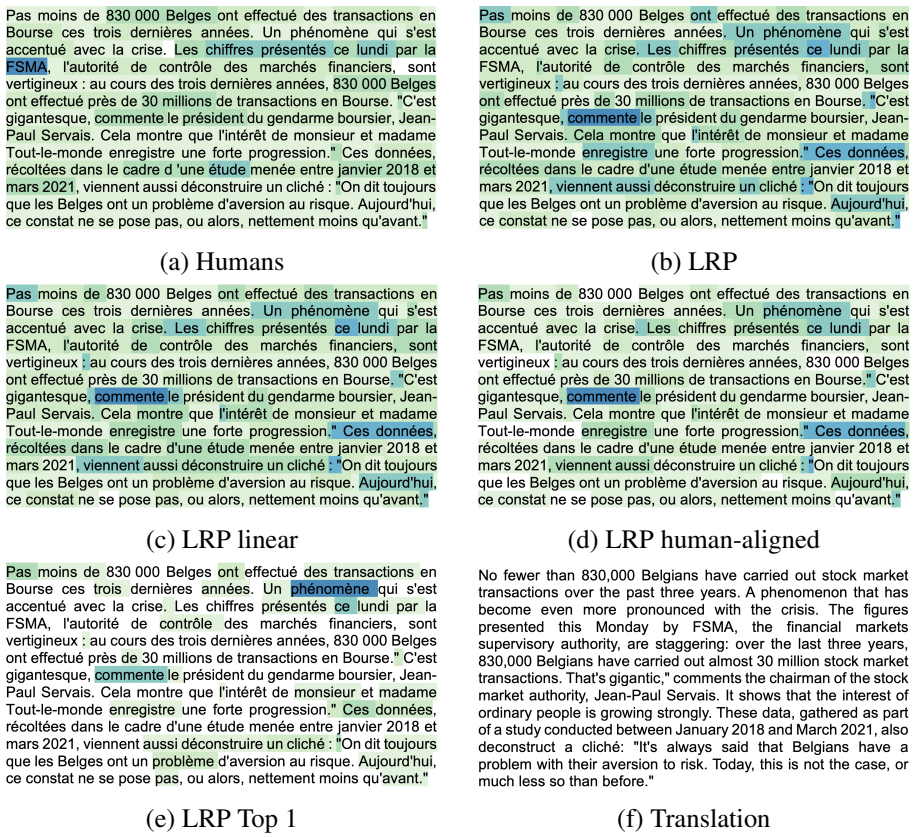


Figure 4.5: Average attention maps of Human and LRP explanations for *News2*

4.4 Classification Results

Although our focus is on the explanations' sensitivity to randomness rather than on the accuracy and variability of predictions, we first examine Table 4.1 to get an overview of the models' and humans' classifications for our target sample. The 10 texts in Table 1 are separated based on their "true label". Since the texts are excepts of press articles, the label is the one attributed to the original article by the media that published it, *News* or *Opin(ion)*. We removed from the experiment the annotations that did not contain any highlighted token and their associated prediction. It leads to some texts having less than 42 predictions, even though there were 42 total annotators in the experiment.

	Human labels (*Opin*, *News*)	LLM labels (*Opin*, *News*)
News 1	1, 39	0, 100
News 2	6, 34	0, 100
News 3	1, 38	0, 100
News 4	2, 34	0, 100
News 5	0, 41	0, 100
Opin 1	42, 0	100, 0
Opin 2	42, 0	43, 57
Opin 3	40, 1	96, 4
Opin 4	40, 0	100, 0
Opin 5	40, 0	100, 0

Table 4.1: Predictions made by humans and CamemBERT models on the 10 texts.

Table 4.1 shows that human annotators appear more unanimous when predicting the class of texts labeled as *opinion*: except for *Opin 3*, all *opinion* texts were classified as such by all annotators. On the other hand, only one *news* text (*News 5*) was classified as *news* by all annotators. This may suggest that the discursive style of the *opinion* class in journalism appears more obvious to human readers than that of the *news* class. However, disagreements are not radical: *News 2* shows the highest imbalance in terms of human predictions, with 6 annotators identifying it as an *opinion* text and 34 as *news*.

Regarding the predictions made by the 100 equivalent fine-tuned CamemBERT models, it appears that they are almost always unanimous: for 8 out of 10 texts, all 100 models agree on their prediction. This was expected, as all models were trained on the same data, with the different seeds only impacting the order in which the 8,000 texts appeared in the training set. Interestingly, and as opposed to humans, the models never disagree on excerpts of *news* texts. Opin 2 shows a very high discrepancy between models, with 43 models predicting the *opinion* class and 57 going for *news*. Because this paper focuses on the variability of human and LRP explanations of similar predictions, we will leave out this text from qualitative analysis, as predictions made by the model are particularly divergent. For quantitative analysis, we will only consider explanations of predictions corresponding to the ground-truth class (43 explanations for Opin 2 and 97 for Opin 3).

4.5 Qualitative Analysis

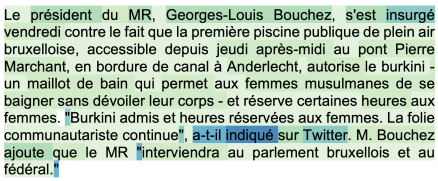
Before turning to quantitative measures of explanation stability, we first examine qualitative differences between human and LRP explanations. The goal here is to highlight structural contrasts in how relevance is allocated across texts, both in terms of distribution at token level and across spans of multiple words.

4.5.1 Linguistic Analysis of Texts

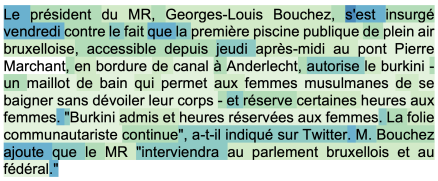
We start our analysis by a visual inspection of aggregated attention maps. The goal is to qualitatively assess how humans and LRP highlight different parts of the text when producing explanations, and to evaluate whether discretization or formatting choices bring LRP explanations closer to human-like patterns.

For this purpose, we compare attention maps that aggregate human explanations (top-N), continuous LRP explanations, their linear and human-aligned versions, as well as formats restricted to *top-1* tokens or to the most salient tokens in discretized vectors. Figures 4.5, 4.6 and 4.7 display examples for three representative texts: two classified as *news* (News 2 and 5) and one classified as *opinion* (*Opin 1*). These texts were selected because they illustrate the clear contrast between explanation formats. We deliberately avoided texts where the models disagreed or produced non-unanimous predictions, ensuring that the visualized comparison between model types is based on the same number of explanations for the same predicted label.

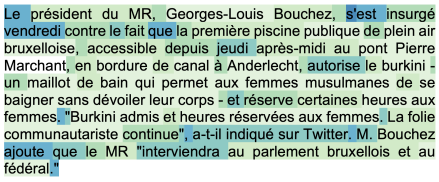
First, the attention maps illustrate how human and LRP explanations diverge in terms of which tokens receive the most relevance on average. For the text in Figure 4.5, humans (Figure 4.5a) appear to explain their predictions for *news* primarily with mentions of information sources (*FSMA*, *chiffres* [figures], *étude* [study]), as well as time markers (*lundi* [Monday], *janvier* [January]) and numerical digits. While some of these patterns align with those observed in Figure 4.5b (e.g., *ce lundi* [this Monday], *ces données* [these data]), the models assign a high relevance to quotation marks and to the quotation verb *commente* [comments]. Those are all recognized markers of the *news* class (Escouflaire, Descampe, Venant, et al. 2024), but are not the ones prioritized in human explanations.



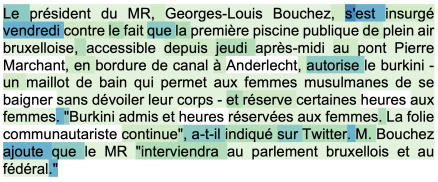
(a) Humans



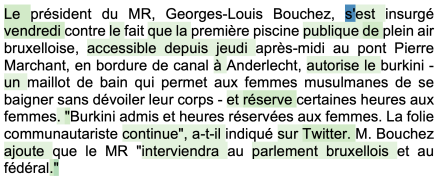
(b) LRP



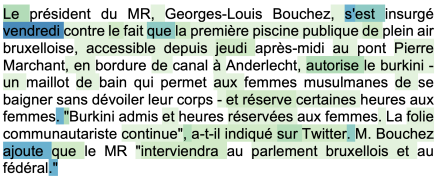
(c) LRP linear



(d) LRP human-aligned



(e) LRP top 1



(f) LRP top Color

The president of the MR party, Georges-Louis Bouchez, protested on Friday against the fact that Brussels' first public open-air swimming pool, open since Thursday afternoon on the canal-side Pierre Marchant bridge in Anderlecht, allows burkini - a swimsuit that allows Muslim women to bathe without revealing their bodies - and reserves certain hours for women. "Burkini allowed and hours reserved for women. The communitarian madness continues," he posted on Twitter. Mr. Bouchez added that the MR "will intervene in the Brussels parliament and at federal level."

(g) Translation

Figure 4.6: Attention maps of different explanation methods for *News5*

News 5 (Figures 4.6), like *News 2*, was classified as news by both humans and models. It further illustrates a recurrent difference between human and LRP aggregated explanations: human often concentrates on a small number of tokens or on a specific section of the text, whereas LRP explanations tends to be more diffuse. Here, both maps 4.6a and 4.6b highlight quotation marks and reporting verbs or clauses (e.g., "a-t-il indiqué sur Twitter" [he said on Twitter], *s'est insurgé* [protested]), but the models also distribute attention across additional tokens such as *vendredi* [Friday], *que* [that], *jeudi* [Thursday], and *autorise* [authorizes].

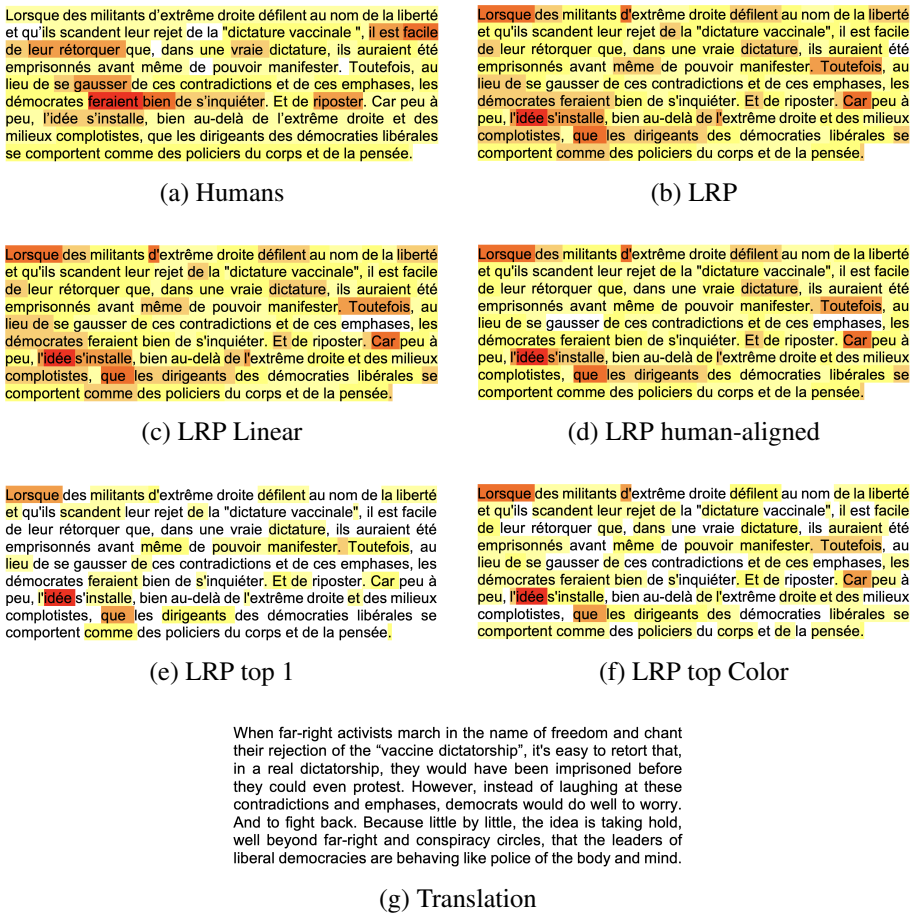


Figure 4.7: Attention maps of different explanation methods for *Opin 1*

For *Opin 1*, where both humans and models classified the text as *opinion*, the explanations are again substantially different. Humans (Figure 4.7a) give a high importance to axiological verbs such as *gausser* [mock], *feraient* bien [would do well], *s'inquiéter* [worry], and *riposter* [retaliate]. In contrast, the most highlighted tokens 4.7b are primarily argumentative connectives (*lorsque* [when], *toutefois* [however], *car* [because]) as well as the sentence *l'idée s'installe [...] que...* [the idea is taking hold [...] that...]. This text also illustrates the importance that transformer models allocate to sentence bound-

aries,i.e., the first and final tokens of sentences. This phenomenon, which was already documented by (Clark et al. 2019), is not reflected in human explanations.

Comparing attention maps of the original LRP explanations with their discretized versions also yields interesting observations. For all three texts, we can observe that there is almost no difference in terms of visualization between the original map (for example Figure 4.5b) and the maps of the linear and human-aligned versions of the same aggregation of explanations displayed in Figures 4.5c and 4.5d. The difference is slightly more visible in Figures 4.6d and 4.7d.

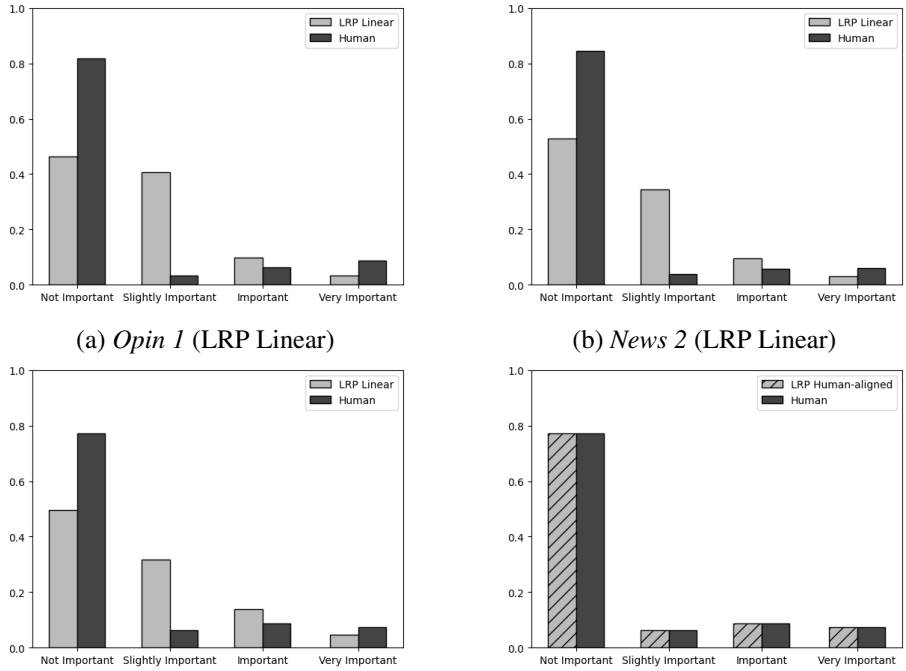
We also considered looking only at the *top-1* tokens of all original LRP maps, as well as at all tokens with the highest value in the discretized human-aligned maps (referred to as "top-color" maps). These formatting choices produce much more noticeable changes in the visual appearance of the maps. *Top-1* maps, such as Figures 4.6e and 4.7e, show a very sharp focus on a few tokens that already stood out in the original maps (Figures 4.6b and 4.7b). In some cases, however, they also bring to the foreground tokens that were less salient in the original maps, as illustrated in Figure 4.5e with the token *phénomène*, which appears consistently across models despite being less emphasized in the non-discretized versions. By contrast, top-color maps (e.g., Figures 4.6f and 4.7f) are sparser and visually closer to human attention maps. They highlight the tokens that models considered most important, but in a way that preserves more continuity and readability than *top-1* maps. In this sense, such human-aligned top-color maps may represent a useful compromise: they remain faithful to LRP explanations in terms of most relevant tokens, while producing patterns that may be more plausible and easier to interpret for human readers at the same time.

4.5.2 Distribution of Relevance Levels in Explanations

We now investigate how relevance levels are distributed in human and LRP explanations. This approach allows us to directly compare how annotators and models distribute relevance over an entire text. Figure 4.8 shows the proportion of words that get assigned to each relevance level by annotators and linearly discretized LRP explanations.

We can see in Figures 4.8a, 4.8b and 4.8c, that the amount of *not important* tokens is always higher for human annotators than for LRP explanations,

regardless of the text. It follows that, all relevance levels considered, annotators tend to assign relevance to fewer tokens than models do. Additionally, the distributions of the three levels of relevance are inverted between humans and CamemBERT models, further confirming the disparities later observed in Section 4.5.3. These results suggest that human annotators and transformer models importantly differ in terms of how they explain, at token level, the same predictions. Figure 4.8d additionally shows how this distribution of relevance levels is used to format the human-aligned versions of the CamemBERT explanations.



(a) *Opin 1* (LRP Linear)

(b) *News 2* (LRP Linear)

(c) *News 5* (LRP Linear)

(d) *News 5* (LRP human-aligned)

Figure 4.8: Proportion of tokens (words) associated to each relevance level.

4.5.3 Distribution of Explanation Segments

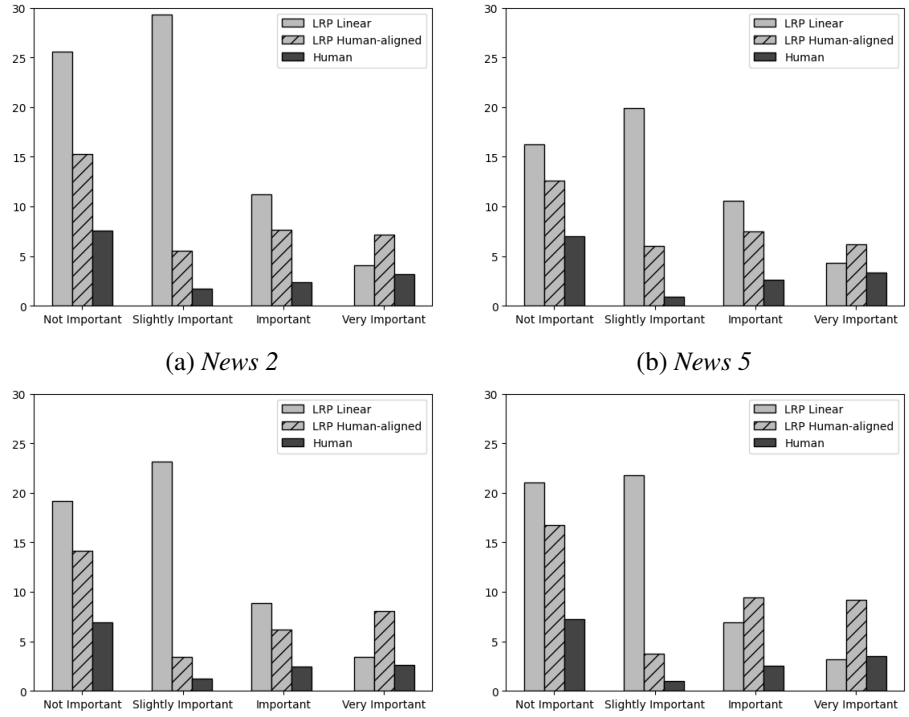
We now analyze how explanations are distributed across contiguous spans of text, so-called explanation segments. For example, in Figure 4.1, *Il est impor-*

tant is the first explanation segment highlighted in yellow (*slightly important*) by the annotator. The goal of this analysis is to compare how human and LRP explanations behave in terms of segment distribution. Figure 4.9 displays the average number of segments (spans of at least one token) for each relevance levels.

The main observation is that humans tend to highlight significantly fewer segments of text compared to the LRP explanations. This trend arises because the relevance values in the LRP explanations are a lot more sparse and volatile from one token to another than the highlights made by the annotators, who tend to consider longer segments of text as either somewhat important or not important at all. For example, Figure 4.6, shows that humans seem to mainly consider the beginning and the end of the text as relevant, while LRP gives attention in a more dispersed way. Another major trend observed in Figure 4.9 is that, proportionally, the LRP explanations attribute low levels of relevance (*not important* or *slightly important*) to a lot of segments. On the contrary, the *very important* level of relevance contains the lowest amount of segments. Annotators tend to leave a lot of segments blank (i.e., they do not highlight them), but are more likely to highlight segments as *very important* than *important*, and *slightly important* is the relevance level that is the least used. One possible explanation for this phenomenon is the difficulty of discerning the relative importance of three relevance levels for human readers. Readers may favor one or two levels (e.g., *not important* and *very important*) when the explanation of a choice through our highlighting task appears more complex. It is also possible that allowing individuals to highlight explanation spans instead of making them choose a different color for each token in the text accentuates this trend. Finally, when using human-aligned discretization, we obtain a distribution significantly closer to the human annotations.

4.6 Quantitative Analysis

While the previous sections focused on qualitative contrasts between human and LRP explanations, we now turn to quantitative measures of their variability. The goal here is to assess, in a systematic way, whether machine-generated explanations are more or less stable than human ones, and to examine how methodological choices such as discretization affect this stability.



(a) *News 2* (b) *News 5*

(c) *Opinion 1* (d) *Opinion 4*

Figure 4.9: Average number of highlighted segments for two *Opinions* and two *News*

4.6.1 Variability of Human vs. LRP Explanations

In this section, we examine how consistent human explanations are compared to those produced by transformer models. The goal is to assess whether humans and LRP explanations tend to highlight the same textual cues when justifying the same predictions, and to what extent the level of agreement among these two types of explanations varies depending on the class of the text. To do so, we use the MCWME metric designed in Chapter 3 and its discrete variant, introduced in Section 4.3.5, to quantify the stability across a set of explanations. We analyze how this stability evolves when considering low (i.e., top-1) or high amount (i.e., top-n) of most relevant tokens.

For both humans and models, we compute the MCWME based on explanations of the true label only, as presented in Table 4.1. This analysis allows us to characterize not only the general variability in explanations, but also their relative stability across the two classes, and to identify cases where humans and models diverge most strongly. We compare the MCWME scores and confidence intervals between all formats of human and LRP explanations described in Sections 4.3.4 and 4.3.6. The explanations and formats being compared are: human explanations, original (non-discretized), linear and human-aligned LRP explanations. We first show a comparison of human and non-discretized LRP explanations in Figure 4.10.

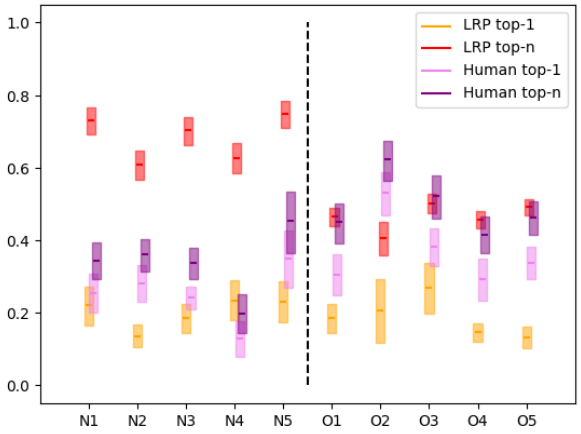
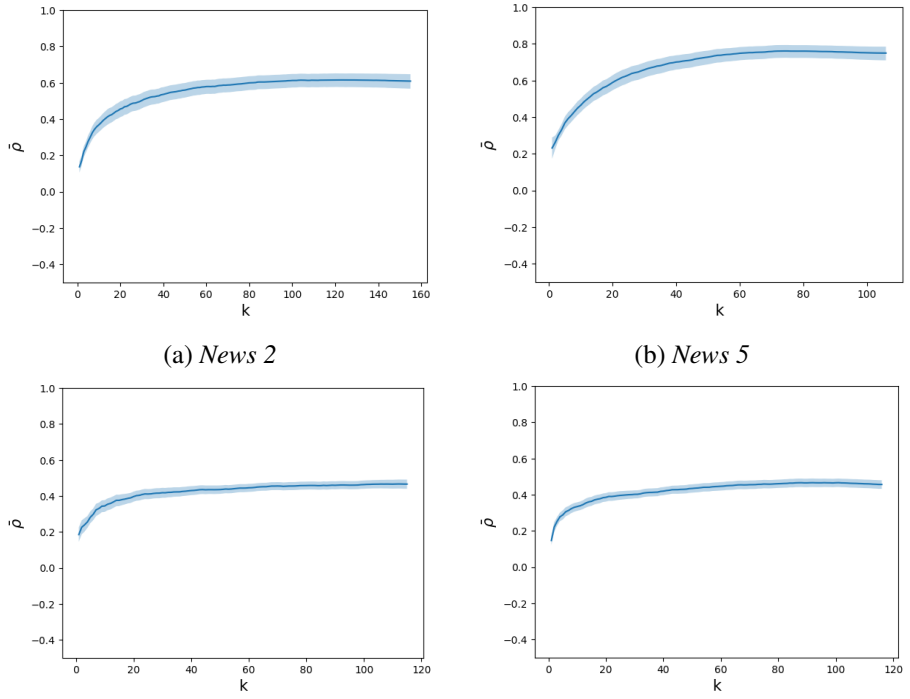


Figure 4.10: Comparison of the MCWME with confidence intervals for top-1 and top-n LRP and Human explanations of the 10 texts.

The first observation arising from Figure 4.10 is that taking only the *top-1* token with the most attention in each explanation systematically produces a lower MCWME score than when all tokens are considered in the calculation. However, this drop is much more severe for CamemBERT than for humans, which suggests that humans tend to agree more on the most important tokens in their explanations than CamemBERT models do. Figure 4.11 displays how the stability of CamemBERT explanations improves in terms of MCWME when increasing the amount of most relevant words considered among each explanation (k). We observe that tendency, across all texts in our experiment.



(a) *News 2*

(b) *News 5*

(c) *Opinion 1*

(d) *Opinion 4*

Figure 4.11: MCWME with respect to the amount of selected top words in each of the explanations for four different texts.

Regarding the class being explained, the boxplots in Figure 4.10 show another interesting difference. When using top-n explanations, LRP leads to higher MCWME score for the *news* class than for the *opinion* one. Figure 4.11 further confirms the difference between the two classes. First, while all the curves display a plateau, it is reached much earlier (around $k = 20$) for *opinions* than for *news* (around $k = 60$). On top of it, the MCWME value at the plateau is higher on average for *news* (between 0.6 and 0.8) than for *opinions* (between 0.4 and 0.5). On the contrary, human annotators seem to agree more on explanations of *opinion* than *news* texts. This may be because *opinion* articles typically contain clear markers of subjectivity and strong stylistic choices, which provide clearer and more consistent cues for interpretation. By contrast, *news* articles tend to follow a more neutral style, less marked by lin-

guistic markers, which may make it harder for human annotators to identify consistent cues for their classification (Koren 2004). Interestingly, the two texts that are the most divisive regarding predictions made by CamemBERT models in Table 4.1 (*Opin* 2 and 3) are also the two texts for which annotators show the highest consistency in their explanations, as shown by the results in Figure 4.10. *Opin* 2 was classified by 43 models as *opinion* and by 57 as *news*, while all humans identified it as *opinion*, and their explanations reach a MCWME score of 0.62. This further confirms that humans and LLMs do not focus on the same elements when predicting the class of a text.

4.6.2 Impact of the Discretization on Explanations Stability

We now examine how the variability evaluated in the previous section changes when continuous LRP explanations are discretized into categorical values. We compare the explanations in their continuous form (i.e., LRP) with their linear versions (i.e., LRP Linear), and further with discretization informed by human annotation distributions (i.e., LRP human-aligned). This step-by-step comparison makes it possible to assess whether discretization reduces the sensitivity to randomness and its dependency to the number of most highlighted tokens (k). It also helps to evaluate whether human-aligned thresholds bring LRP explanations variability closer to the one observed across the annotators. Figure 4.12 displays a comparison between the LRP and LRP Linear explanations.

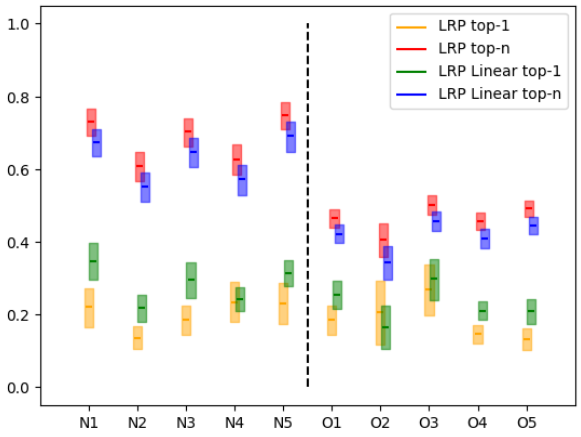


Figure 4.12: Comparison of the MCWME with confidence intervals for top-1 and top-n LRP and LRP Linear explanations of the 10 texts.

It reveals a consistent pattern: for $k = 1$, the MCWME values for the linearly discretized explanations are higher than the one for their continuous counterparts, while for $k = N$, the MCWME values for the discretized explanations are lower. This dual shift effectively reduces the gap between the stability scores for $k = 1$ and $k = N$, bringing them closer to the values observed in human explanations (cf. Figure 4.10 and 4.14). Figure 4.13 shows that the introduction of human-based discretization accentuates this effect: discretizing the attention values based on human-aligned thresholds yields even smaller gaps with the MCWME values observed for human explanations.

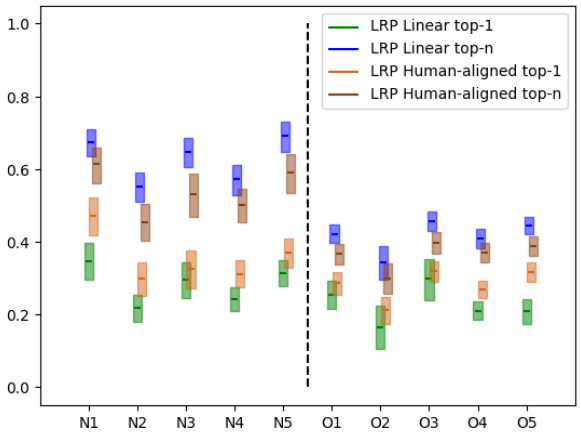


Figure 4.13: Comparison of the MCWME with confidence intervals for top-1 and top-n LRP Linear and LRP human-aligned explanations of the 10 texts.

Figure 4.14 shows the final comparison between human and human-based discretized explanations. We observe that for the *Opinions* 1, 4 and 5 and *News* 2 and 5 there is not even a significant difference between human and human-aligned LRP explanations in terms of stability. For other *News*, human-aligned explanations remain significantly more stable than human ones. For *Opinion* 2 and 3, the results are more difficult to interpret, but these texts are the one for which all models do not agree, which may explain why they follow a different pattern.

Based on our observations, we conclude that, in our case study, discretizing the LRP explanations, in particular when informed by human annotation, moderates the sensitivity to randomness when focusing on a small number of most highlighted tokens. For high values of k, it leads to the opposite behavior

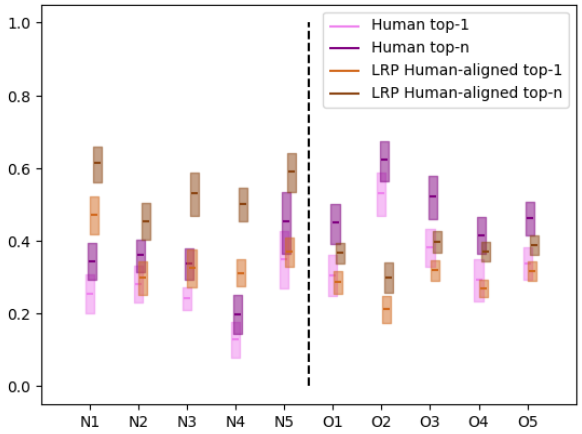


Figure 4.14: Comparison of the MCWME with confidence intervals for top-1 and top-n Human and LRP human-aligned explanations of the 10 texts.

and tends to increase the explanation sensitivity to randomness. In both cases, it leads to stability values that are more similar to the ones observed across the human annotations.

4.7 Chapter's Conclusions

In this chapter, we compared the variability observed across explanations provided by equivalent models with that observed across human annotations. We first started by showing that, despite predicting the same label most of the time, the models and annotators provided various explanations. We then highlighted that the human annotations displayed less variability than model explanations when considering fewer words, but that the opposite trend was observed when considering all the words in the texts.

We followed by showing in section 4.6.1 that humans tend to provide more concentrated explanations, highlighting few tokens and often focusing on a small number of linguistically salient cues of *news* and *opinion* classes. In contrast, LRP explanations are more diffuse and dispersed, with relevance assigned to a wide range of tokens. Interestingly, the class of the text also plays a role in explanation stability. For humans, explanations of *opinion* predictions were slightly more consistent than those of *news*, though the difference

was small. For LRP, however, *news* explanations were consistently more stable than *opinion* explanations, especially when considering all tokens. This was also reflected in our qualitative analysis of attention maps, as we observed that humans and models did not focus on the same types of tokens for their predictions of *news* or *opinion*. These observations suggest that humans and transformer models may prioritize different types of information when performing the same task, reflecting expected differences between cognitive and algorithmic processing.

In light of both the qualitative and quantitative results, the answer to the question "are model explanations more variable than human ones?" is not straightforward. On one side, we showed that LRP explanations are stylistically very different from human ones: they are more diffuse, spread over many tokens, and less concentrated on a few salient cues. Then, quantitative trends highlighted that LRP sometimes displays higher sensitivity than humans, especially when considering only the most highlighted tokens. However, this difference diminishes, and even reverses, when taking into account larger portions of the explanation. In other words, while machine explanations are not systematically more variable than human explanations, their variability follows different trends depending on how explanations are represented and compared.

Finally, our analysis revealed that discretizing the LRP explanations brings them substantially closer to human ones in term of stability. Only applying uniform breakpoints to LRP explanations already lowers the stability difference with human explanations, and human-aligned discretization further narrows the gap. When visualizing average attention maps and through qualitative analysis, we showed that human-aligned maps resembled human attention maps more closely than LRP ones. This supports the idea that mild explanation post-processing can meaningfully improve the alignment between LRP outputs and human reasoning. It however comes at the cost of a reduced faithfulness.

These findings have implications for explainability research in NLP. Raw LRP explanations may be faithful indicators of internal model reasoning, but they are not necessarily well-aligned with human explanations. Incorporating human-aligned constraints into explanation visualization of transformer models could lead to improvements in that regard. However, answering the question whether it is desirable or not to have models that show as low explanation stability as human annotators is not straightforward and will depends on the definition of plausibility we opt for. On the one hand, observing explanations

that are sensitive to randomness can be argued to be more plausible as it better reflects what is observed across human annotators. On the other hand, if the sensitivity to randomness leads to harder to interpret explanation distributions for the end users, then it can be argued to have a negative impact on plausibility.

We emphasize that this study, limited to the CamemBERT model and LRP explanations, may not be representative of the entire field of explainability applied to transformer architectures, but we believe that it still yields interesting insights to better understand the potential weaknesses of these approaches in general. In the bigger picture, we hope that it will help understanding the impact of explanation sensitivity to randomness on model explainability.

With that said, we limited our study to a French binary journalistic text classification task (news vs. opinion), using on a small sample of 10 texts only. On top of it, across all our analyses, both humans and fine-tuned CamemBERT models produced explanations that varied from one text to another, asking for more data to be taken into consideration to be able to draw broader conclusions in the future. It is possible that other tasks would yield different patterns of agreement or variability. It is also possible that other explainability methods or other models will be less aligned with our findings. Furthermore, LRP explanations are restricted to token-level attributions and do not capture potential relations or dependencies between words, which may limit the alignment with human reasoning processes. In addition, human annotations in this experiment were themselves constrained by the in-house highlighting interface and by the predefined importance categories, which may have influenced the way annotators expressed their explanations. This raises the inevitable broader question of whether human and model explanations, given their different formats and constraints, are ever fully comparable.

While we studied the stability difference between LRP and human explanations, we did not evaluate the plausibility of single explanations, and in particular the possible change in plausibility between two explanations of equivalent models on the same text. One interesting experiment to run would therefore be to expose human readers to different types of explanations. One could then evaluate if it impacts the reader to predict one class or the other. Such an experiment would allow us to verify whether the presented post processing techniques indeed lead to explanations that are easier to understand. It could also help to evaluate if individuals would be impacted differently by different

explanations coming from equivalent models. Observing such results would again alert on the potential harmfulness of only providing one explanation map when performing experiments about explainability.

Finally, we left aside text explanations given by generative models, such as ChatGPT. While we believe that running a similar study on these models may be of interest, we still lack a methodology to do so. In particular, in our case, where we have interest in evaluating the impact of the training randomness on explanations, the inherent non determinism of text generators when providing explanations would be a big concern. We therefore leave this idea as a potential further work.

Chapter 5

Explanation Sensitivity to Randomness Dependencies

The content of this chapter has been published in (Bogaert and Standaert 2024b) and also contains results obtained by Romain Loncour during his master thesis[1] that the author had the chance to partially supervise.

5.1 Motivations

In the last chapter, we showed that the variability across the explanations provided by models trained using different random seeds was not following the same tendency than human inter-annotator disagreement. We end our work by exploring the link between various factors and explanation sensitivity to randomness. In particular, while the previous chapter showed that post-processing the explanations can lead to a significant change in term of explanation stability, we now aim at studying how variations in the training data or the model's training capacity can impact it. To run our analysis, we make use of the tools we previously designed during this thesis: the MCWME stability metric combined with a noise analysis through boxplots.

[1] Available under request

5.2 Change in the data

5.2.1 Methodology

We next study the impact of changes in the training (and explained) data on explanation sensitivity to randomness. The datasets used to train the models will be specified in each of the experiments. When needed, we make use of synthetic datasets to perform a trivial classification task. These datasets are needed in order to isolate the impact of the targeted changes, which can be very difficult to do using real-world data. It allows to avoid biases by studying datasets that are as identical as possible, aside from the subject of study.

As previously, we use the CamemBERT-base model (Martin et al. 2020) when dealing with French text classification. To classify English pieces of text, we use the RoBERTa-base model (Y. Liu et al. 2019), which shares the same architecture but was pre-trained on an English corpora. In all cases, we fine-tune 200 models using the same hyper-parameters (which are the default ones, aside from the learning rate of 2×10^{-5}, the batch size of 16 and the number of epoch of 1), and a different random seed each time. We then select a subset of $m = 100$ equivalent models, whose accuracy does not significantly differ on our test set and compatible texts from the test set of the two datasets. We finally generate word-level explanations using the Layer-wise Relevance Propagation method (Chefer et al. 2021) as done previously.

5.2.2 Syntactic dependencies: Word order

Since the pre-training of large language models is always done on coherent sentences, we wanted to evaluate if using a different text structure during the fine-tuning would increase the explanation sensitivity to randomness. In our first experiment we therefore investigate the impact that shuffling the sentences' words in a dataset has on the explanations' stability.

To run our experiment, we use two datasets: the first one is composed of a set of 10 000 different 10 words long sentences, split in two classes, where only one word differs. To avoid semantically incorrect sentences when switching words, we make use of first names. The first class always contains the first name "John", while it is replaced by the first name "James" in the second one. The second dataset is composed of the exact same sentences, but the words composing them are shuffled. It follows that these two datasets are composed of the exact same word distribution, but that the order of those words varies

from one to the other. In both cases, the task is trivially to classify between texts containing the first name "John" or "James". We then select two (i.e., one for each training dataset) subsets of $m = 100$ equivalents models. All models reach 100% test accuracy on their respective task. We finally compute the MCWME metric for 80 different texts.

To compare text specific MCWME values obtained in different setups, we need a way to combine the MCWME obtained on different sets of texts. To do so, we use the format displayed in Figure 5.1. Each x-coordinate corresponds to a MCWME value and its confidence interval on a given text. Since each ordered text in the first dataset has a corresponding shuffled text in the second one, we can pair them together and proceed to multiple comparisons on the same plot. To allow for multi class visualization, the results for each class are separated by a dashed line. We for now only focus on binary classification, but this visualization could be extended to more.

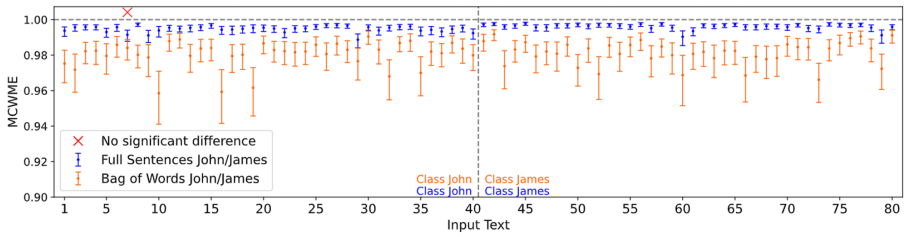


Figure 5.1: MCWME comparison plot between shuffled and non-shuffled sentences. The texts of the left class always contain the first name "John" vs. the first name "James" for the class on the right. Since the MCWME can range from 0 to 1, the span of the Y axis is adapted for space and formatting issues. A red cross on top of the plot highlights a non significant MCWME difference

Figure 5.1 shows the MCWME comparison between ordered and shuffled (i.e., Bag of words or Bow) sentences. We first see close to perfect explanation stability when the words are not shuffled. This is quite reassuring, as if that was not the case, that would mean that the explanations provided a significant relevance value to multiple words in the text. Since the task is built to only have one discriminant word the model can rely on to perform its task, and since models always reach perfect accuracy, we know that they have to rely on that word (and that word only) heavily. Would the explanations not be close to perfectly stable, the only conclusion we could have come up with

would be that the explanation method is not reflecting properly the model's behavior. While a close to perfect explanation stability does not prove that the explanation method accurately reflects the models behavior (and no empirical analysis will ever do), this first observation does at least not invalidate our trust in LRP's capacities. Figure 5.2 and 5.3 further confirm that all explanations indeed only provide a high relevance value to the discriminant word for the task.

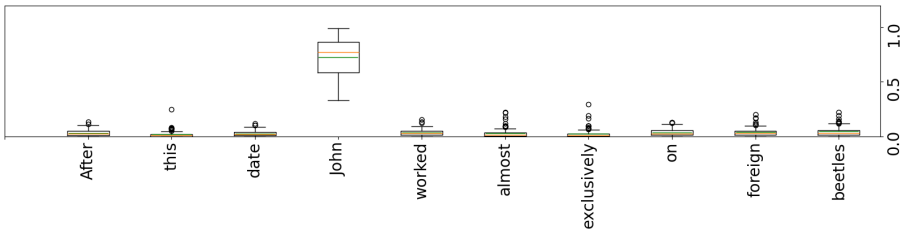


Figure 5.2: Explanation boxplot for a text containing the first name "John". The green line represents the median word value across all explanations, the orange one represents the average one.

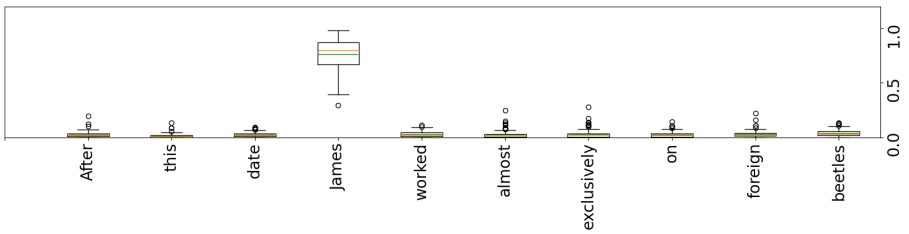


Figure 5.3: Explanation boxplot for the same text but with the first name "James".

Regarding our research question, we observe that all MCWME comparisons between one text and its respective counterpart[2] show significant differences. We conclude that shuffling the words during the fine-tuning leads to a significantly higher explanation sensitivity to randomness. While the explanation stability in both setups remains close to perfect, this result is rather surprising as the word order was not needed to perform the task in either of the

[2] Aside from one, highlighted by a red cross on top

setups. We hypothesize that this difference arises because transformer models can't get rid of the relations between words, implemented through the attention mechanism, when they perform their task. This is true even when it only provides detrimental information as in our experiment.

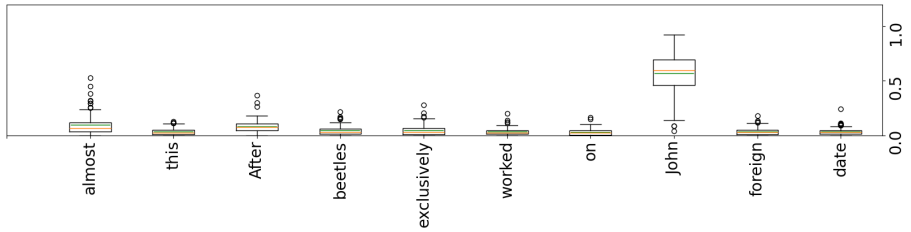


Figure 5.4: MCWME comparison plot between sentences ending by a period or not. The texts of the left class always contain the first name "John´´ vs. the first name "James´´ for the class on the right.

5.2.3 Syntactic Dependencies: Period or not

We next study the impact of putting a period at the end of all the sentences contained in the ordered dataset. Such an addition supposedly has no impact on the task difficulty, and all the sentences remain ordered. They are still composed of the same words, aside from one: the name "John" that is replaced by the name "James" in the other class. The only difference between our two datasets is the period at the end of each sentence in the second dataset. Figure 5.5 shows a MCWME comparison plot for all the texts in our dataset. It shows that adding a period to all the sentences leads to a lower explanation stability than when having no period in any of the sentences.

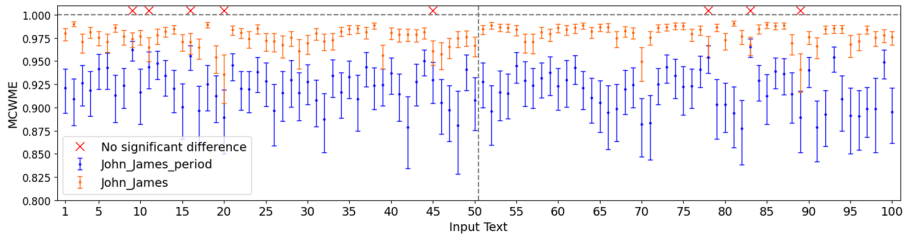


Figure 5.5: Explanation boxplot for the same text but with a peridod at the end.

When looking at an explanation boxplot, displayed in Figure 5.6, we see that the last word (i.e., the word to which the period was added) gathers more relevance than previously. While this is again quite surprising, we refer to (Clark et al. 2019) who hypothesizes that the relation of a word with the end-of-sentence token [SEP] is used by the model to perform a "no-op" operation. Adding a period at the end of the sentences could make this behavior appear, as this marker is semantically linked to sentence separations in human languages, and mechanically linked to the [SEP] token internally. It is however also possible that simply adding a punctuation marker to all sentences triggers specific attention heads that were not used previously. These internal processes, despite being useless for the task, end up impacting the LRP explanations.

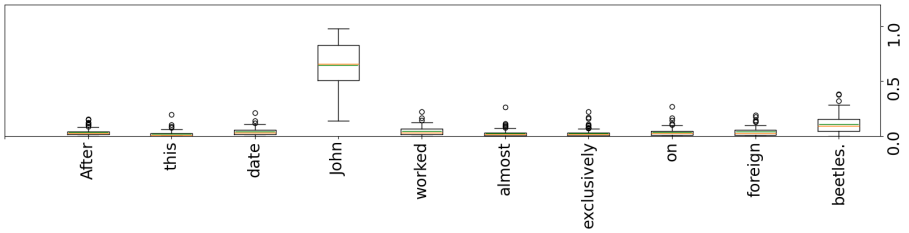


Figure 5.6: Explanation boxplot for the same text but with a period at the end.

5.2.4 Class dependency: absence of discriminant markers

One of the fundamental characteristic of LRP explanations is that they can only provide a relevance value to the words present in the text. In this next experiment, we investigate the explanations of classes that are characterized by the absence of discriminant words which, by design, cannot be highlighted in the text. To run our experiment, we use two datasets: the first one is the same ordered dataset as in the previous sections. The second one is built by keeping the class corresponding to the first name "John" as is but by replacing the first name "James" by a random word in the other class.[3] The task is still trivial, but the label is now determined solely by the presence or absence of the first name "John". Figure 5.7 shows a MCWME comparison plot between the two setups.

[3]We also investigated simply removing the first name in the other class, but the size of the text would become a discriminant factor as well, and the sentence's coherence would be impacted equally as it is when adding a random word.

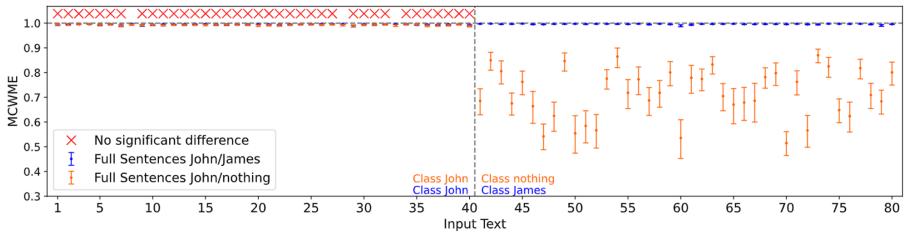


Figure 5.7: MCWME comparison plot. The texts of the left class always contain the first name "John" vs. no first name for the class on the right.

We observe that the MCWME is significantly lower for the class without any discriminant marker which again confirms that explanation sensitivity to randomness can be class specific, as already observed in the previous chapter. Interestingly, we observe MCWME values around 0.7, while random explanations would provide values close to 0. We hypothesize that this value arises because all words are not considered as equally irrelevant. Despite the average explanation being "flatter", the words with the highest average relevance seem to be at the beginning and end of the sentences and around the replaced word. Figure 5.8 shows such an explanation.

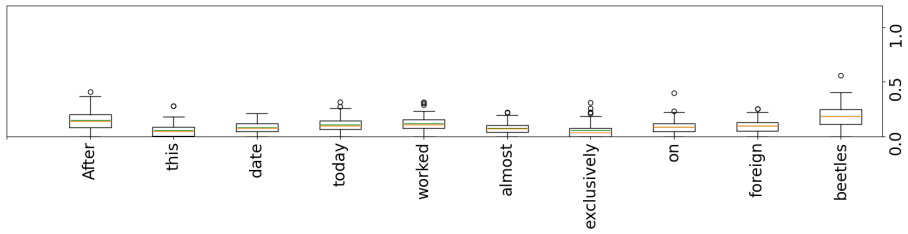


Figure 5.8: Explanation boxplot for a text where the word "John" was replaced by the word "today".

5.2.5 Task dependencies : ArXiv vs InfOpinions

We finally compare the explanation stability across two different tasks. The first one consists in categorizing paper abstracts in the astro-physic or mathematic category. To train our models, we use the ArXiv dataset [4] in which

[4] https://www.kaggle.com/datasets/Cornell-University/arxiv

we only keep 5000 abstract with the Astro-ph.GA and 5000 with the Math.NT tags. The texts are 148.43 tokens long on average. The second task consists in classifying press articles into the *news* and *opinion* categories, for which we use the InfOpinion dataset (Bogaert, Escouflaire, et al. 2023b) to train our models, as done previously. The texts are 338.6 tokens long on average. The models trained on the Arxiv dataset reach 99.8% accuracy on the test set on average, while models trained on InfOpinion reach around 96%. Since we have no way to pair the texts from the InfOpinion and ArXiv datasets together, we pair them randomly for the sake of simplicity. Figure 5.9 shows a MCWME comparison plot over the two tasks.

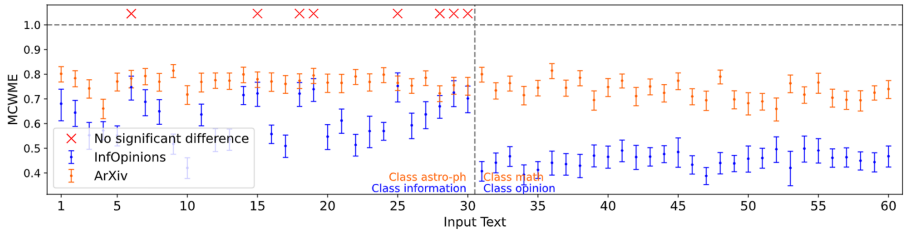


Figure 5.9: MCWME comparison plot for explanations based on the InfOpinion task and the ArXiv task.

We first confirm that the explanation stability is different between the two classes of the InfOpinion dataset, as observed previously on a smaller subset of texts. While it is possible that this trend comes, as for our synthetic use case, from one class having less or no distinctive words, we can not conclude on it. It is also possible that the length of the text plays a role, since *opinions* tend to be longer than *news* in general (and inside our corpus as well). We therefore leave the investigation of the reasons behind this difference as a potential further work.

We also observe a significant difference in the explanation stability on both tasks. At the very least, it means that the explanation sensitivity to randomness can be (and probably often is) task dependent. We hypothesize that this gap comes from the more discriminant vocabulary between the classes in the ArXiv dataset than in the one in the InfOpinion dataset, where a more in depth understanding of the relations between the different words is needed. This hypothesis is made increasingly plausible by the fact that the models perform better on the ArXiv dataset than on the InfOpinion one, reflecting that one task

is easier than the other for the models. Other factors can nevertheless play a role in this difference as well, such as the length or the language of the texts.

5.3 Change in the model's capacity

We now turn toward the impact of the model's training capacity on explanation sensitivity to randomness. More specifically, we are interested in its relation with the fine-tuning surface (FTS) that we define as the number of bits modified during the fine-tuning. The FTS is an interesting measure since it is directly related to the quantity of information that is learnable by the model during the fine-tuning phase.

5.3.1 Methodology

We run all our experiments on the InfOpinion dataset. Since the dataset only contains French texts, we use the CamemBERT French pre-trained transformer model (Martin et al. 2020), that we adapt to the task by either fully fine-tuning it or partially fine-tuning it (cf. Section 1.2). In the first case, we fully fine-tune the model on the training set during 2 epochs. In the second case, we first probe the internal representation of each text from the training set at the last layer of the pre-trained model without retraining it. As Peters et al. (2019), we then use the first token's representation as a text embedding and train a RoBERTa classification head during 20 epochs on top of these embeddings. We refer to these alternatives as the *fine-tuned* and *frozen* models. The frozen model naturally leads to a lower fine-tuning surface by reducing the amount of parameters affected by the fine-tuning.

These two types of models can then be quantized, which consist in reducing the size of some of the model's parameters or weights (Dettmers et al. 2023; Jacob et al. 2018). This process also reduces the fine-tuning surface, but this time by lowering the amount of bits used by some of the parameters, mainly in the linear layers of the model. We restrict our study to post-training quantization without quantization-aware training which means that we don't retrain our models after their quantization. The two most common ways to quantize a model are to use 8-bit or 4-bit floating point parameters (FP8 or FP4) and to use 4-bit NormalFloat (NF4). We use the FP4 and NF4 techniques as implemented in the bitsandbytes library.[5] It follows that we consider 6 types

[5] https://github.com/TimDettmers/bitsandbytes

of models for our study: the *fine-tuned* and the *frozen* ones, used either with no quantization or with a FP4 or a NF4 quantization.

To obtain our explanations, we train each of the *fine-tuned* and *frozen models* on 200 different random seeds. We then quantize them using either no technique, NF4 or FP4 quantization. For each of these 6 types, we keep the 100 models with the closest performances. We finally explain some texts of our test set via the LRP explanation method.

5.3.2 Accuracy and fine-tuning surface analysis

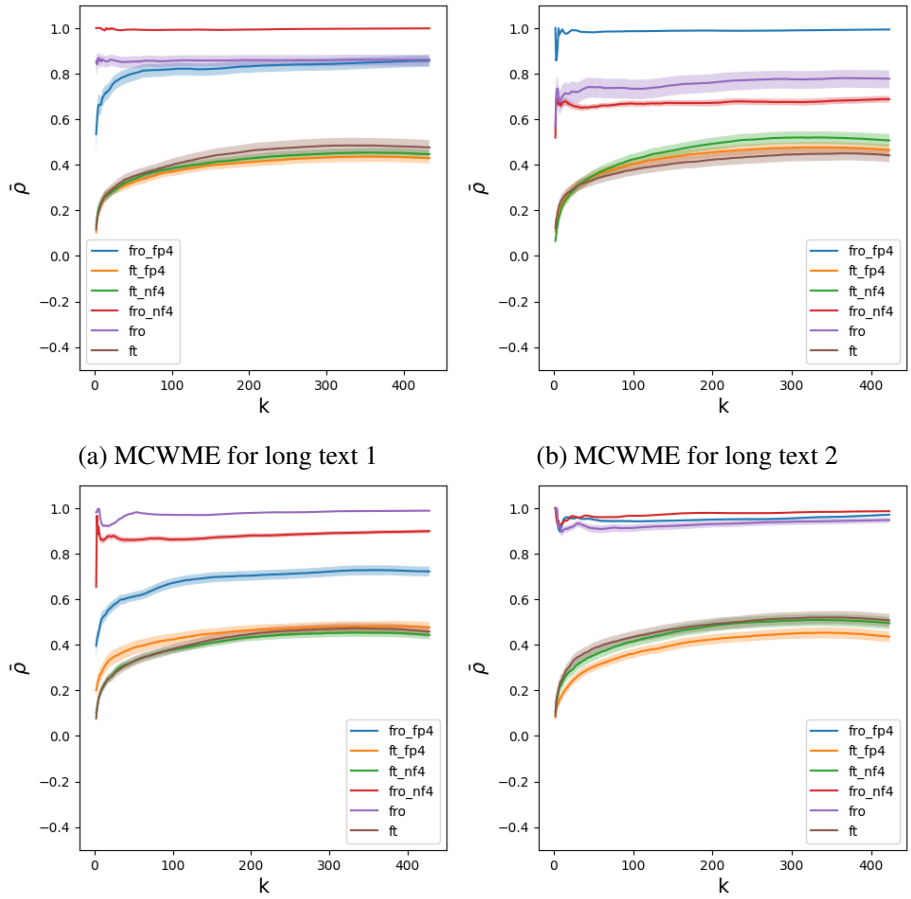
	Avg. accuracy	Fine-tuning surface
Ft	0.960 ± 0.06	3.5×10^9
Frozen	0.960 ± 0.03	1.8×10^7
Ft_q(FP4)	0.967 ± 0.05	7.4×10^7
Ft_q(NF4)	0.961 ± 0.06	7.4×10^7
Frozen_q (FP4)	0.952 ± 0.04	2.4×10^6
Frozen_q (NF4)	0.951 ± 0.05	2.4×10^6

Table 5.1: Accuracy and fine-tuning surface of the 100 equivalent models.

From Table 5.1, we can first see that reducing the amount of trainable bits of the models, whether it is by freezing some of the parameters or by quantizing some of the weights, does not lead to a significant drop in accuracy. This is rather surprising, as we divide the fine-tuning surface by approximately 1000 when going from the fully-fine-tuned model to the quantized frozen ones. It however remains substantial as the models still have 20 times more trainable bits than we have data in our entire training set.

5.3.3 Fine-tuning surface dependencies

We first show MCWME plots with respect to the number of top-words k for different texts. Figure 5.10 and 5.11 show the results for long and short texts respectively. We can first observe that for all texts, their is at best a minor difference between the curves corresponding to the explanations of the 3 *fine-tuned* models. These curves follow the same tendency as well: they start at their lowest value, then grow up quickly and end with a nearly completely flat plateau to reach their highest value at $k = n$. Based on these observations, we conclude that the quantization seems to have little to no impact on fully *fine-tuned* models in term of explanation sensitivity to randomness.



(a) MCWME for long text 1

(b) MCWME for long text 2

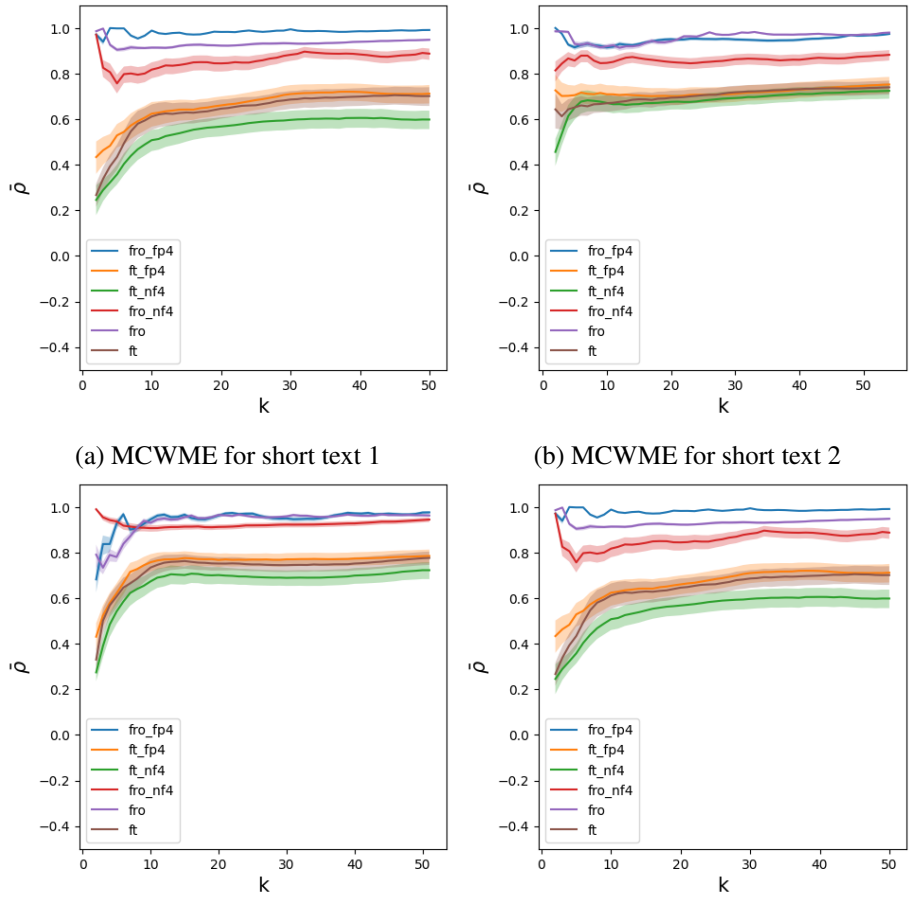
(c) MCWME for long text 3

(d) MCWME for long text 4

Figure 5.10: MCWME with respect to k for 4 different long texts

It also appears that the explanations of such models exhibit a significantly lower stability than the explanations corresponding to the *frozen* models. The MCWME value is around 0.5 for long texts and 0.7 for short ones. All curves corresponding to the *frozen* models display a higher explanation stability for all values of k. We conclude that lowering the amount of parameters impacted by the *fine-tuning* indeed lowers the explanation sensitivity to randomness. Finally, we observe varying tendencies across the frozen models. For short texts, the NF4 quantization always leads to a lower explanation stability compared

to the two other alternatives, which appear to be very similar to one another. This quantization technique also leads to very different trends, sometimes even inverted from all the others, peaking at low values of k, decreasing quickly and then stabilizing around its $k = n$ value.



(a) MCWME for short text 1

(b) MCWME for short text 2

(c) MCWME for short text 3

(d) MCWME for short text 4

Figure 5.11: MCWME with respect to k for 4 different short texts

For long texts, the 3 alternatives can display significantly different results, but none of them is better for all texts. Aside for the FP4 quantization technique, the curves are rather flat, showing no dependency with respect to k.

We follow up by illustrating the distribution of the relevance that LRP assigns to each word of a given text. Figure 5.12 provides boxplots for all the models on one text from our test set. The three bottom (resp., upper) rows display the sensitivity to the training randomness of the *fine-tuned* (resp., *frozen*) model and its quantized version.

We can first observe that the different models, despite their similar accuracy, do not assign relevance to the same words on average. For example, the explanations of *frozen* models attribute more relevance on words 2 and 4 (Policiers/Policeman and Chargé/Rushed) than the *fine-tuned* models, while the *fine-tuned* models attend more to the end of the text (Selon les journalistes/According to the journalists). We can also notice that the box-plots corresponding to the *frozen* model and its quantized versions are thinner than the ones of the *fine-tuned* models. Finally, the quantization has little impact on the top words considered by the models. However, while it lowers the fine-tuning surface, it tends to increase the box's sizes (i.e., the word level noise), in particular for the *frozen* model.

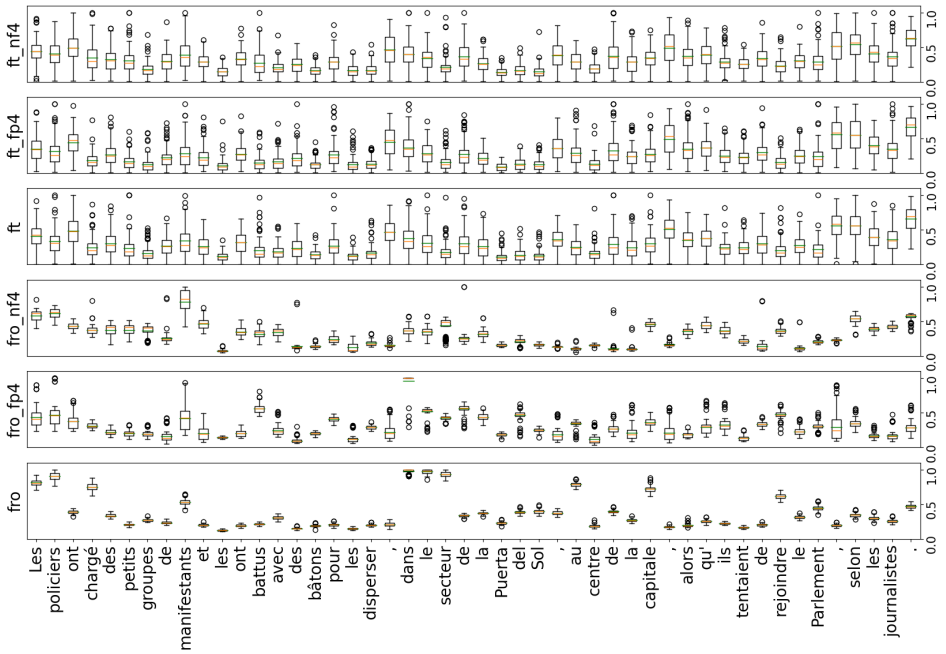


Figure 5.12: Boxplot for a short piece of News.

We conclude that these two interventions (i.e., partial freezing and weight quantization), while they both reduce the fine-tuning surface, impact the explanation sensitivity to randomness very differently. On the one end, partially freezing the model during the fine-tuning tends to drastically reduce explanation sensitivity to randomness. On the other end, the impact of post-training model quantization varies depending of the text instances, but does not have that big of an impact in general and can even lead to a higher explanation sensitivity to randomness.

5.4 Chapter's conclusion

In this last chapter, we leveraged the tools developed during this thesis to evaluate the impact that various modifications have on explanation sensitivity to randomness. After showing that we have no reason to doubt the capacities of the LRP explanation method as it provides both very stable and faithful explanations on a trivial task, we focused on two different types of modifications: on the data side, and on the model side.

Using synthetic data and trivial classification tasks, we first showed that shuffling the words in every sentence leads to lower explanation stability. We also observed that adding a dot at the end of every sentence leads to the same behavior. While both the word order and the period at the end, despite not being discriminant factors between the two classes may end up being used by the model to perform its task, it raises questions about what we intend to see in the explanations.

More precisely, it asks whether explanations should provide a high relevance value to discriminant words only, or to all the words used in contextualization operations performed inside the model but not discriminant by themselves. For example, nothing prevents the model from using the discriminant key word in combination with any other word to perform a trivial task. This combination would in its turn also become a discriminant feature, but with two elements instead of one. If multiple models proceed that way to classify a piece of text, faithful explanations would naturally be unstable (as all words in the sentence can be used in pair with the discriminant one, and different seeds are likely to use different words). On the other hand, if we expect these models to follow the Occam's razor principle, then they should only rely on the unique discriminant key word and provide a high relevance value to that word solely. Faithful explanation reflecting this behavior would be perfectly stable.

Based on this reasoning, two conclusions follow. First, if we consider that simple explanations (e.g., only highlighting the discriminant key word) are more plausible than more convoluted ones (e.g., highlighting the discriminant key word in combination with other words), then stable, plausible and faithful explanations can only be obtained if the model follows the Occam's razor principle to perform its task. Second, if we posit that the LRP explanations are perfectly faithful, then we showed that our model does not follow the Occam's razor principle (otherwise, the explanations would be perfectly stable regardless of the small modifications we applied), which, based on the first conclusion can be detrimental to its explainability. It therefore raises questions about the possibility of combining the excellent performances of transformers on difficult tasks while pushing them to use simple reasoning on easy ones.

Removing the discriminant markers naturally also leads to a lower explanation stability. While not surprising, it also raises interesting questions. Namely, in real world use-cases, what is the cause of the explanation stability difference between different classes or tasks ? We showed that data that do not follow the same structure as what was seen during the pre-training, punctuation or factors that cannot be reflected in the explanations such as the absence of markers[6] can all lead to significant differences. It follows that understanding the reasons why explanation sensitivity to randomness is varying from one setup to another is not trivial at all in real word use-cases. This problem is enforced even more by the text dependency of the explanation sensitivity to randomness.

Model-wise, we evaluated whether the fine-tuning surface is a good proxy for this sensitivity. Our results provide empirical evidence that it is not. On the one hand, simplifying the models by quantization does not forcibly reduce the explanations' sensitivity to the training randomness. This raises the question whether other model simplifications could lead to different results. On the other hand, freezing the models significantly reduces this sensitivity. However, such a freezing is hiding the sensitivity issue more than it is impacting it. It takes the frozen part of the model as a ground truth whereas it is also the outcome of a pre-training, the explanations of which are possibly affected by an equally high sensitivity to the training randomness (even harder to characterize due to computational reasons). It should therefore be viewed as a possible separation of duties, which only makes sense if the frozen part of the model

[6]The text length would be a good example of such a feature as well and therefore an interesting extension of this work.

used to generate text embeddings is itself explainable via other (e.g., linguistic) means.

Eventually, and from a more methodological viewpoint, our results are again based on simple (word-level, first-order, univariate) explanations (Bogaert and Standaert 2024a). Whether other types of (possibly more complex) explanations or explanation methods would lead to similar results could be an interesting research direction in the future.

Chapter 6

Conclusion

This thesis explored the still understudied topic of explanation sensitivity to randomness.

We started by defining the problem in Chapter 2. We showed that models fine-tuned on the same training set but with different randomness (in the form of a seed), although they yield similar accuracies, can provide different explanations for texts on which they give the same prediction. We then argued that, since there is no reason to prefer one model over another in this context, restricting ourselves to the explanation of a single model would be insufficient. We showed that this phenomenon could (positively or negatively) impact different features related to the model explainability, and that further study was therefore needed to characterize explanation sensitivity to randomness.

We then tackled the question of the metrics to use to quantify explanation sensitivity to randomness in Chapter 3. To do so, we first formalized the word-level univariate first-order or $(1, 1, 1)$ plausibility assumption, in order to define the type of explanations we had interest in studying as well as possible alternatives. We then investigated the use of information theoretic metric and leveraged their flaws to discuss what our stability metric needed to reflect. We concluded by showing that the Mean Correlation With the Mean Explanation (MCWME) metric helped alleviate these flaws, in particular when used in combination with word-level noise visualization in the form of boxplots.

Thanks to these tools, we ran our first characterization of explanation sensitivity to randomness in Chapter 4. More precisely, since different human annotators can provide different explanations as well, we wanted to evaluate

whether the explanation sensitivity to randomness was comparable with the inter-annotator (in)stability. To answer our question, we first ran a human annotation experiment based on a *news* vs. *opinion* classification task. We then compared the variability observed across explanations provided by Layerwise Relevance Propagation (LRP) explanations of equivalent CamemBERT models with that observed among human annotators. In all our analyses, both humans and fine-tuned CamemBERT models produced explanations that varied from one instance to another. We also showed that the patterns and sources of this variability differed interestingly between the two. Finally, post-processing the model explanations led to an explanation sensitivity to randomness closer to the variability observed across human annotators.

Seeing that post-processing the explanations had an impact on explanation sensitivity to randomness, we studied the impact of changes in the data and in the model's amount of fine-tuned bits in Chapter 5. We first showed that we have no reason to doubt the capacities of the LRP explanation method as it provided both very stable and coherent explanations on a trivial task. We then highlighted that the context's plays a role in explanation sensitivity to randomness, even when it is not needed for the task. We followed by observing that the absence of discriminant markers in one of the classes leads to a significant increase in term of explanation sensitivity to randomness. We finally characterized the explanation sensitivity to randomness of models trained on two different tasks, observing clear stability differences but being unable to conclude on the reasons. Finally, we showed that the model's number of fine-tuned bits has an impact on the explanation sensitivity to randomness, but we found no relations between the size of the model's parameters and the explanation stability.

While all these findings provide valuable insights, they are mainly based on a particular combination of the CamemBERT transformer model and the Layerwise Relevance Propagation explanation method applied to a French *opinion* vs *information* classification. It follows that the generality of our observations should be extended. While there is not reason to believe that what we showed in this work only appears in the setups we tried, we still lack a more extensive research to confirm this natural intuition. The examination of other corpora, especially in languages other than French, would be interesting. The study of other NLP tasks would be consolidate our results as well. For example, the detection of 'fake news', which is also recognized in the field as being particularly complex (Vargo et al. 2018; Zellers et al. 2019) could be inter-

esting to work on, and would be a nice come back toward this thesis initial field of research, but with another approach. Other modalities, such as image classification could be investigated as they are also concerned by explanation sensitivity to randomness (Watson et al. 2022).

At a more technical level, the evaluation of other language models and other explanatory methods with possibly different formats would also be necessary. In this work, we only investigated (1,1,1) explanations in the form of attention maps. This stands in contrast to explanations produced by generative models in the form of 'chain-of-thought' reasoning which aim at being plausible, but they are intrinsically non-deterministic for a given model instance and therefore cannot guarantee faithfulness to the underlying decision process (Turpin et al. 2023). We emphasize again that the field of Explainable Artificial Intelligence is for now mostly based on empirical analysis and that a more theoretical framework may be needed to be able to draw more robust conclusions. At the very least, we hope that this work will push for more formalization of explanation sensitivity to randomness in the future, regardless if it is to evaluate a model's capacities, to study a specific task or as a way to test the behavior of new explanation method.

As a track for possible further works, and on top of all the suggested extensions at the end of the different chapters, we note that while we showed that there was variability in the explanations, we did not evaluate whether this variability was meaningful to human readers. In particular, we showed that it did not correspond to what's observed between various annotators both in terms of MCWME value and word highlighting trends. We however did not investigate whether the different explanations composing this distribution had different explainability properties. Answering this question would require to have a large scale explanation review experiment. This experiment would also allow to check whether explanations clustered together based on our stability value indeed share similar traits according to human reviewers. It could finally provide at least a partial answer to the question "Is explanation sensitivity to randomness detrimental to plausibility ?".

Bibliography

Abnar, Samira and Willem H. Zuidema (2020). "Quantifying Attention Flow in Transformers". In: *ACL*, pp. 4190–4197.

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. (2023). "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774*.

Adebayo, Julius, Justin Gilmer, Ian Goodfellow, and Been Kim (2018). "Local explanation methods for deep neural networks lack sensitivity to parameter values". In: *arXiv preprint arXiv:1810.03307*.

Alhindi, Tariq, Smaranda Muresan, and Daniel Preoţiuc-Pietro (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.

Antoun, Wissam, Virginie Mouilleron, Benoît Sagot, and Djamé Seddah (2023). "Towards a Robust Detection of Language Model-Generated Text: Is ChatGPT that easy to detect?" In: *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles, TALN 2023 - Volume 1 : travaux de recherche originaux - articles longs, Paris, France, June 5-9, 2023*. Ed. by Christophe Servan and Anne Vilnat. ATALA, pp. 14–27. URL: https://aclanthology.org/2023.jeptalnrecital-long.2.

Arras, Leila, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek (2017). "Explaining Recurrent Neural Network Predictions in Sentiment Analysis". In: *WASSA@EMNLP*. ACL, pp. 159–168.

Artstein, Ron (2017). "Inter-Annotator Agreement". In: *Handbook of linguistic annotation*, pp. 297–313.

Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek (2015). "On pixel-wise expla-

nations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7, e0130140.

Balota, David, Maura Pilotti, and Michael Cortese (2001). "Subjective Frequency Estimates for 2,938 Monosyllabic Words". In: *Memory & cognition* 29, pp. 639–647.

Bansal, Naman, Chirag Agarwal, and Anh Nguyen (2020). "SAM: The Sensitivity of Attribution Methods to Hyperparameters". In: *CVPR Workshops*. Computer Vision Foundation / IEEE, pp. 11–21.

Basile, Valerio, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. (2021). "We need to consider disagreement in evaluation". In: *Proceedings of the 1st workshop on benchmarking: past, present and future*. Association for Computational Linguistics, pp. 15–21.

Benamara, Farah, Maite Taboada, and Yannick Mathieu (2017). "Evaluative language beyond bags of words: Linguistic insights and computational applications". In: *Computational Linguistics* 43.1, pp. 201–264.

Bethard, Steven (2022). "We Need to Talk about Random Seeds". In: *CoRR* abs/2210.13393.

Bhattacharjee, Amrita and Huan Liu (2023). "Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text?" In: *SIGKDD Explor.* 25.2, pp. 14–21. DOI: 10.1145/3655103.3655106. URL: https://doi.org/10.1145/3655103.3655106.

Bibal, Adrien, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin (2022). "Is Attention Explanation? An Introduction to the Debate". In: *ACL (1)*, pp. 3889–3900.

Binder, Alexander, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek (2016). "Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers". In: *ICANN (2)*. Vol. 9887. Lecture Notes in Computer Science. Springer, pp. 63–71.

Bogaert, Jérémie, Antonin Descampe, and François-Xavier Standaert (2025). "Consolidating Explanation Stability Metrics". In.

Bogaert, Jérémie, Louis Escouflaire, Marie-Catherine de Marneffe, Antonin Descampe, François-Xavier Standaert, and Cédrick Fairon (2023a). "Sensibilité des explications à l'aléa des grands modèles de langage: le cas de la classification de textes journalistiques". In: *Trait. Autom. Des. Langues* 64.3.

— (2023b). “TIPECS: A Corpus Cleaning Method using Machine Learning and Qualitative Analysis”. In: *International Conference on Corpus Linguistics (JLC)*.
— (2024). “Explanation sensitivity to the randomness of large language models: the case of journalistic text classification”. In: *arXiv preprint arXiv:2410.05085*.
Bogaert, Jérémie and François-Xavier Standaert (2024a). “A Question on the Explainability of Large Language Models and the Word-Level Univariate First-Order Plausibility Assumption”. In: *Responsible Language Models (ReLM)*.
— (2024b). “Does a Reduced Fine-Tuning Surface Impact the Stability of the Explanations of LLMs?” In: *ESANN 2024 - Proceedings*.
Bonin, Patrick, Alain Méot, and Aurélia Bugaiska (2018). “Concreteness Norms for 1,659 French Words: Relationships with Other Psycholinguistic Variables and Word Recognition Times”. In: *Behavior research methods* 50, pp. 2366–2387.
Breiman, Leo (2001). “Statistical modeling: The two cultures (with comments and a rejoinder by the author)”. In: *Statistical science* 16.3, pp. 199–231.
Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). “Language Models are Few-Shot Learners”. In: *NeurIPS*.
Carroll, John B. (1964). “Language and Thought”. In: *Reading Improvement* 2.1, p. 80.
Cavalcanti, Anderson Pinheiro, Rafael Ferreira Mello, Dragan Gašević, and Fred Freitas (2024). “Towards explainable prediction feedback messages using BERT”. In: *International Journal of Artificial Intelligence in Education* 34.3, pp. 1046–1071.
Chefer, Hila, Shir Gur, and Lior Wolf (2021). “Transformer Interpretability Beyond Attention Visualization”. In: *CVPR*. Computer Vision Foundation / IEEE, pp. 782–791.
Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D Manning (2019). “What Does BERT Look at? An Analysis of BERT’s Attention”. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and*

Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019. Ed. by Tal Linzen, Grzegorz Chrupala, Yonatan Belinkov, and Dieuwke Hupkes. Association for Computational Linguistics, pp. 276–286. DOI: 10.18653/V1/W19-4828. URL: https://doi.org/10.18653/v1/W19-4828.

Conroy, Nadia K., Victoria L. Rubin, and Yimin Chen (2015). "Automatic Deception Detection: Methods for Finding Fake News". In: *Proceedings of the Association for Information Science and Technology* 52.1, pp. 1–4.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-Vector Networks". In: *Mach. Learn.* 20.3, pp. 273–297. DOI: 10.1007/BF00994018. URL: https://doi.org/10.1007/BF00994018.

Cover, Thomas M. and Joy A. Thomas (2006). *Elements of information theory (2. ed.)* Wiley.

Danilevsky, Marina, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen (2020). "A Survey of the State of Explainable AI for Natural Language Processing". In: *AACL/IJCNLP*. ACL, pp. 447–459.

Dathathri, Sumanth, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu (2020). "Plug and Play Language Models: A Simple Approach to Controlled Text Generation". In: *ICLR*. OpenReview.net.

de Bodt, C., D. Mulders, M. Verleysen, and J. A. Lee (2020). "Fast Multiscale Neighbor Embedding". In: *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–15. DOI: 10.1109/TNNLS.2020.3042807.

Dehghani, Mostafa, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby (2023). "Scaling Vision Transformers to 22 Billion Parameters". In: *CoRR* abs/2302.05442. DOI: 10.48550/ARXIV.2302.05442. arXiv: 2302.05442. URL: https://doi.org/10.48550/arXiv.2302.05442.

Desrochers, Alain and Glenn Thompson (2009). "Subjective Frequency and Imageability Ratings for 3,600 French Nouns". In: *Behavior research methods* 41.2, pp. 546–557.

Dettmers, Tim, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer (2023). "QLoRA: Efficient Finetuning of Quantized LLMs". In: *NeurIPS*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/n19-1423. URL: https://doi.org/10.18653/v1/n19-1423.

Efron, Bradley and Robert Tibshirani (1993). *An Introduction to the Bootstrap*. Springer.

Escouflaire, Louis (2022). "Identification des Indicateurs Linguistiques de la Subjectivité les Plus Efficaces pour la Classification d'Articles de Presse en Français". In: *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 2: 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL)*, pp. 69–82.

Escouflaire, Louis, Jérémie Bogaert, Antonin Descampe, Cédrick Fairon, and RTBF (2023). "The RTBF Corpus: a Dataset of 750,000 Belgian French News Articles Published between 2008 and 2021". In: *International Conference on Corpus Linguistics (JLC)*.

Escouflaire, Louis, Antonin Descampe, and Cédrick Fairon (2024). "Unveiling Subjectivity in Press Discourse: A Statistical and Qualitative Study of Manually Annotated Articles". In: *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics* 34.

Escouflaire, Louis, Antonin Descampe, Antoine Venant, and Cédrick Fairon (2024). "La subjectivité dans le journalisme québécois et belge: transfert de connaissance inter-médias et inter-cultures". In: *35èmes Journées d'études sur la Parole (JEP 2024) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2024) 26ème Rencontre des étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2024)*. Vol. 2. ATALA & AFPC, pp. 12–13.

Esser, Frank and Andrea Umbricht (2014). "The evolution of objective and interpretative journalism in the Western press: Comparing six news systems since the 1960s". In: *Journalism & Mass Communication Quarterly* 91.2, pp. 229–249.

Frénay, Benoît and Michel Verleysen (2014). "Classification in the Presence of Label Noise: A Survey". In: *IEEE Trans. Neural Networks Learn. Syst.* 25.5, pp. 845–869.

Gémes, Kinga, Ádám Kovács, Markus Reichel, and Gábor Recski (2021). "Offensive Text Detection on English Twitter with Deep Learning Models and Rule-Based Systems". In: *FIRE (Working Notes)*. Vol. 3159. CEUR Workshop Proceedings. CEUR-WS.org, pp. 283–296.

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith (2018). "Annotation Artifacts in Natural Language Inference Data". In: *NAACL-HLT (2)*. Association for Computational Linguistics, pp. 107–112.

Han, Kai, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, and Dacheng Tao (2023). "A Survey on Vision Transformer". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 45.1, pp. 87–110.

Hanawa, Kazuaki, Sho Yokoi, Satoshi Hara, and Kentaro Inui (2021). "Evaluation of Similarity-based Explanations". In: *ICLR*. OpenReview.net.

Hassan, Naeemul, Fatma Arslan, Chengkai Li, and Mark Tremayne (2017). "Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster". In: *KDD*. ACM, pp. 1803–1812.

Hayes, Andrew F. and Klaus Krippendorff (2007). "Answering the Call for a Standard Reliability Measure for Coding Data". In: *Communication Methods and Measures* 1.1, pp. 77–89.

Herman, Bernease (2017). "The Promise and Peril of Human Evaluation for Model Interpretability". In: *CoRR* abs/1711.07414.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Comput.* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Jacob, Benoit, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew G. Howard, Hartwig Adam, and Dmitry Kalenichenko (2018). "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference". In: *CVPR*. Computer Vision Foundation / IEEE Computer Society, pp. 2704–2713.

Jacovi, Alon and Yoav Goldberg (2020). "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In: *ACL*, pp. 4198–4205.

Jiang, Nan-Jiang and Marie-Catherine de Marneffe (2022). "Investigating reasons for disagreement in natural language inference". In: *Transactions of the Association for Computational Linguistics* 10, pp. 1357–1374.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei (2020). "Scaling Laws for Neural Language Models". In: *CoRR* abs/2001.08361. arXiv: 2001.08361. URL: https://arxiv.org/abs/2001.08361.

Katari, Rohan and Madhu Bala Myneni (2020). "A survey on news classification techniques". In: *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*. IEEE, pp. 1–5.

Koren, Roselyne (2004). "Argumentation, enjeux et pratique de l'«engagement neutre»: Le cas de l'écriture de presse". In: *Semen. Revue de sémio-linguistique des textes et discours* 17.

Koufakou, Anna, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti (2020). "HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language". In: *WOAH*. ACL, pp. 34–43.

Kovaleva, Olga, Alexey Romanov, Anna Rogers, and Anna Rumshisky (2019). "Revealing the Dark Secrets of BERT". In: *EMNLP/IJCNLP (1)*. ACL, pp. 4364–4373.

Krüger, Katarina R, Anna Lukowiak, Jonathan Sonntag, Saskia Warzecha, and Manfred Stede (2017). "Classifying news versus opinions in newspapers: Linguistic features for domain independence". In: *Natural Language Engineering* 23.5, pp. 687–707.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut (2020). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: https://openreview.net/forum?id=H1eA7AEtvS.

Latif, Siddique, Aun Zaidi, Heriberto Cuayáhuitl, Fahad Shamshad, Moazzam Shoukat, and Junaid Qadir (2023). "Transformers in Speech Processing: A Survey". In: *CoRR* abs/2303.11607.

Latkin, Carl A, Lauren Dayton, Justin C Strickland, Brian Colon, Rajiv Rimal, and Basmattee Boodram (2023). "An assessment of the rapid decline of trust in US sources of public information about COVID-19". In: *Vaccine Communication in a Pandemic*. Routledge, pp. 22–31.

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain (2018). "The Science of Fake News". In: *Science* 359.6380, pp. 1094–1096.

Lehmann, Erich Leo and Joseph P. Romano (2008). *Testing Statistical Hypotheses, Third Edition*. Springer texts in statistics. Springer. ISBN: 978-0-387-98864-1.

Li, Jiwei, Xinlei Chen, Eduard H. Hovy, and Dan Jurafsky (2016). "Visualizing and Understanding Neural Models in NLP". In: *HLT-NAACL*. ACL, pp. 681–691.

Li, Qian, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He (2020). "A Survey on Text Classification: From Shallow to Deep Learning". In: *CoRR* abs/2008.00364.

Liu, Nelson F., Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith (2019). "Linguistic Knowledge and Transferability of Contextual Representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, pp. 1073–1094. DOI: 10.18653/v1/n19-1112. URL: https://doi.org/10.18653/v1/n19-1112.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692.

Lyu, Qing, Marianna Apidianaki, and Chris Callison-Burch (2022). "Towards Faithful Model Explanation in NLP: A Survey". In: *CoRR* abs/2209.11326.

Manna, Supriya and Niladri Sett (2024). "Faithfulness and the Notion of Adversarial Sensitivity in NLP Explanations". In: *arXiv preprint arXiv:2409.17774*.

Martin, Louis, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot (2020). "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault. Association for Computational Linguistics, pp. 7203–7219. DOI: 10.18653/v1/2020.acl-main.645. URL: https://doi.org/10.18653/v1/2020.acl-main.645.

McAleese, Stephen and Mark T. Keane (2024). "A Comparative Analysis of Counterfactual Explanation Methods for Text Classifiers". In: *CoRR* abs/2411.02643.

McCullagh, Peter and John A. Nelder (1989). *Generalized Linear Models*. Springer. ISBN: 978-1-4899-3242-6. DOI: 10.1007/978-1-4899-3242-6. URL: https://doi.org/10.1007/978-1-4899-3242-6.

McHugh, Mary L (2012). "Interrater reliability: the kappa statistic". In: *Biochemia medica* 22.3, pp. 276–282.

Mieleszczenko-Kowszewicz, Wiktoria, Kamil Kanclerz, Julita Bielaniewicz, Marcin Oleksy, Marcin Gruza, Stanislaw Wozniak, Ewa Dziecіol, Przemyslaw Kazienko, and Jan Kocon (2023). "Capturing human perspectives in NLP: Questionnaires, annotations, and biases." In: *NLPerspectives@ ECAI*.

Miller, Tim (2019). "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: *Artif. Intell.* 267, pp. 1–38.

Mohammad, Saif M. and Peter D. Turney (2013). "Crowdsourcing a Word-Emotion Association Lexicon". In: *Comput. Intell.* 29.3, pp. 436–465.

Montavon, Grégoire, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller (2019). "Layer-Wise Relevance Propagation: An Overview". In: *Explainable AI*. Vol. 11700. Lecture Notes in Computer Science. Springer, pp. 193–209.

Müller, Sebastian, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke (2023). "An Empirical Evaluation of the Rashomon Effect in Explainable Machine Learning". In: *Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference, ECML PKDD 2023, Turin, Italy, September 18-22, 2023, Proceedings, Part III*. Ed. by Danai Koutra, Claudia Plant, Manuel Gomez Rodriguez, Elena Baralis, and Francesco Bonchi. Vol. 14171. Lecture Notes in Computer Science. Springer, pp. 462–478. DOI: 10.1007/978-3-031-43418-1_28. URL: https://doi.org/10.1007/978-3-031-43418-1%5C_28.

Murdoch, W. James, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu (2019). "Interpretable Machine Learning: Definitions, Methods, and Applications". In: *CoRR* abs/1901.04592.

New, Boris, Christophe Pallier, Marc Brysbaert, and Ludovic Ferrand (2004). "Lexique 2: A New French Lexical Database". In: *Behavior Research Methods, Instruments, & Computers* 36.3, pp. 516–524.

Newman, Nic, Richard Fletcher, Craig T Robertson, A Ross Arguedas, and Rasmus Kleis Nielsen (2024). *Reuters Institute digital news report 2024*. Reuters Institute for the study of Journalism.

Peters, Matthew E., Sebastian Ruder, and Noah A. Smith (2019). "To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks". In: *Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019, Florence, Italy, August 2, 2019*. Ed. by Isabelle Augenstein, Spandana Gella, Sebastian Ruder, Katharina Kann, Burcu Can, Johannes Welbl, Alexis Conneau, Xiang Ren, and Marek Rei. Association for Computational Linguistics, pp. 7–14. DOI: 10.18653/v1/w19-4302. URL: https://doi.org/10.18653/v1/w19-4302.

Pirali, Camille, Thomas François, and Núria Gala (2022). "PADDLe: a Platform to Identify Complex Words for Learners of French as a Foreign Language (FFL)". In: *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pp. 46–53.

Plank, Barbara (2022). "The'Problem'of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation". In: *arXiv preprint arXiv:2211.02570*.

Polignano, Marco, Valerio Basile, Pierpaolo Basile, Giuliano Gabrieli, Marco Vassallo, and Cristina Bosco (2022). "A Hybrid Lexicon-Based and Neural Approach for Explainable Polarity Detection". In: *Inf. Process. Manag.* 59.5, p. 103058.

Reif, Emily, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim (2019). "Visualizing and measuring the geometry of BERT". In: *Advances in Neural Information Processing Systems* 32.

Reif, Emily, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim (2019). "Visualizing and Measuring the Geometry of BERT". In: *NeurIPS*, pp. 8592–8600.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (Aug. 2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. en.

arXiv:1602.04938 [cs, stat]. URL: http://arxiv.org/abs/1602.04938 (visited on 09/29/2022).

Ruchansky, Natali, Sungyong Seo, and Yan Liu (2017). "CSI: A Hybrid Deep Model for Fake News Detection". In: *CIKM*. ACM, pp. 797–806.

Sadasivan, Vinu Sankar, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi (2025). "Can AI-Generated Text be Reliably Detected? Stress Testing AI Text Detectors Under Various Attacks". In: *Trans. Mach. Learn. Res.* 2025. URL: https://openreview.net/forum?id=OOgsAZdFOt.

Sap, Maarten, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith (2021). "Annotators with attitudes: How annotator beliefs and identities bias toxic language detection". In: *arXiv preprint arXiv:2111.07997*.

Schudson, Michael (2001). "The objectivity norm in American journalism". In: *Journalism* 2.2, pp. 149–170.

Sen, Cansu, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke A. Rundensteiner (2020). "Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words?" In: *ACL*, pp. 4596–4608.

Sharma, Karishma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu (2019). "Combating Fake News: A Survey on Identification and Mitigation Techniques". In: *ACM Trans. Intell. Syst. Technol.* 10.3, 21:1–21:42.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu (2017). "Fake News Detection on Social Media: A Data Mining Perspective". In: *SIGKDD Explor.* 19.1, pp. 22–36.

Silver, N Clayton and William P Dunlap (1987). "Averaging Correlation Coefficients: Should Fisher's Z Transformation be Used?" In: *Journal of applied psychology* 72.1, p. 146.

Singh, Roshan, Soon Ae Chun, and Vijay Atluri (2020). "Developing machine learning models to automate news classification". In: *Proceedings of the 21st Annual International Conference on Digital Government Research*, pp. 354–355.

Srinivas, Suraj and François Fleuret (2019). "Full-Gradient Representation for Neural Network Visualization". In: *NeurIPS*, pp. 4126–4135.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic Attribution for Deep Networks". In: *ICML*. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 3319–3328.

Tandoc, Edson, Zheng Wei Lim, and Richard Ling (2018). "Defining "Fake News"". In: *Digital Journalism* 6.2, pp. 137–153.

Todirascu, Amalia (2019). "Genre et classification automatique en TAL: le cas de genres journalistiques". In: *Linx. Revue des linguistes de l'université Paris X Nanterre* 78.

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel Bowman (2023). "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting". In: *Advances in Neural Information Processing Systems* 36, pp. 74952–74965.

Vargo, Chris J., Lei Guo, and Michelle A. Amazeen (2018). "The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016". In: *New Media Soc.* 20.5, pp. 2028–2049.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *NIPS*, pp. 5998–6008.

Vernier, Matthieu, Laura Monceaux, and Béatrice Daille (2009). "DEFT'09: détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique". In: *Atelier Défi Fouille de Textes (DEFT'09)*, pp. 101–112.

Voita, Elena, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov (2019). "Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned". In: *ACL (1)*. Association for Computational Linguistics, pp. 5797–5808.

Wankhade, Mayur, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni (2022). "A survey on sentiment analysis methods, applications, and challenges". In: *Artificial Intelligence Review* 55.7, pp. 5731–5780.

Watson, Matthew, Bashar Awwad Shiekh Hasan, and Noura Al Moubayed (2022). "Agree to Disagree: When Deep Learning Models With Identical Architectures Produce Distinct Explanations". In: *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*. IEEE, pp. 1524–1533. DOI: 10.1109/WACV51458.2022.00159. URL: https://doi.org/10.1109/WACV51458.2022.00159.

Weber-Genzel, Leon, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank (2024). "VariErr NLI: Separating annotation error from human label variation". In: *arXiv preprint arXiv:2403.01931*.

Wiebe, Janyce, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin (2004). "Learning subjective language". In: *Computational linguistics* 30.3, pp. 277–308.

Wu, Zhengxuan and Desmond C. Ong (2021). "On Explaining Your Explanations of BERT: An Empirical Study with Sequence Classification". In: *CoRR* abs/2101.00196.

Xu, Wen, Jing He, and Yanfeng Shu (2020). "Transfer Learning and Deep Domain Adaptation". In: *Advances and Applications in Deep Learning*. Ed. by Marco Antonio Aceves-Fernandez. IntechOpen. Chap. 3.

Yeh, Chih-Kuan, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar (2019). "On the (in) fidelity and sensitivity of explanations". In: *Advances in neural information processing systems* 32.

Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi (2019). "Defending Against Neural Fake News". In: *NeurIPS*, pp. 9051–9062.

Zhang, Harry (2004). "The Optimality of Naive Bayes". In: *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA*. Ed. by Valerie Barr and Zdravko Markov. AAAI Press, pp. 562–567. URL: http://www.aaai.org/Library/FLAIRS/2004/flairs04-097.php.

Zhou, Xinyi and Reza Zafarani (2020). "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities". In: *ACM Comput. Surv.* 53.5, 109:1–109:40.

Zhou, Yichu and Vivek Srikumar (2022). "A Closer Look at How Fine-tuning Changes BERT". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Association for Computational Linguistics, pp. 1046–1061. DOI: 10.18653/v1/2022.acl-long.75. URL: https://doi.org/10.18653/v1/2022.acl-long.75.

Zhou, Yilun, Marco Túlio Ribeiro, and Julie Shah (2022). "ExSum: From Local Explanations to Model Understanding". In: *NAACL-HLT*. Association for Computational Linguistics, pp. 5359–5378.

Zini, Julia El and Mariette Awad (2023). "On the Explainability of Natural Language Processing Deep Models". In: *ACM Comput. Surv.* 55.5, 103:1–103:31.

Chapter 7

Appendix

Can Fake News Detection be Accountable? The Adversarial Examples Challenge

(Extended Abstract)

Jérémie Bogaert, Quentin Carbonnelle,
Antonin Descampe, François-Xavier Standaert

UCLouvain, Louvain-la-Neuve, Belgium

Abstract

Automated fake news detection is an important challenge in view of the increasing ability of statistical language models to generate large amounts of (possibly fake) articles, so that recognizing them manually becomes unrealistic. Yet, the reliable deployment of such automated detection tools would require ensuring that they are accountable. Algorithmic accountability is known to be difficult to reach, especially when adversarial behaviors aim to make algorithms deviate from their expected mode of operation. In this paper, we illustrate with a case study that this challenge is further amplified in contexts where the labeling of the articles is prone to errors, which is the case of fake news detection.

1 Introduction

The proliferation of fake news on social media platforms has created an increasing need of solutions to detect them. While the generation of fake news and the necessary fact checking that it implies started as a mostly manual process (e.g., discussed in [17]), recent advances in natural language generation with machine learning algorithms have amplified the risk of so-called neural fake news [19]. The massive amount of articles that such tools can generate makes manual fact checking unrealistic and raises the question of their automated detection. As recently surveyed in [1, 20], solutions combining natural language processing and machine learning are attractive for this purpose, due to their ability to capture various features of newspaper articles. Besides, the sensitive nature of the fake news detection problem also requires that its deployment comes with guarantees of algorithmic accountability [3]. But as discussed in [2], such guarantees are hard to obtain in contexts where adversarial behaviors can target the robustness of machine learning classifiers, which is the case of fake news detection.

In this paper, we therefore study the robustness of fake news detection against adversarial examples [4, 8]. For this purpose, we first build a database of fake and reliable news. As already observed in the literature, this step is inherently challenging since there is no strict definition of what a fake news is [6]. We deal with this difficulty by combining a public dataset of fake news (`https://www.kaggle.com/mrisdal/fake-news`), which were scraped from blacklisted websites over a period of 30 days around the 2016 US election, and reliable news collected from the New York Times and the Guardian over approximately the same period. We additionally explore both data sets to show that they cover similar topics. While inevitably imperfect, this database is then used to show that standard machine learning classifiers can detect fake news with a good accuracy, but are also easy to fool with adversarial examples. Interestingly, it appears that the difficulty to define what a fake news is (possibly combined with label errors in the training sets) gives adversaries additional opportunities to generate fake news classified as reliable. So our results question again the possibility

to rely on accountable algorithms for sensitive tasks that can be targeted by adversarial behaviors and suggest different research avenues related to fake news in general.

We note that the topic of fake news is widely multidisciplinary [7] and even the more technical topic of automated fake news detection is already covered by a broad literature (e.g., the already mentioned [1, 19, 20] but also [5, 11, 13, 12] to name a few). Our contribution is not to improve automated fake news detection algorithms but to illustrate the difficult interplay between such algorithms and the need of algorithmic accountability when adversarial behaviors are considered. As a natural starting point in this direction, we show that there exist (for now simple) examples of fake news detectors that are weak against such adversarial behaviors, raising the question of how to prevent them, with technical and non-technical means. In other words, we put forward an under-discussed risk for the reliable deployment of such systems.

The rest of the paper is structured as follow. We start by describing the database we used for our investigations and discussing its unavoidable limitations in Section 2. We follow by showing simple examples of supervised fake news detection tools that can detect fake news with reasonable accuracy in Section 3. We finally exhibit how to craft adversarial examples against these classifiers in Section 4. We conclude by analyzing the impact of these findings and tracks for further research in Section 5.

2 Building a fake news database

Research on fake news detection has often been limited by the quality of existing datasets and their specific application contexts [12]. As already mentioned, this is mostly due to the difficulty of precisely defining what a fake news is, making their labeling for supervised learning challenging [6]. Besides, our goal to investigate adversarial examples is typically calling for "not too short" texts, excluding popular databases such as [18]. To the best of our knowledge, the database that comes the closest to our needs is the fake news corpus (`https://github.com/several27/FakeNewsCorpus`). However, preliminary investigations suggested quite significant dissimilarities between the topics of reliable and fake articles (namely, football for fake articles and politics for reliable ones). This similarity makes it unsuitable for our purposes, since it implies risks to classify the articles based on their topic more than their fake/reliable nature. We therefore prepared our investigations by attempting to build a better database.

Concretely, we started from a public dataset of 3500 fake news from Kaggle (`https://www.kaggle.com/mrisdal/fake-news`), which were scraped from 207 blacklisted websites over a period of 30 days around the 2016 US election. Some preliminary processing was applied to remove unwanted features of the articles (e.g., meta-data that is not in the original articles and was added by the blacklisted websites). No website contributed to more than 1% of the database to ensure that imperfect scraping, preprocessing or labeling for some websites cannot significantly impact the overall results. We then tried to build a complementary database of reliable news, covering the same period of time. For this purpose, we used the API developed by the New York Times and The Guardian, and scraped papers about World news and US news for the intended period. We had to slightly increase the window of time (Oct. 16 to Dec. 4 for the reliable news vs. Oct. 26 to Nov. 25 for the fake news), in order to collect 3500 articles (1500 from the New York Times, 2000 from The Guardian).

We performed a preliminary exploration of the topics covered by this database using the Term Frequency–Inverse Document Frequency (TF-IDF) statistic. It is a standard tool for information retrieval or summarization, which indicates the relevance of the words in some documents [10]. We first launched it on the full database to identify the most relevant words overall. Next, we launched it on the fake news corpus, leading to the results of Figure 1 (where the X axis lists the 40 most relevant words of the full

database). It confirms the coverage of the US elections (e.g., with Trump and Clinton in the first places), but also highlights the weight of some neutral words like 'said'.

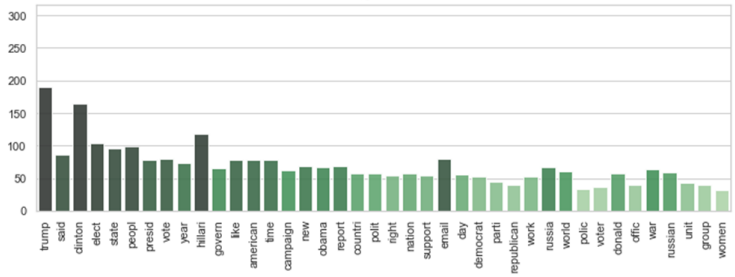


Figure 1: Evaluation of the fake news database's topics with TF-IDF score.

We finally computed the same TF-IDF scores for the reliable news, which are given in Figure 2 and confirm the coverage of similar topics. They also highlight some specific features of the fake news corpus already: for example the more frequent use of the (misspelled) first name Hilari or the importance of the word e-mail (relating to an ongoing affair of Mrs. Clinton's e-mail leaks during the 2016 elections).

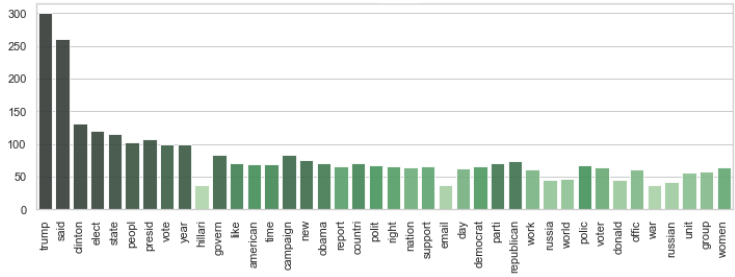


Figure 2: Evaluation of the reliable news database's topics with TF-IDF score.

In order to confirm that the resulting (7000-paper) database did not contain obvious parasitic patterns, we additionally visualized the data by feeding the t-distributed Stochastic Neighbor Embedding (t-SNE) tool of [16] with the vectors output by the TF-IDF transform.* t-SNE projects each high-dimensional object towards a two-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability. As illustrated in Figure 3, the distribution of the topics is similar for fake and reliable articles while, for example, the separation between World and US (reliable) news can be distinguished in the right part of the figure. It suggests that the topics do not create obvious (parasitic) ways to discriminate fake and reliable news. So while the evaluations in this section do admittedly not provide any formal guarantee that no such patterns exist, we assume in the following that this database is good enough for our purposes.

* Reduced to 500 dimensions thanks to truncated Singular Value Decomposition (SVD). We also tried larger number of dimensions but it did not change our main observations.

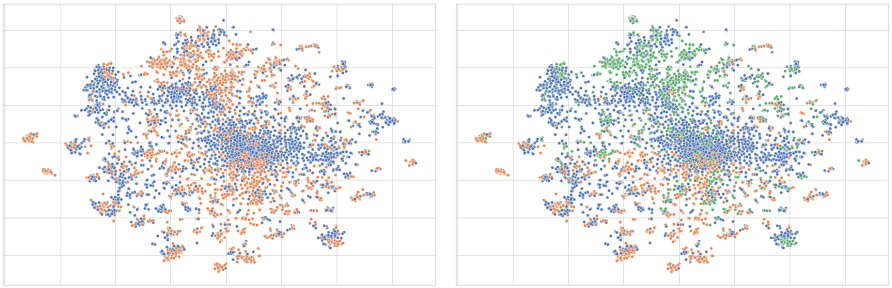


Figure 3: Visualization of the news' topics with t-SNE. Left: fake news (•) and reliable news (•). Right: fake news (•), reliable World news (•) and reliable US news (•).

3 Exemplary classifiers

The next step of our investigations was to build statistical classifiers for fake and reliable news, thanks to supervised machine learning. We considered various options for this purpose: logistic regression, naive Bayes, random forests and Long Short-Term Memory (LSTM). For place constraints, we focus our following descriptions on logistic regression and LSTM (the other classifiers did not significantly affect our main conclusions).

Concretely, all our classifiers were built starting with the same preprocessing: we removed punctuation, non-alphanumeric characters, multiple whites spaces, websites, stop words and short words). For the logistic regression, we then vectorized the articles as bag of words using TF-IDF while for the LSTM we used a Word2Vec model trained on our full dataset [9]. As Google's Word2Vec (`https://code.google.com/archive/p/word2vec/`), we fixed the embedding to 300-dimensional vectors. The rationale behind this choice is that the LSTM can take advantage of the words' order. The model hyperparameters were then tuned using a 5-fold cross-validation.

The accuracy of these classifiers is illustrated in Figure 4. It is significantly better than a random guess in both cases, and reaches $> 90\%$ for the logistic regression (we expect that the LSTM would reach this accuracy or even improve it with more training samples). Despite further optimizations are certainly possible, we assume these values to be sufficient for trying to decrease them with adversarial examples.

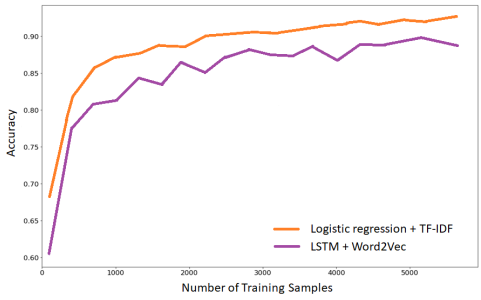


Figure 4: Learning curve of exemplary fake news detectors.

4 Crafting adversarial examples

An adversarial example is an input, unknown to the target machine learning algorithm, that makes it deviate from its public specifications and is purposely generated by an adversary having an interest in such a deviated behavior [4]. In the context of fake news detection, the goal of the adversary is to force the misclassification of a fake news as reliable, despite not changing its semantic content from the human perception viewpoint. The main question we tackle next is whether such adversarial examples can be crafted for the detectors of the previous section. In this context, the adversarial goal could vary from producing such adversarial examples at a low rate, possibly taking advantage of some manual processing, or to produce them at a high rate, automatically. As a first step to show the existence of a risk, we consider the easiest (low rate with manual intervention) context. We briefly discuss the relaxation of this context at the end of the section. The threat model could also vary from white box access to the models (i.e., knowing their parameters) to only black box access (i.e., only being able to observe input/output pairs). We discuss both options in the following.

4.1 Methodology

The high-level approach we used to craft adversarial examples at low rate is in 3 steps:

1. Identify Reliable Hot Words (RHW), which are the words having high probability to push the classification of any article towards the reliable class.
2. Identify the Target's Fake Hot Words (TFHW), which are the words having high probability to push the classification of a target article towards the fake class.
3. Human intervention to replace TFHW by RHW in a semantic-preserving manner.

The identification of the RHW and TFHW was performed using high-level ideas similar to [8, 2]. In the white box setting, we applied the Fast Gradient Sign Method (FGSM) which essentially boils down to maximizing the classifier's loss function, then using the gradient information to add or remove words independent of their position.[†] This gradient can be computed analytically for the logistic regression, and we used the numerical estimation provided by TensorFlow (`https://www.tensorflow.org/`) for the LSTM. In the black box setting, we used a more straightforward approach where we just removed words one by one and feed the classifier with the modified articles in order to identify words that impact the detection probability the most.

4.2 Experimental results

We will discuss the type of results we obtain with Example 1, which is initally detected as a fake news with 90% probability by the logistic regression, and 65% probability by the LSTM. As aforementioned, the first step in trying to fool the detection is to identify RHW and TFHW. RHW are identified on the reliable news corpus while TFHW are identified on the target article only. We illustrate this step with our black box approach applied to the beginning of the example (namely, the words *In what is being described as another 'bizarre'*), where the short words 'In', 'what', 'is', 'being' and 'as' do not impact the classification, while removing the words 'described' and 'bizarre' respectively decrease its probability of being detected as a fake news by 2.8% and 3.5%. By extending such an approach to the whole article (and the whole corpus of reliable news for RHW), we can then build lists of TFHW and RHW.

[†] This position could be exploited in the case of the LSTM, which we leave as a scope for further investigations (a direct exploitation would result in a quite computationally intensive strategy).

> In what is being described as another 'bizarre' attempt to sabotage her own campaign, Hillary Clinton has desecrated a series of beloved US symbols, including punching a bison, setting fire to the Stars & Stripes and spitting at Jerry Seinfield. The Presidential hopeful seems determined to make a series of unprovoked errors, not least of which was agreeing to Bill hosting a sleepover for a group of Girl Guides. Short of dressing the Statue of Liberty in a Burka, Mrs Clinton has lurched from one PR blunder to another. Commented one journalist: 'The Presidential race is entering the final furlong and if Mrs Clinton was horse – and before you can say Benghazi – she's gone from bookie's favourite to an ingredient at the local glue factory'. Having already become the unwitting focus of various health scares and FBI investigations, Mrs Clinton's campaign is as orderly as a Marx Brothers movie. Her lead in the polls has been cut as video emerges of her lighting a cigar with a rolled up Bill of Rights, then proceeding to take a dump on the White House lawn. Hillary's erratic behaviour has seen her sing the Star-Spangled Banner in Korean, dress as Oprah Winfrey for Halloween and pebble-dash Mount Rushmore. Remarked a flummoxed advisor: 'She keeps doing the unthinkable – like making Donald Trump electable'. Share this story...

Example 1: Sample of the fake news database.

The main high-level observations that can be extracted from these lists are:

1. That they contain both semantically tainted words (e.g., 'Hilary', 'FBI', 'Donald') and semantically neutral words (e.g., 'said', 'commented', 'campaign').
2. That there is a higher proportion of semantically tainted words for TFHW.
3. That the lists of (ordered) words identified as RHW or TFHW for the logistic regression and the LSTM, significantly overlap but are not identical.

We can then build adversarial examples, e.g., for the (simpler) LSTM:

> [...] 'bizarre' attempt to sabotage her own campaign, ~~Hillary~~ **Mrs** Clinton has desecrated
> [...] Mrs Clinton's campaign is as ~~orderly~~ **neat** as a Marx Brothers movie [...]

By changing only two words (which do not affect the meaning of the article), it is now classified as a fake with only 45% probability (so as reliable with 55% probability). Interestingly, exactly those changes are not sufficient to misclassify the article with the logistic regression (which still classifies it as a fake, but with a probability reduced from 90% to 55%). Yet, another change of the second word does the job (suggesting that as usual with adversarial examples, they have a certain level of transferability):

> [...] 'bizarre' attempt to sabotage her own campaign, ~~Hillary~~ **Mrs** Clinton has desecrated
> [...] Remarked a flummoxed ~~advisor~~ **minister**: 'She keeps doing the unthinkable [...]

Since based on manual interventions, we did not craft such examples for large number of articles. We repeated the process for 5 fake news and succeeded to force a misclassification by changing a maximum of 15 words. The LSTM classifier was usually fooled with slightly less words which we assume is due to its initially lower accuracy.

Overall, and despite drawing general conclusions based on such small-scale examples is hard, one important observation is that in contrast with the case study of [2] where adversarial examples had to be mostly based on neutral words (to avoid being easy to spot by the human perception), fooling a fake news detection algorithm can be done with more tainted words (e.g., by changing 'Hillary' into 'Mrs' in our example). In other words, the fact that the fake or reliable nature of an article does not have a definition as clear as the topics used in the case study [2] makes the adversary's task easier.

4.3 Towards automation

Our handmade experiments naturally raise the question whether adversarial examples can be automated. As a first step in this direction (i.e., to evaluate the sensitivity of

our fake news detectors to semantically-neutral modifications), we evaluated a greedy strategy where we just substituted words by synonyms (sometimes causing syntax problems). One example obtained for the logistic regression is given below:

> [...] symbols, including punching a ~~bison~~ **buffalo** [...] [...] a group of ~~Girls~~ **Woman** guides
> [...] various health scares and FBI ~~investigations~~ **inquiry**, Mrs Clinton's campaign [...]

Out of 100 test articles, we could misclassify 22% (resp., 32%) with the LSTM (resp., logistic regression) classifier (presumably because our greedy strategy is better suited to models that do not exploit the words' order). sing advanced statistical language models should lead to more adversarial opportunities, which we leave as an open problem.

5 Conclusions and open problems

Advances in digital media are responsible for journalists to loose the monopoly on information production and dissemination, and raise new legitimacy concerns (e.g., regarding whether journalism can still offer quality and reliable news in the digital era) [15]. Algorithmic accountability is one of the emerging (and difficult to reach) goals aiming to mitigate such concerns [3]. In this work, we confirm that ensuring algorithmic accountability with robustness against adversarial behaviors is especially challenging in contexts such as fake news detection where the definition of optimization criteria is inherently fuzzy due to the lack of well defined classes. Our results are preliminary in many respects and therefore suggest various directions for further investigations. First, the difficult collection of a fake news / reliable news database would be improved by designing a tool able to automatically generate large amounts of fake news from reliable ones (e.g., by biasing them in a given direction). It would allow a better analysis of the syntactic or semantic patterns that fake news detection can exploit. Second, and as already mentioned, the generation of adversarial examples would benefit from automation in order to enable larger-scale experiments. Third, the evaluation of countermeasures (i.e., fake news detection tools able to cope with adversarial examples) is an important long-term goal as well. Early discussions in [4, 2] suggest that a purely technical solution may not be possible. The perspective of such negative conclusions therefore questions the need of complementary approaches, e.g., relying on the trust in the journalists who write stories more than on content, as suggested in [14].

References

[1] Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1):1–4, 2015.

[2] Antonin Descampe, Clément Massart, Simon Poelman, François-Xavier Standaert, and Olivier Standaert. Automated news recommendation in front of adversarial examples and the technical limits of transparency in algorithmic accountability. AI & Society, 2021.

[3] Nicholas Diakopoulos. Algorithmic accountability. Digital Journalism, 3(3):398–415, 2015.

[4] Ian J. Goodfellow, Patrick D. McDaniel, and Nicolas Papernot. Making machine learning robust against adversarial inputs. Commun. ACM, 61(7):56–66, 2018.

[5] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In KDD, pages 1803–1812. ACM, 2017.

[6] Edson C. Tandoc Jr., Zheng Wei Lim, and Richard Ling. Defining "fake news". Digital Journalism, 6(2):137–153, 2018.

[7] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. Science, 359(6380):1094–1096, 2018.

[8] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. Deep text classification can be fooled. In IJCAI, pages 4208–4215. ijcai.org, 2018.

[9] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In ICLR (Workshop Poster), 2013.

[10] Juan Ramos. Using tf-idf to determine word relevance in document queries. In Proceedings of the First Instructional Conference on Machine Learning, pages 29–48, 2003.

[11] Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI: A hybrid deep model for fake news detection. In CIKM, pages 797–806. ACM, 2017.

[12] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. ACM Trans. Intell. Syst. Technol., 10(3):21:1–21:42, 2019.

[13] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. SIGKDD Explor., 19(1):22–36, 2017.

[14] David Sterrett, Dan Malato, Jennifer Benz, Liz Kantor, Trevor Tompson, Tom Rosenstiel, Jeff Sonderman, and Kevin Loker. Who shared it?: Deciding what news to trust on social media. Digital Journalism, 7(6):783–801, 2019.

[15] Jingrong Tong. Journalistic legitimacy revisited. Digital Journalism, 6(2):256–273, 2018.

[16] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of Machine Learning Research, 9(86):2579–2605, 2008.

[17] Chris J. Vargo, Lei Guo, and Michelle A. Amazeen. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. New Media Soc., 20(5):2028–2049, 2018.

[18] William Yang Wang. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In ACL (2), pages 422–426. Association for Computational Linguistics, 2017.

[19] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In NeurIPS, pages 9051–9062, 2019.

[20] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Comput. Surv., 53(5):109:1–109:40, 2020.

Automatic and Manual Detection of Generated News: Case Study, Limitations and Challenges

Jérémie Bogaert
UCLouvain
Belgium
jeremie.bogaert@uclouvain.be

Marie-Catherine de Marneffe
The Ohio State University
USA
demarneffe.1@osu.edu

Antonin Descampe
UCLouvain
Belgium
antonin.descampe@uclouvain.be

François-Xavier Standaert
UCLouvain
Belgium
fstandae@uclouvain.be

ABSTRACT

In this paper, we study the exploitation of language generation models for disinformation purposes from two viewpoints. Quantitatively, we argue that language models hardly deal with domain adaptation (i.e., the ability to generate text on topics that are not part of a training database, as typically required for news). For this purpose, we show that both simple machine learning models and manual detection can spot machine-generated news in this practically-relevant context. Qualitatively, we put forward the differences between these automatic and manual detection processes, and their potential for a constructive interaction in order to limit the impact of automatic disinformation campaigns. We also discuss the consequences of these findings for the constructive use of natural language generation to produce news items.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; *Supervised learning by classification*; *Machine learning approaches*; • **Applied computing** → *Annotation*.

KEYWORDS

Fake news detection, digital journalism

ACM Reference Format:
Jérémie Bogaert, Marie-Catherine de Marneffe, Antonin Descampe, and François-Xavier Standaert. 2022. Automatic and Manual Detection of Generated News: Case Study, Limitations and Challenges. In *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation (MAD '22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3512732.3533589

1 INTRODUCTION

The spreading of misinformation, disinformation and fake news on social media platforms has been a topic of increasing interest over recent years [**?**]. Current advances in Natural Language Generation (NLG), based on neural machine learning, make it easy to generate massive amounts of seemingly consistent texts that can subsequently contribute to increase information disorder. There has therefore been work for automatically detecting misleading content under the term of "fake news detection". However, it is important to note that this general research direction covers quite different operational goals (see for example [**? ? ?**] for surveys and [**? ? ? ?**] for a few technical case studies).

One possible goal is to gauge the veracity of a news item. But this goal faces the (non-technical) challenge of defining what a fake news is [**?**]. Besides, many different elements can be taken into account to detect whether news are genuine or fake: the source of the information and the sites, people or platforms relaying them, the author's reputation, a fact-checking process confronting the content with an external knowledge base, the style and the lexical field found in the text, etc. Among these elements, some are subjective (e.g., the author's or publishing sites' reputation) and most are based on contextual information not found in the news content itself. As a result, identifying labeled data sets for supervised machine learning in the context of fake news detection is inherently error-prone, which potentially implies label noise [**?**]. It also makes the reasons for which news are detected as fake using such data sets hard to interpret.

In order to circumvent these difficulties, fake news detection is sometimes restricted to the (better defined) problem of distinguishing machine-generated news from human-generated ones [**?**]. Such a step back is interesting since it provides a basis for the detection of neural potential fake news, or more generally for the detection of mass disinformation relying on automatic text generation. We follow this line of work and extend it from two perspectives.

First, we study quantitatively the news generated by a state-of-the-art language generation model called `Grover` [**?**]. While such an evaluation is generally provided in papers introducing new language models (including `Grover`), we extend it in two directions. On the one hand, we observe that news in general, and fake news in particular, are usually targeting recent and fast evolving topics. We therefore build an experiment where classifiers are required to detect whether submitted news are human- vs. machine-generated, for text that is more and more remote from the training database. We consider different classifiers for this purpose (namely logistic

MAD '22, June 27–30, 2022, Newark, NJ, USA

ACM ISBN 978-1-4503-9242-6/22/06…$15.00
https://doi.org/10.1145/3512732.3533589

regression, support vector machine, naive Bayes, and long short-term memory). On the other hand, we evaluate a manual detection experiment performed over four weeks, with 30 participants, each of them asked to evaluate 10 human- or machine-generated pieces of news per week. Unsurprisingly, our results confirm that domain adaptation is a challenge for language (as for other) models [**?**]: text generated by `Grover` for unseen topics is identified as machine-generated both by simple automatic tools and by manual detection. These results thus raise the question of how fast such language generation models can be adapted to newsfeeds, e.g., thanks to emerging plug-and-play language models [**?**].

Second, we investigate qualitatively the decisions made by the two detection processes (automatic and manual), and the reasons for which news are labeled as human- or machine-generated by the annotators in our manual-detection experiment. We first observe that news items that are incorrectly classified are not systematically the same for the two types of detection. We then leverage the reasons our annotators gave as open answers in our manual-detection experiment to explain their label choice (machine- or human-generated). Practically, we propose a systematization of the reasons in four types (syntax, semantic, background, other), each of them split into four subtypes (detailed in subsection 5.1). Discussing the possibility (or lack thereof) of exploiting these reasons in automatic detection tools, we highlight a potential complementarity between the automatic and manual detection of machine-generated news. We finally conclude the paper by motivating such a combined detection with relevant examples taken from our manual detection experiment.

The rest of the paper is structured as follows. In Sections refsec:design and 3, we describe the two main steps of our experiment, namely the selection of human-generated news and the generation of news with a language model, on topics that are increasingly remote from the training database. In section 4, we detail the quantitative part of our results and confirm the challenge of domain adaptation for language models in front of both automatic and manual detectors. In section 5, we detail the qualitative part of our results and exhibit differences between the automatic and manual detection of machine-generated and human-generated news. We conclude by putting forward the possibility of combined detection as an interesting direction for further research.

2 NEWS SELECTION & GENERATION

In this paper, we aim to evaluate the language generation capacity of the `Grover` language model for news that are increasingly remote from the training database. In this section, we describe our methodology for generating such news. For this purpose, we first identify some of the main topics covered by the `RealNews` database on which `Grover` was trained [**?**]. We then explain the (more and more challenging) categories of news that we used in our experiment.

2.1 News topics

`Grover` has been released in 2019 and was trained on the `RealNews` database. This database is composed of 37 million news which are a subset of `CC-NEWS`, a larger dataset available via common crawl, updated daily and containing news collected from information websites all over the world.[1] Each piece of news contains multiple fields such as the publication date, the author(s)' name(s), the title and the main text. `Grover` is able to generate using only some of these fields, like the title and the date. It is available in three versions (base, large and mega). We next use the base version. The motivations for selecting `Grover` are twofold: first, it is a state-of-the-art language generation model available in open source; second it allows comparison with the NeurIPS 2019 paper of Zellers et al. Yet, other language models could be considered, which we leave as an open problem.

In order to gain some understanding of the `RealNews`' topics, we applied Latent Dirichlet Allocation (LDA) on a fraction of its articles. LDA is a statistical model that allows clustering by assuming each data point to be a mixture of some unobserved groups [**?**]. For text data, the components of the mixture are the topics and they are identified thanks to the presence of representative words in each document. We used the `pyLDAvis` package for this purpose.[2] Figure 1 illustrates the results that we obtained for exemplary clusters and topics. For each cluster (on the left of the figure), LDA outputs a list of associated words (on the right of it).

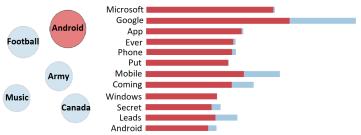


Figure 1: LDA illustration. Exemplary clusters are on the left. Associated words are on the right, with the ratio between their number of appearances in the "Android" cluster (in red) and their number of appearances in any clusters (in blue).

Concretely, we used LDA to identify 25 clusters from which we manually isolated 15 ones such that their associated words were easily and naturally connected with a well identified topic. Namely: *Black Friday, Facebook, Wall street, Football, American Senate, Air Crash, Olympic Games, Android, Healthcare, Army, Music, Immigration, Christmas, Education* and *Canada*. Note that we primarily selected these (timeless) topics because they are quite unrelated to each other. For completeness, we additionally selected 5 event-triggered topics: *Metoo Movement, Notre-Dame Fire, Trump's Election, Ebola* and *Brexit*. As will be discussed in section 4, the timeless or event-triggered nature of the topics does not have a significant impact on our main conclusions.

Besides these 15+5 topics, we also chose topics outside the `RealNews` database. For this purpose, we adopted the even simpler approach of selecting articles of `CC-NEWS` that are subsequent to the ones in `RealNews`. As illustrated in Figure 2, the `RealNews` articles' were all published before 2019 (more precisely April 2019 which is `Grover`'s release date). Therefore, we manually selected 5 other topics that took place after this date: *Joe Biden's Election, Covid19, Beiruth Explosion, Georges Floyd's Death* and *Capitol Invasion*.

[1] https://commoncrawl.org.
[2] https://github.com/bmabey/pyLDAvis.

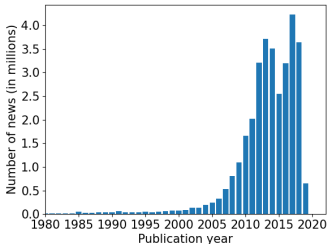


Figure 2: Publication years of `RealNews` articles.

2.2 News categories

The topics' selection described above provided us with the human-generated news for our experiment. We now detail how we produced their machine-generated counterpart. As already mentioned, `Grover` can generate news based on their title and date only, but can additionally exploit other fields. We thus defined our news categories as follows:

- **In RealNews-Full.** Human-generated news from the `RealNews` database. Machine-generated news created using all the fields of the human-generated news. This category focuses on the topics: *Black Friday, Facebook, Wall Street, Football & American Senate.*
- **In RealNews-Title/Date.** Human-generated news fr- om the `RealNews` database. Machine-generated news created using the date/title fields of the human-genera- ted news. This category focuses on the topics: *Air crash, Olympic Games, Android, Healthcare & Army.*
- **Topic in RealNews.** Human-generated news outside the `RealNews` database on topics seen in it. Machine-generated news created using the date/title fields of the human-generated news. Selected topics: *Music, Immigration, Christmas, Education & Canada.*
- **Topic out of RealNews:** Human-generated news outside the `RealNews` database, on topics that are not seen in it. Machine-generated news created using the date/title fields of the human-generated news. Selected topics: *Joe Biden's Election, Covid19, Beiruth Explosion, Georges Floyd's Death & Capitol Invasion.*

The first two categories are aimed to determine whether the addition of fields beyond the title/date of machine-generated news has a significant impact on their quality. The last two ones are increasingly remote from `Grover`'s training database and are aimed to answer our research question: how well can a language model deal with domain adaptation? Coming from `CC-NEWS`, the articles in these categories have the same format as those in the two first categories.

3 MACHINE-GENERATED NEWS DETECTION

We now move to the description of the automatic and manual detections that we aim to evaluate in our experiment.

3.1 Automatic detection

We trained four machine learning models in order to classify human-generated and machine-generated news:

- **Logistic Regression** (LR) is a staple for classification tasks in general, and it is often used for text classification [**?**]. It models the relation between a response variable and one or more explanatory variables as a logistic function. Logistic regression is flexible, easy to use and usually leads to interpretable results [**?**].
- **Support Vector Machines** (SVM) are another widely used algorithm for classification problems [**?**]. Intuitively, they implement the idea that the inputs to classify can be non-linearly mapped to a high-dimension feature space where a linear decision surface is constructed in order to discriminate pairs of classes [**?**].
- **Naive Bayes** (NB) is a probabilistic classifier that assumes independence between the features to apply Bayes' Theorem. It is one of the simplest models, but can often achieve surprisingly good results [**?**].
- **Long Short-Term Memory** (LSTM) is a more complex model that became popular over the last years [**? ?**]. In contrast with the previous tools, it is able to deal efficiently with sequentially dependent data (which is usually the case of news) [**?**]. LSTM is often considered as a baseline deep learning approach.

As usual for text classification, all the news we used for training and testing models were pre-processed using standard techniques. First, we tokenized news into words and assigned part-of-speech to each word. We then removed all the not fully alphabetical words (e.g., numbers) and we lemmatized the remaining ones using the WordNetLemmatizer from the nltk.stem package.[3] We finally vectorized these lemma into numbers. For the LR, SVM and NB models, we used the simple Term Frequency-Inverse Document Frequency (TF-IDF) statistic to produce vectors of words' weights [**?**]. For the LSTM model, we used the words embedding layer available in the Keras package.[4] The rationale behind this choice is that the LSTM can take advantage of the words' order.

3.2 Manual detection

In order to evaluate the previous automatic detection in front of a manual detection and put forward their potential differences, we set up a manual detection experiment with 30 participants during 4 weeks. At a high level, we asked participants to decide whether pieces of news they read were human-generated or machine-generated and to give a confidence score for their decision. Additionally, an open field allowed them to explain their choices. We built the experiment according to the following partition rules:

- Each participant reviews 10 news items per week.
- Each of these news item is on a different topic.
- The 10 news are balanced: they contain 5 human-generated and 5 machine-generated news.
- Participants never receive both a human-generated news and its machine-generated counterpart.

[3] https://www.nltk.org/api/nltk.stem.html.
[4] https://keras.io/api/.

Participants were not informed of these partition rules.

As usual in statistical evaluations, we designed our experiment to be able controlling different plausible sources of variance that could affect our conclusions. That is, while we are primarily interested in determining whether the (automatic and manual) detection becomes harder with the more and more challenging categories of news defined in the previous section, we also investigated other explanatory variables. Namely, we first analyzed the impact of the type of topics (i.e., timeless or event-triggered) and the weeks (to see whether the participants' detection ability evolves over time). We next analyzed the impact of the title. For this purpose, participants were asked to give a first decision without seeing the news' title and to confirm (or change) it after seeing the title. Finally, we split the participants into groups according to their background (junior engineering, senior engineering, humanities and life sciences). Each group reviewed the same news items, meaning that each item is annotated 3 times (once per group). Participants were not rewarded.

Concretely, this experiment was run on a website that we designed rather than on platforms such as Amazon Mechanical Turk.[5] On the one hand, it allows having a better control of the experiment plan. On the other hand, it limits the amount of answers since the selection of participants and what was requested of them was more constrained.

4 QUANTITATIVE ANALYSIS

In this section, we present the quantitative results of our experiment. We start by detailing how we selected the parameters of our machine learning models and give one model's learning curves for illustration. We next describe our main result, namely the impact of the news categories on the accuracy of the automatic and manual detections. We note that we use the (easiest to interpret) accuracy metric because of our balanced experiment plan. We finally discuss the impact of the other explanatory variables of our experiment.

Model parameters. In order to avoid overfitting, we built all our models based on the following process. For each news category, we first extracted a validation set of 400 news out of the 2,000 ones available. Within the 1,600 remaining items, we then selected the model parameters by using a 4-fold cross validation (in which all the sets are balanced and contain the same number of human- and machine-generated news). For all models, the best parameters were obtained by performing a grid search (i.e., testing parameter combinations). For the LSTM, we additionally selected the number of epochs such that the training and test losses don't diverge for more than 5 epochs. Having found the best parameters, we finally re-trained the model on the 1,600 (training and test) news and evaluated their accuracy on the validation set.

The training curves of the LR model for the different news categories, where we artificially reduce the size of the training set, are in Figure 3. We observe a good convergence and similar results were obtained for the other models. For the last point of the curves, we additionally report the following (Gaussian) confidence interval:

$$\left[\overline{\alpha} - z * \sqrt{\frac{\overline{\alpha} * (1-\overline{\alpha})}{n}}, \overline{\alpha} + z * \sqrt{\frac{\overline{\alpha} * (1-\overline{\alpha})}{n}}\right],$$

where $\overline{\alpha}$ is the estimated accuracy, z is set to 1.96 for a 95% confidence interval and $n = 400$ is the size of the validation set. Similar results were obtained with bootstrap confidence intervals (i.e., without the Gaussian assumption).

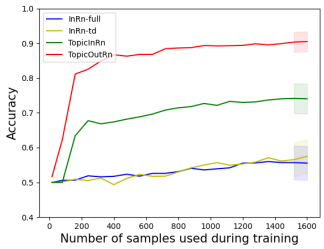


Figure 3: Learning curves of the LR detection tool.

Results. Figure 4 summarizes the results we obtained for our four machine learning models (i.e., the last point of their training curves) and for the manual detections of our experiment reported with the same confidence interval.

As far as the automatic detection is concerned, we observe that all the models lead to similar results and the three main categories of news (i.e., "in `RealNews`", "Topic in `RealNews`" and "topics out of `RealNews`") indeed lead to increasingly easy detections, reflected by a higher accuracy. In particular, the "topic out of `RealNews`" category leads to detections with an accuracy of approximately 90%. By contrast, the addition of all the fields (besides the title and the date that are always used) to the news generation does not have a significant impact on the detection (see for example "inRn-full" vs. "inRn-td" in Figure 4). The same observation holds for the type of topics, as illustrated in Appendix, Figure 8, where we additionally analyzed the "topic in `RealNews`" category with both timeless and event-triggered topics.

Comparing these accuracies with the ones reported in [**?**], where `Grover` is used in order to detect the news it generated, we observe that much simpler models allow detecting machine-generated news when domain adaptation is required (i.e., for the "topic out of `RealNews`" category). So these results support the claim that detecting machine-generated news on fast evolving topics, that typically require domain adaptation, is feasible even without access to complex (and expensive) models. By contrast, and unsurprisingly, the accuracies we reach for the "in `RealNews`" and "Topic in `RealNews`" categories are significantly lower than in [**?**], presumably due to the simpler models we use.

As for the manual detection, the results obtained are with lower statistical confidence due to the limited amount of samples we could evaluate (100 per category vs. 2,000 per category for automatic

[5] https://www.mturk.com/.

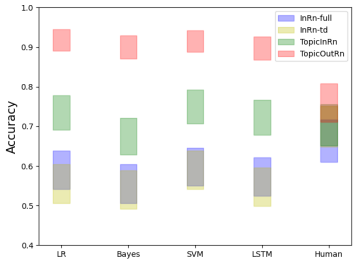


Figure 4: Accuracy of automatic and manual tools to detect human-generated from machine-generated news.

detection). Yet, we observe that the accuracy of the "in `RealNews`" category is confidently lower than the one of the "topics out of `RealNews`" category. Interestingly, we also observe in Figure 5 that the accuracy of the manual detectors depends on the (real) label of the news to classify, and is significantly higher when the items are human-generated, which is in contrast with the automatic detection, where human-generated and machine-generated news are classified with similar accuracies. So this figure provides a first indication that automatic and manual detection may not exploit the same features of the news, which we will further discuss in the following section.

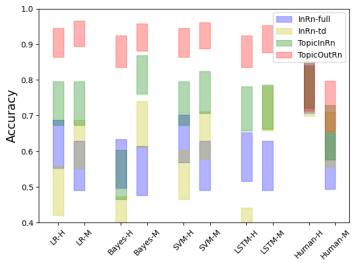


Figure 5: Accuracy of automatic and manual detection tools evaluated on machine- and human-generated news.

Other variables. Besides the dependency of the detection on the main categories, we evaluated the potential dependency between the labels predicted and other potential explanatory variables (i.e., the types of topics, the week, the titles and the groups). None of these variables was found to have a statistically significant impact on the predicted labels with the amount of samples collected in our experiment.

5 QUALITATIVE ANALYSIS

In this last section, we finally analyze justifications that participants gave during the manual detection experiment and give some illustrative cases where automatic detection and manual detection lead to different outcomes.

5.1 Justification types and subtypes

We categorized the reasons given by the participants into three main types: *Syntax*, *Semantics* and *Background*, ranging from more local & internal to more global & external.

For the *Syntax* type, we considered 4 sub-types: *Punctuation & Connectors* (PC) for the kind and amount of punctuation & connectors; *Words Repetitions* (WR) for the repetitions of certain words; *Patterns' Repetitions* (PR) for the repetition of syntactic patterns (e.g., sentence constructions); *Grammar & Typos* (GT) for spelling or grammatical mistakes.

For the *Semantics* type, we considered 4 sub-types: *Sentence Meaning* (SM) for meaningless sentences; *Transitions* (T) for logical transitions between sentences; *Goal & Structure* (GS) for the definition of a main point in the news items and a text structure that is coherent with respect to that main point; *Internal Coherence* (IC) for inconsistencies between pieces of information appearing in the item.

For the *Background* type, we considered 4 sub-types: *External Coherence* (EC) for inconsistencies with respect to the reader's knowledge; *Writing Style* (WS) for the subjective evaluation of the writing skills; *Vocabulary & Expressions* (VE) for the use of particular vocabulary words or typical expressions; *References* (R) for the presence and validity of links, citations, ..., referred to within the news item.[6]

To these three main types, we add an *Others* type, either for reasons that relate to specific features of our models or when several of the previous types are present in the justification It contains 4 sub-types as well: *Numbers* (N) for the presence of numbers that trigger either an internal (e.g., 3+2 = 6) or external (e.g., 107% of the people) inconsistency; *Subjectivity & Opinion* (SO) for the expression of a viewpoint or opinion in the news; *Believable* (B) for news that in general do not contradict the reader's prior knowledge, whether related to the *Syntax*, *Semantic* or *Background* types; *Comparisons* (C) for justifications made by comparing different news.

Note that these types and sub-types of justifications can be used positively (resp., negatively) to argue in favor of (resp., against) the fact that a news item was machine-generated.

Based on this taxonomy, we represent the distribution of the reasons' types given by the participants during the experiment in Figure 6. While, as in the previous section, we lack samples to turn these histograms into quantitative conclusions, we can nevertheless extract two useful observations. First, this figure shows that the types of reasons used by the participants are not the same for machine-generated and human-generated news. This complements the indication given by Figure 5 that the accuracy of manual detection differs for machine-generated and human-generated news. Second, it shows that the detection of machine-generated news

[6] WS and VE sub-types are background as they relate to the participants' a priori expectation of how humans and computers generate news.

relies more on local types of reasons (i.e., *Syntax* and *Semantics*) while the detection of human-generated news relies more on global ones (i.e., *Background* and *Others*).

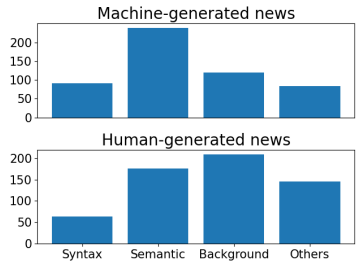


Figure 6: Histogram of the participants' justification types for machine-generated and human-generated news.

The more detailed Figure 7 leads to similar observations while also questioning the possibility to improve automatic detection tools with more advanced machine learning models and the possibility to improve manual detections by combining them with an automated detection assistant.

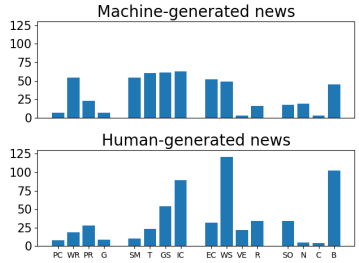


Figure 7: Histogram of the participants' justification sub-types for machine-generated and human-generated news.

Regarding the first possibility, it is worth noticing that among the reasons currently listed by the participants in our manual detection experiment, only a few are exploitable by the simple models we used as automatic detectors. For example, our pre-processing step removes the punctuation, the connectors, all the non fully alphabetical words (e.g., URLs, numbers, ...) and the inflection forms (cancelling potential grammar mistakes). Furthermore, our three first models (i.e., LR, SVM and NB) rely on a TF-IDF vectorizer, implying that they leverage words' frequencies that are only reflected in some justification sub-types (e.g., *Comparisons*, *Words Repetitions*, *Vocabulary & Expressions*, and to some extent *Writing Style* and *Subjectivity & Opinion* if related to the words frequencies). By being able to capture sequential text features, the LSTM can in theory exploit more justifications sub-types like *Pattern Repetitions*, *Sentence Meaning*, *Transitions* and to some extent *Goal & Structure* and *Internal Coherence*, but would require more training data for this purpose.

More positively, the simplicity of our automatic detection tools emphasizes their potential complementarity with manual detection. That is, none of the participants in our experiment mentioned the analysis of words' frequencies (or other statistical measures that can be extracted by machine learning algorithms) in their justifications. This suggests that (even simple) automatic detectors could be an interesting asset to help manual detection (e.g., by generating warning signals for news that would be worth special attention).

5.2 Manual vs. automatic detection

As an illustration of the potential complementarity between the manual and automatic detection of machine-generated vs. human-generated news, we finally discuss a few examples where these two types of detection lead to contradictory outcomes. In order to identify them, we first listed these cases exhaustively (e.g., in Table 1 and Table 2 where we can see that a higher detection accuracy leads to a higher similarity between manual and automatic detection).[7]

Models	Output =		Output ≠		
	Hg/Hg	Mg/Mg	Hg/Mg	Mg/Hg	Total
LR	28	21	11	40	100
Bayes	30	20	12	38	100
SVM	30	18	14	38	100
LSTM	42	14	18	26	100

Table 1: Agreement between manual and automatic detections for the "In `RealNews`-Full" Category. Hg/Hg: both predict human-generated, Mg/Mg: both predict machine-generated, Hg/Mg: automatic detection predicts human-generated and manual detection predicts machine-generated, Mg/Hg: automatic detection predicts machine-generated and manual detection predicts human-generated.

Models	Output =		Output ≠		
	Hg/Hg	Mg/Mg	Hg/Mg	Mg/Hg	Total
LR	46	35	10	9	100
Bayes	46	32	13	9	100
SVM	25	34	11	30	100
LSTM	38	36	9	17	100

Table 2: Agreement between manual and automatic detections for the "Topic out of `RealNews`" Category.

[7] Since every news was reviewed by 3 participants (one per group), the manual detection decision in the table is taken as the majority vote.

A first example is given in Appendix B, subsection B.1. This machine-generated news item is correctly classified by all the models but incorrectly classified by all the participants of our experiment. The main reason given for this (incorrect) decision is that the writer's viewpoint was given in some sentences (highlighted in red), which participants considered as typically human. A second example is given in Appendix B, subsection B.2. This machine-generated news is also incorrectly classified by all the participants. Here the reason was that the names of the authors (highlighted in red) appear at the beginning and at the end of the items, which participants considered as a proof of consistency. Interestingly, both examples suggest that the a-priori understanding of the participants about what language models can achieve (or lack thereof) can play a role in their decisions, although this was not detected quantitatively in our experiments. Hence, repeating such experiments and considering clearly defined groups of participants with a limited understanding of language models and groups that are trained to use them, with more samples, would be an interesting follow-up work.

Finally, a third example is given in Appendix B, subsection B.3. This machine-generated news is correctly identified as such by all the participants but incorrectly classified by all the models. This time, participants mostly underlined a lack of goal and structure in the news, giving the impression that the sentences are weirdly connected. As mentioned in subsection 5.1, such semantic reasons are among those that are unlikely to be captured by our simple models, leaving the investigation of more complex models with more training data as another interesting follow-up question.

6 CONCLUSIONS

In this paper, we investigated different challenges raised in the news field by the recent improvements of automatic natural language generation based on language models.

First of all, in relation to the quality of machine-generated news and their likelihood to be perceived as human-generated, our experiments confirm the difficulty for language models to deal with domain shifting. Generating a seemingly consistent text on a topic outside those processed during training remains inherently challenging and requires further investigation in domain adaptation techniques like plug-and-play language models [**?**]. The implications of this observation in the news field are two-fold. On the one hand, the amount of time required for training or domain adaptation makes it less easy to spread harmful disinformation on recent event-triggered topics. On the other hand, for professional newsrooms, it also makes the perspective of using these techniques in a constructive manner quite faraway at this stage. Automatic content production has a great potential in journalism, whether it be for "speeding up production, increasing breadth of coverage, enhancing accuracy or enabling new types of personalization" [**?**], but without further advances, it will likely remain based on a combination of templates and structured data rather than on language models.

Second, in relation to the detection of machine-generated news, our results indicate a complementarity between the automatic and manual detection methods. Humans and machines do not focus on the same aspects, and our results confirm that computing power should be used for what it does best: scale and speed. Counting words in a huge amount of texts and detecting hidden patterns in the way sentences are built is definitely something computers do better and faster than humans. Therefore, enhancing manual detection based on semantic or background knowledge elements with statistical syntactic indications that are automatically computed appears to be the most efficient way to detect such generated news. In newsrooms, besides finding and spreading news items, debunking and verifying information has always been part of the normative role journalists assign themselves [**?**]. With the advent of technologies enabling massive amount of credible machine-generated text, this role needs to evolve and take a computational turn. New skills need to enter the newsroom and this is yet another example of how journalism practices are not replaced but displaced by the recent improvements of automation in the production workflow. Whether machines will eventually be able to exploit semantic and background elements just as humans naturally do remains an open question. However, even a positive answer will not empty the role of journalists of what is the very essence of the profession: being accountable for the information disseminated and the editorial decisions that are taken. This hybridation of professional practices with accountability remaining on the human side is in line with the "human still in the loop" perspective recently put forward by Milosavljević and Vobič [**?**]. It emphasizes a pressing need to embed innovations into established human–machine relations in the news production process, which is amplified by the technical challenges of making algorithms accountable [**?**]. It also follows Bucher's general observation that algorithms "do not eliminate the need for human judgment and know-how in news work; they displace, redistribute, and shape new ways of being a news worker" [**?**].

Acknowledgments. François-Xavier Standaert is a senior associate researcher of the Belgian Fund for Scientific Research (FNRS-F.R.S.). This work has been funded in parts by the Walloon Region's ARIAC/TRAIL research project.

REFERENCES

[] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022.
[] Taina Bucher. 2018. *If...Then: Algorithmic Power and Politics.* Oxford University Press.
[] Fabrice Colas and Pavel Brazdil. 2006. Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. In *Artificial Intelligence in Theory and Practice, IFIP 19th World Computer Congress, TC 12: IFIP AI 2006 Stream, August 21-24, 2006, Santiago, Chile (IFIP, Vol. 217)*, Max Bramer (Ed.). Springer, 169–178. https://doi.org/10.1007/978-0-387-34747-9_18
[] Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen. 2015. Automatic Deception Detection: Methods for Finding Fake News. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.
[] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Mach. Learn.* 20, 3 (1995), 273–297. https://doi.org/10.1007/BF00994018
[] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *ICLR*. OpenReview.net.
[] Antonin Descampe, Clément Massart, Simon Poelman, François-Xavier Standaert, and Olivier Standaert. 2022. Automated news recommendation in front of adversarial examples and the technical limits of transparency in algorithmic accountability. *AI Soc.* 37, 1 (2022), 67–80.
[] Nicholas Diakopoulos. 2019. *Automating the News: How Algorithms Are Rewriting the Media* (harvard university press ed.).
[] Benoît Frénay and Michel Verleysen. 2014. Classification in the Presence of Label Noise: A Survey. *IEEE Trans. Neural Networks Learn. Syst.* 25, 5 (2014), 845–869.
[] David J Hand and Keming Yu. 2001. Idiot's Bayes—not so stupid after all? *International statistical review* 69, 3 (2001), 385–398.
[] Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by

ClaimBuster. In *KDD*. ACM, 1803–1812.
[] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
[] David W. Hosmer and Stanley Lemeshow. 2000. *Applied Logistic Regression, Second Edition*. Wiley. https://doi.org/10.1002/0471722146
[] Edson C. Tandoc Jr., Zheng Wei Lim, and Richard Ling. 2018. Defining "Fake News". *Digital Journalism* 6, 2 (2018), 137–153.
[] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The Science of Fake News. *Science* 359, 6380 (2018), 1094–1096.
[] Gang Liu and Jiabao Guo. 2019. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing* 337 (2019), 325–338. https://doi.org/10.1016/j.neucom.2019.01.078
[] Marko Milosavljević and Igor Vobič. 2019. Human Still in the Loop. *Digital Journalism* 7, 8 (Sept. 2019), 1098–1116. https://doi.org/10.1080/21670811.2019.1601576
[] Tomas Pranckevicius and Virginijus Marcinkevicius. 2017. Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Balt. J. Mod. Comput.* 5, 2 (2017). https://doi.org/10.22364/bjmc.2017.5.2.05
[] Juan Ramos. 2003. Using TF-IDF to Determine Word Relevance in Document Queries. In *Proceedings of the First Instructional Conference on Machine Learning*. 29–48.
[] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *CIKM*. ACM, 797–806.
[] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating Fake News: A Survey on Identification and Mitigation Techniques. *ACM Trans. Intell. Syst. Technol.* 10, 3 (2019), 21:1–21:42.
[] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor.* 19, 1 (2017), 22–36.
[] Stephen J. A. Ward. 2018. Epistemologies of Journalism. In *Journalism*, Tim P. Vos (Ed.). De Gruyter, 63–82. https://doi.org/10.1515/9781501500084-004
[] Wen Xu, Jing He, and Yanfeng Shu. 2020. Transfer Learning and Deep Domain Adaptation. In *Advances and Applications in Deep Learning*, Marco Antonio Aceves-Fernandez (Ed.). IntechOpen, Chapter 3.
[] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *NeurIPS*. 9051–9062.
[] Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A C-LSTM Neural Network for Text Classification. *CoRR* abs/1511.08630 (2015). arXiv:1511.08630 http://arxiv.org/abs/1511.08630
[] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* 53, 5 (2020), 109:1–109:40.

A ADDITIONAL FIGURES

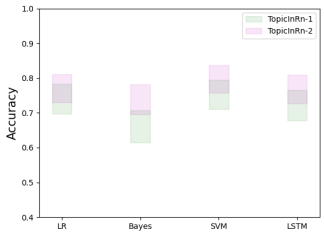


Figure 8: Accuracy of automatic and manual detection tools for timeless and event-triggered types of topics.

B EXEMPLARY NEWS

B.1 First example

Candid Obama admits: healthcare website glitches are a gift to critics

This story was updated at 2:12 p.m.

WASHINGTON – Barack Obama is expected to invoke [...]

If Obama starts off speaking about "infrastructure problems," he may as well be announcing that a healthcare act was the foundation of the federal government's economic success. Because of the website's weakness and the software shutdown, millions of people have been unable to use the healthcare exchanges or their federal employer plans for weeks and months.

As we move deeper into the election, Obama is [...]

That would be a good idea.

B.2 Second example

Beirut explosions create a dilemma for the world

By Marja Sokneski and Olga Rosendahl

MUZAFFARABAD, August 9 (Reuters) – As the Iranian nuclear talks in Rome wind down, Iranian leaders are under increasing pressure to commit more to resolving the Middle East's biggest nuclear crisis. [...]

"Iran was in Syria mainly to ease the crisis. What we need is a durable settlement that all sides can live with." He did not give more details.

(Reporting by Marja Sokneski and Olga Rosendahl; Editing by Catherine Evans)

B.3 Third example

16 feared dead after hot air balloon crashes in Texas

The International Space Station air ambulance has landed at the base of the Texas Air National Guard base, where the disaster happened Tuesday afternoon (July 30).

National Guardsmen [...]

A 14-foot air ambulance was taking part in a training exercise at the base that day that saw a hot air balloon shoot a hole in the grass before crashing into the base's hilly terrain.

Crews on site watched the shooter [...]

a 100-ton oxygen tank stuck above the ground, said Lt. Scott Anderson, spokesman for the station.

"I don't think we lost a single lifter or crew member," Anderson said. [...]

With the air ambulance's parachute, a safe launch position, and the power of the air rescue truck, the rocket typically takes about 10 minutes to complete.

KXAN requested an interview from the air force, as well as a telephone interview with the agency.

The RTBF Corpus : a dataset of 750,000 Belgian French news articles published between 2008 and 2021

Louis Escouflaire [1,2], Jérémie Bogaert [3], Antonin Descampe [1] et Cédrick Fairon [2]
[1] CENTAL - UCLouvain, [2] ORM - UCLouvain, [3] ICTEAM - UCLouvain
(louis.escouflaire/jeremie.bogaert/antonin.descampe/cedrick.fairon)@uclouvain.be
International Conference on Corpus Linguistics 2023 (JLC23)

Introduction

News corpora provide a large and diverse representation of written language that is essential for various fields of linguistics, such as lexicology, discourse analysis, sociolinguistics, and for training language models in NLP (Tognini-Bonelli, 2021). Their diachronic aspect allows for an investigation of language evolution over time and of its back and forth influence on society, making them valuable resources for linguistic research and teaching (Hilpert & Gries, 2016). Likewise, in media and journalism studies, news corpora are important both for qualitative and quantitative research (Conboy, 2007).

In this paper, we introduce the RTBF Corpus, a large diachronic corpus of 767,204 Belgian French news articles published between 2008 and 2021 by the Belgian public service media RTBF. We present the contents and structure of the corpus, along with the different layers of metadata available for each text. We also describe the three different versions of the articles available in the corpus (depending on the cleaning and preprocessing steps applied to the text). The RTBF corpus is freely available online in CSV format [1], for research and teaching purposes only.

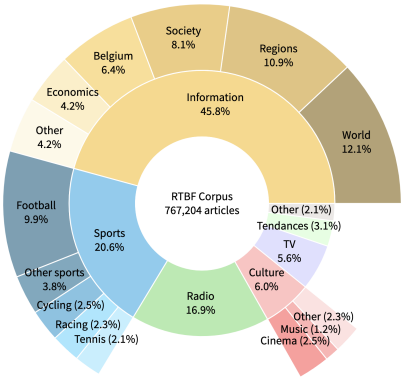


FIGURE 1 – Article distribution per feed (in percentages).

The RTBF Corpus

There exist few open-source French news corpora of considerable size and interest. The *Est Républicain* Corpus, released in 2009, is to our knowledge the largest freely available French news corpus with around 149 million words (Gaiffe & Nehbi, 2009). The articles it contains were published between 1999 and 2003 and between 2006 and 2011, and cover mostly local news of the Eastern part of France (Seddah et al., 2012), which may restrict the variety of topics mentioned in the corpus.

The RTBF Corpus is a freely available Belgian French news corpus, like the *Est Républicain* corpus. However, the RTBF Corpus contains more than 214 million words, which makes it 1.4 times larger. RTBF is a national media which covers all kinds of topics, among which international and Belgian

1. https ://dataverse.uclouvain.be/dataset.xhtml ?persistentId=doi :10.14428/DVN/PEVSSI

news, sports and culture. The corpus is chronologically continous (unlike the *Est Republicain* corpus) : the articles were published between 2008 and 2021, allowing for longitudinal investigations over the whole 2010 decade and more, and for analyses on news coverage of recent events (Escouflaire et al., 2022 ; Escouflaire et al., 2023). The corpus can also be a useful resource for experiments involving machine learning or requiring large amounts of data, such as news genre classification. Such large-scale resources are limited for the French language, making this corpus a useful asset.

The news director of RTBF officially handed us the rights over this dynamic corpus, which will be updated regularly with new articles published on the RTBF website. The corpus will be available online for free downloading, if the user agrees to follow the terms of use which are attached to the data. In short, the corpus may exclusively be used for research or teaching purposes (no commercial usage is permitted), it must be exploited in an ethical manner, and its user must inform RTBF of any planned publications related to the results and conclusions obtained from the corpus, and the media has the ability to request relevant comments and observations made by them to be included.

Corpus description

RTBF (*Radio-télévision belge de la Communauté Française*) is the public service broadcasting organization for the French-speaking community of Belgium. As a public service media, it is directly funded by the Belgian government and has three main missions : inform, educate, and entertain a public as large as possible in the French-speaking Belgian community. Besides operating television channels and radio stations, RTBF also operates a news website since 2008, on which web-only press articles are published daily. Through scientific collaboration with RTBF, we received access to the entirety of the web articles published on their website from 2008 to 2021. As shown in Figure 2, the publication rate has changed over the years, likely due to editorial movements inside the media. As of January 2023, between 1,000 and 1,500 articles are uploaded every week on their website.

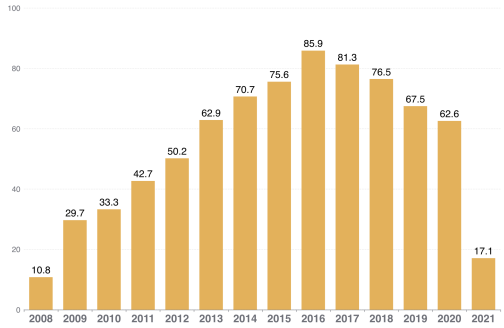


FIGURE 2 – Article distribution per year (in thousands of articles).

The corpus contains 767,204 articles, for a total of 214,072,620 words. The mean amount of words per article is 279. The full text of each article is in the *text* column of the corpus, along with seven columns consisting of different pieces of metadata attached to the article : *ID*, *title*, *publication date*, *signature*, *feed*, s*category* and *keyword*. We present the contents of all columns in more detail :

- ***ID*** : The article's internal ID, which is a number containing between 3 and 8 digits generated by the media source. This ID can be exploited to easily find its original page and layout on the RTBF website. A straight Google search with the format "*RTBF [ID]*" will return the article's link as the first result.
- ***title*** : The article's title, with an average length of 10 words. The lead paragraph and subtitles of the article are part of the main *text* column.

- ***publication date*** : The day on which the first version of the article was published on the RTBF website. The content of some articles may have been updated after their publication, but this information is not present in the corpus.

- ***signature*** : The name of the journalist who wrote the article, the radio or TV program it was originally derived from (cfr. *feed) further*, or the original source which the information is attributed to. About 43% of all articles in the corpus are signed solely or in collaboration with *Belga* or *AFP* (Agence France-Presse), two renowned press agencies delivering straight news. A great number of press agency dispatches are published daily on the RTBF website, with or without enrichments made by RTBF journalists. 1% of all articles were signed by a person or an entity who signed only one article in the corpus.

- ***feed*** : The structural and topical section of the RTBF website in which the article was published. The entire corpus is divided into 7 feeds : *Information*, *Sports*, *Radio*, *Culture*, *TV*, *Tendances* and *Other*. Their percentage distribution in represented in the inner circle of Figure 1. The *Information* feed, which is also the default feed on the RTBF website, contains mostly news belonging to the *information* genre of journalism (Grosse, 2001), i.e. content that is usually considered objective and factual by the source and by readers. Most press agency dispatches are found in this feed. However, the *Information* feed also contains 5,000 articles belonging to the *opinion* genre, i.e. op-eds or columns (those articles are marked with the *chronicles* or *opinions* category tag). The *Sports* and *Culture* feeds also include some amount of opinionated articles (*chronicles)*, representing respectively 416 (0.2%) and 2,800 (6%) of their articles. Then, the *Radio* and *TV* feeds contain articles which were derived from programs broadcasted on RTBF radio or television channels. *Tendances* contains lifestyle articles (e.g. fashion, food, technology, health, travel). The *Other* feed groups articles that were not classified into the 6 other feeds, among which many press releases and weather reports.

- ***category*** : Further classification related to the article's topics. Categories can help to group articles into more specific themes, as shown on the outer circle of Figure 1. Categories with small amounts of articles (e.g. *Basketball*, *Gaming*) were not represented on Figure 1, but were instead grouped into the *Other* category of the feed (*e.g. Other sports*). For the *Radio* and *TV* feeds, categories (not shown on Figure 1) are related to the specific TV channel or radio station from which the broadcast-related articles were derived (e.g. "Matin Première").

- ***keyword*** : Word or phrase attributed to some articles by the journalist or source who wrote them to further annotate their content (e.g. *Europe*, *environment*, *Internet*). About 28,5% articles in the corpus were assigned a keyword.

Corpus cleaning and preprocessing

Three different versions of the article's full text are available in the corpus. Each version contains the text at different stages of data cleaning and preprocessing :

- *HTML* text : The raw version of the text directly received from the RTBF, containing HTML tags. This version of the text can be used for research dealing with metatextual information (e.g. words highlighted in bold or italic, headings).

- *cleaned* text : The version of the text after multiple steps of data cleaning. To get this version, we removed HTML tags and fixed bugs related to text encoding of special characters. Some metatextual elements are still included in this version, such as signatures added at the end of some texts and links referring to other articles of the website.

- *preprocessed* text : The version obtained after multiple preprocessing steps. Most signatures at the end of articles (e.g. *with Belga*) and links to other articles (e.g. *Read also : ...*) were systematically removed from the articles. Numbers and URLs were replaced by placeholder tokens, <NUM> and <URL>. This version should be used carefully, as some elements that

may be relevant for some experiments could be missing from the text. Those preprocessing steps were initially derived and designed for the creation of the InfOpinion corpus (*Bogaert* et al., *2023*), a subcorpus of the RTBF corpus fit for a text classification task.

Notes

The RTBF Corpus was initially created in the context of a PhD program carried out at UCLouvain, in partnership with RTBF. The research is funded by FRS-FNRS (Belgian National Fund for Scientific Research) and conducted by PhD student Louis Escouflaire at ILC (Institute for Language and Communication) under the guidance of Antonin Descampe (Observatory for Research on Media and Journalism) and Cédrick Fairon (Center for Natural Language Processing).

References

Bogaert, J., Escouflaire, L., de Marneffe, M.-C., Descampe, A., Fairon, C. & Standaert, F.-X. (2023). TIPECS : A corpus cleaning method using machine learning and qualitative analysis. *International Conference on Corpus Linguistics 2023 (JLC23).*

Conboy, M. (2007). *The language of the news.* London : Routledge.

Escouflaire, L., Descampe, A., Fairon, C. (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 2021. *JADT 2022 : 16th International Conference on Statistical Analysis of Textual Data.* Naples, Italy.

Escouflaire, L., Descampe, A., Lits, G., Fairon, C. (2023). Analyzing the semantic evolution of bias in French news articles using word embeddings. *Digital Humanities Benelux 2023.* Brussels, Belgium.

Gaiffe, B. & Nehbi, K. (2009). Le corpus de l'Est Républicain. *Technical report, Atilf.* http ://www.cnrtl.fr/corpus/estrepublicain/.

Hilpert, M., & Gries, S. T. (2016). Quantitative approaches to diachronic corpus linguistics. The Cambridge handbook of English historical linguistics, 36-53.

Küppers, A., & Ho-Dac, L. M. (2011). Un corpus de presse francophone pour l'étude de l'impact d'Internet sur les pratiques langagières. *CJC Praxiling : Corpus, données, modèles : approches qualitatives et quantitatives.*

Seddah, D., Candito, M., Crabbé, B., & Anguiano, E. H. (2012). Ubiquitous usage of a broad coverage French corpus : Processing the Est Républicain corpus. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3249-3254.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at work.* Amsterdam/Philadelphia : John Benjamins.

TIPECS : A corpus cleaning method using machine learning and qualitative analysis

Louis Escouflaire [1,2], Jérémie Bogaert [3], Marie-Catherine de Marneffe [1], Antonin Descampe [2], François-Xavier Standaert [3] et Cédrick Fairon [1]
[1] CENTAL - UCLouvain, [2] ORM - UCLouvain, [3] ICTEAM - UCLouvain
(louis.escouflaire/jeremie.bogaert/antonin.descampe/cedrick.fairon)@uclouvain.be
International Conference on Corpus Linguistics 2023 (JLC23)

1. Introduction

We present TIPECS (**T**rain, **I**nfer **P**redictions, **E**xplain, **C**lean and **S**tart again), a corpus cleaning method relying on a mixed approach between machine learning and manual analysis. The aim of our dataset cleaning approach is to remove tokens or segments that are considered as discriminant features by a classification model trained on a given dataset for a given task, but that cannot be generalized to other similar tasks or datasets. Such elements are referred to as *artifacts* (Gururangan et al. [2018]). For example, when classifying news articles as opinion or information, some press agencies signatures, such as Belga, will be clear indicators for the information class, but might not appear in other datasets. To make the dataset as valuable as possible for the task overall, these artifacts should be removed. We showcase TIPECS on a French information vs. opinion news classification task. The code for our method and the dataset created will both be available on GitHub. [1]

2. The TIPECS method

TIPECS is a semi-automatic method to iteratively remove, from a given dataset, tokens that can be considered as artifacts for a given task. Figure 1 illustrates the process. The intuition behind the method is to use a powerful model overfitting on artifacts and biases to highlight them and remove them from the dataset. The method is applicable to all kinds of textual datasets and using various types of machine learning models. The different steps of the TIPECS method are the following :

- **T**rain : Feed the training data to the model to teach it to correctly classify the items.
- **I**nfer **P**redictions : Use the trained model to predict a class for some unseen items.
- **E**xplain : Use a token-level model explanation method to find the top-n tokens that influence the most the model's predictions (n can be determined by the user, default value is 20).
- **C**lean : If artifacts are found through qualitative analysis of the n-most discriminant tokens, add preprocessing steps to remove them from the dataset.
- **S**tart again : Repeat the process until the n-most discriminant tokens are only relevant items.

Some constraints must be respected to obtain valuable results with TIPECS. First, the machine learning model used has to be able to identify discriminant features for the classification task. Transformer models are therefore candidates of choice for our method, given their tendency to overfit on various biases and artifacts (Gururangan et al. [2018]). Second, the model has to be able to explain its predictions at token level, for example, with transformer models, using perturbation methods or attention probing (Bibal et al. [2022]). Finally, since the training, the prediction

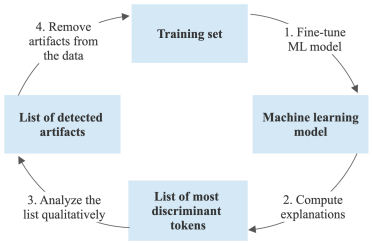


FIGURE 1 – TIPECS framework to remove artifacts from a dataset.

1. Link to GitHub (anonymized).

inference and the explanation steps have to be carried out at each cycle, these three steps should not be too computationally intensive.

3. Case study

3.1. The InfOpinion corpus

The InfOpinion corpus comes from the RTBF corpus, a collection of 750,000 news articles published by the Belgian French-speaking public service media RTBF (Radio-télévision belge de la Communauté Française) on their website between 2008 and 2021. This corpus was built for training and evaluating a classification model distinguishing between pieces from the journalistic *opinion* genre (such as editorials, commentaries, reviews), which are considered to be subjective, and pieces belonging to the *information* genre (press agency dispatches, news articles), meant to be more objective texts (Grosse [2001]). This binary categorization relies solely on the articles' annotation by the RTBF as either *opinion* or *information*.

The InfOpinion corpus consists of 10,000 articles published between 2012 and 2021 on the INFORMATION feed of the RTBF Corpus. It is a balanced dataset containing on the one hand 5,000 articles classified by the RTBF as *Op-eds* ("Chroniques") or *Opinions*, and on the other hand 5,000 articles randomly selected from the *Belgium*, *World*, and *Society* categories, which tackle similar topics to those discussed in the *Op-eds* category. Each article in the InfOpinion corpus has multiple layers of metadata, which are described further in (*Anonymized* et al., [*in press*]) : ID, title, publication date, signature, feed, category and keyword. An additional column is devoted to the article's genre (or class), which is either *information* or *opinion*.

3.2. Applying TIPECS

We showcase TIPECS by using it to preprocess the InfOpinion corpus with *information* vs. *opinion* classification in mind. We use CamemBERT, a French adaptation of the RoBERTa transformer model (Liu et al. [2019], Martin et al. [2020]) and an explanation method based on a variation of layer-wise propagation (LRP). The idea behind LRP is to explain a deep neural network prediction by assigning a relevance score to each input feature (i.e. token) by starting from the prediction and back-propagating to the input using conservation constraints (Bach et al. [2015]). We use Chefer et al. [2021]'s variation of the LRP method, which we modified to be able to work with CamemBERT's architecture. All other parameters remained unchanged from the original implementation. These settings allow us to obtain good and fast results with a single GPU (about 40 minutes per TIPECS cycle).

The InfOpinion corpus is first divided into a training set (80%), a development set (10%) and a test set (10%), all balanced between the *information* and *opinion* classes. We **train** the base version of CamemBERT on the training set during 2 epochs to classify *opinion* and *information* articles. We then **infer predictions** with the trained CamemBERT model on the development set and **explain** those predictions at token level with LRP by computing attention maps for the 1,000 articles in the development set. To avoid analyzing too specific tokens, we only take into account tokens with at least 10 occurrences in the dataset. For each token, words and punctuation signs included, we compute its mean relevance value across all the articles in which it appears. All tokens are then ranked by average relevance value. Table 1 shows the 20 most discriminant tokens obtained after the first iteration for both the *opinion* and *information* classes.

The next step consists in manually examining the two lists of the most discriminant tokens for the *opinion* and *information* classes, to find which elements are not relevant to the classification task. Then, we add preprocessing rules to **clean** the dataset by removing the unwanted tokens. Finally, after each TIPECS cycle, we **start again** : the CamemBERT model is fine-tuned on the newly cleaned version of the InfOpinion dataset, and all the steps in the TIPECS iterative process are repeated until no artifacts remain in the two lists of most discriminant tokens, i.e. until all tokens are judged potentially relevant and generalizable for explaining predictions in our task. In this case study, we stopped after 5 iterations.

Opinion				Information			
cinéaste *filmmaker*	provocation	film	foi *faith*	AFP</s>	"</s>	visant *aiming*	samedi *Saturday*
sinon *otherwise*	bref *in short*	actualité *news*	pari *bet*	Belga</s>	jeudi *Thursday*	indiqué *indicated*	Belga
latin	[	ah	imaginaire *imaginary*	précise *specifies*	expliqué *explained*	mardi *Tuesday*	bière *beer*
média *media*	puisqu'il *because he*	prenons *[we] take*	raconter *to tell*	pompiers *firefighters*	précisant *specifying*	mercredi *Wednesday*	concrètement *concretely*
#	cinéma *cinema*	incapable *unable*	décidément *decidedly*	rapporte *reports*	.</s>	précisé *specified*	ajouté *added*

TABLE 1 – Top-20 most discriminant tokens (based on LRP explanations) for the *opinion* and *information* classes after the first TIPECS iteration (from top to bottom, left to right).

In the first iteration, we identified multiple potential artifacts in the dataset related to the texts' structures, such as press agency or author signatures (e.g. *Belga*), links and references to similar articles (*Read also...*), and structural tokens (#). We therefore added a preprocessing step to keep only the text and remove the metadata surrounding it. We also replaced URLs with a unique tag (<URL>) to avoid model overfitting on the various hyperlinks present in the data. Similarly, we replaced all numbers and digits with a tag (<NUM>), to prevent the model from using meaningless numerical values during its predictions while still allowing it to highlight their presence or frequency.

Throughout the iterations of the process, we also observed that some topics were significantly more frequent in *opinion* articles than in *information* articles, in particular tokens associated with cultural topics (e.g. *cinema, film*). To address this, we modified the information dataset by replacing 10% of the articles with articles from the CULTURE feed of the RTBF corpus.

We found that text length was also playing a major role in the model's predictions. *Information* articles are on average shorter than *opinion* pieces. Since the inherent input limit for the CamemBERT model is of 512 tokens, the end-of-sequence character (</s>) appeared more often after a final punctuation sign (.,! ,? or ...) in *information* articles, as seen in Table 1. On the contrary, it appeared more often in the middle of sentences for *opinion* articles (because the input texts were too long). To avoid this bias, we implemented a dynamic truncation preprocessing step allowing to cut long articles after the last complete sentence (i.e. ending with .,?,! or ...) ending before the token limit. This way, the model can no longer discriminate articles based on the end-of-sequence character.

Table 2 shows the final lists of most discriminant tokens we obtained on the development set after 5 TIPECS iterations. Note that since both classes do not exactly treat the same topics, words such as *hospital* or *iPhone* appear in the lists. They are part of semantic fields inherently more prevalent in their respective class (*information* and *opinion*), and thus cannot be considered artifacts in this task, as they could generalize to other datasets.

This case study shows how TIPECS can be used to identify artifacts in a classification task dataset to prune it from unwanted bias during subsequent experiments.

Opinion				Information			
verbe *verb*	latin	église *church*	imaginons *[we] imagine*	affiche *poster/displays*	indique *indicates*	samedi *Saturday*	métrage *footage*
puisqu'il *because he*	sondage *poll*	iPhone	humour *humor*	mardi *Tuesday*	jeudi *Thursday*	vendredi *Friday*	ajouté *added*
patrons *managers*	paraît *seems*	budgétaires *budgetary*	médiatique *media-related*	hôpital *hospital*	mercredi *Wednesday*	passagers *passengers*	précisé *specified*
conservateurs *conservatives*	articles	mots *words*	découvre *discovers*	indiqué *indicated*	située *located*	lundi *Monday*	ONG *NGO*
syndical *union*	expression	intéresse *interests*	déficit *deficit*	exposition *exhibition*	aérienne *aerial*	AFP	précise *specifies*

TABLE 2 – Top-20 most discriminant tokens (based on LRP explanations) for the *opinion* and *information* classes after 5 iterations of the TIPECS method (from top to bottom, left to right).

Acknowledgments

This project was carried out in the context of two PhD programs carried out at UCLouvain, funded by FNRS (Belgian National Fund for Scientific Research) and by TRAIL (Trusted AI Labs).

References

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics, 2018. doi : 10.18653/v1/n18-2017.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. Is Attention Explanation ? An Introduction to the Debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 3889–3900, Dublin, Ireland, 2022. Association for Computational Linguistics. doi : 10.18653/v1/2022.acl-long.269.

Ernst-Ulrich Grosse. Evolution et typologie des genres journalistiques. Essai d'une vue d'ensemble. *Semen. Revue de sémio-linguistique des textes et discours*, (13), 2001.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa : A robustly optimized BERT pretraining approach. *arXiv preprint arXiv :1907.11692*, 2019.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, 2020. doi : 10.18653/v1/2020.acl-main.645. arXiv :1911.03894 [cs].

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7) : e0130140, 2015.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021.

Fine-Tuning is not (Always) Overfitting Artifacts

Jeremie Bogaert[1], Emmanuel Jean[2],
Cyril de Bodt[1] & François-Xavier Standaert[1] *

1- ICTEAM/UCLouvain - Belgium. 2- Multitel - Belgium

September 9, 2025

Abstract

Since their release, transformers, and in particular fine-tuned transformers are widely used for text related classification tasks. However, only a few studies try to understand how fine-tuning actually works and existing alternatives, such as feature-based transformers, are often overlooked. In this work, we study a French transformer model, CamemBERT, to compare the fine-tuned and feature-based approaches in terms of their performances, interpretability and embedding space. We observe that while fine-tuning has a limited impact on performances in our case study, it significantly affects the intepretability (by better isolating words that are intuitively connected to the classification task) and embedding space (by summarizing the majority of the relevant information into a fewer dimensions) of the results. We conclude by highlighting open questions regarding the generalization potential of fine-tuned embeddings.

1 Introduction

Since their release a few years ago, transformers [11] have been of interest for the whole Natural Language Processing (NLP) community. Models such as BERT [6] and its variants (RoBERTa [12], AlBERT [13], ...) achieve state-of-the-art results on numerous classification tasks and datasets. These models are typically first pre-trained on big amounts of data in order to learn how to transform texts into embeddings, and then adapted to an end task. To do so, [6] presented two different approaches: the fine-tuning approach, *where a simple classification layer is added to the pre-trained model and all parameters are jointly fine-tuned on a down stream task*, and the feature-based approach, often referred to as the frozen one, *where fixed features are extracted from the pre-trained model and fed to another model*. While the former is usually preferred, and may marginally improve the model performances [8], both models

* Work supported by the Service Public de Wallonie Recherche, grant n°2010235-ARIAC by DIGITALWALLONIA4.AI. FXS is Senior Research Associate of the F.R.S.-FNRS.

reach similar results on many classification tasks [5, 9]. However, as many other complex models, neural networks are prone to learn statistical quirks in data, which can results in adopting shallow heuristics that succeed for most training examples, instead of learning underlying generalizations that they are intended to capture [1, 4], which we denote as overfitting artifacts. It raises the question of the impact of fine-tuning in this respect, which this paper tackles by assessing different properties of fine-tuned vs frozen models based on two French text classification tasks using CamemBERT [14]. We first observe quantitatively that, for both tasks, both approaches reach similar performances and none of them overfits data. We then use recent explanation methods [7] to identify the most important words for both frozen and fine-tuned models. We find that fine-tuned models rely on more "intuitively-relevant" words for the considered classification tasks than the frozen ones, leading to more interpretable results. We also visualize the impact of fine-tuning on the data representation through nonlinear dimension reduction methods [3], observing that the embedding structure of the fine-tuned models leads to much more separated clusters and concentrated information across fewer dimensions. As the fine-tuned approach outperforms the frozen one in all regards for our case studies, we conclude by discussing the possible loss of generality that fine-tuned embeddings can lead to (compared to frozen ones). For this purpose, we evaluate the performance loss in the context of a task A when using a fine-tuned embedding obtained for a different task B. We observe that the loss in performances is negligible whereas the impact on the embedding space is more significant, raising the challenge of better understanding such a context from the interpretability viewpoint as an interesting open problem.

2 Tasks, datasets, and experiment design

We use two datasets to conduct our experiments. The first one is a homemade dataset, called the InfOpinion dataset (which will be made public upon acceptance). It contains 10,000 news extracted from the "Radio-Télévision Belge de la communauté Française" (RTBF) corpus and it was built to train and evaluate a classification model distinguishing between texts from the journalistic *opinion* genre (such as editorials, commentaries, reviews), which are considered to be subjective, and texts belonging to the *information* genre (press agency dispatches, news articles), meant to be more objective. This binary categorization relies solely on the articles' annotation by the RTBF as either *opinion* or *information*. The second dataset is the Allocine dataset (`https://huggingface.co/datasets/allocine`). It is composed of 200,000 movie reviews that can be positive (50.02%) or negative (49.98%). Both datasets are split in 3 parts: a training set (80%), a validation set (10%) and a test set (10%). For both datasets, the task is to predict the binary category of a given text.

We start our experiments by training both the fine-tuned and frozen models. In the first case, we fine-tune the model on the training set during 2 epochs, as presented in [6]. In the second case, we use the base version of CamemBERT

[14] as our transformer model. We first probe the internal representation of each text from the training set at the last layer of the pre-trained model, just before the RoBERTa classification head. As in [5], we then use the first token's representation as a text embedding and train a RoBERTa classification head during 20 epochs on top of these embeddings. The model accuracy is evaluated at each epoch on the validation set and, at the end of the training process, on the test set. Once both models are trained, we use the Layer-wise Relevance Propagation (LRP) method [7] to collect word-level explanations for every text. We compute the mean relevance across all texts for each word and compare the top words for our two models. We finally study the latent representations of both the frozen and the fine-tuned models. We use multi-scale t-SNE [3] to visualize the text embeddings (which are in 798-D) in 2-D, while also analyzing the concentration of information through Principal Component Analysis (PCA). We conclude by initiating a discussion on the loss of generality induced by fine-tuning, by using text embeddings fine-tuned on one classification task for another task. The code of our experiments will be made public upon acceptance.

3 Results and discussion

3.1 Model performances

Table 1 reports the accuracies of both models on both datasets. For the Allocine dataset, the fine-tuned model is slightly (but statistically significantly, $p<<0.01$) better, confirming previous works like [5, 9]. For the Infopinion dataset, the differences are not even statistically significant ($p = 0.34$). We also note that the performances on the test set are similar to the one on the training set (given in parentheses), suggesting that no overfitting of the data occurs.

	Accuracy fine-tuned	Accuracy frozen
Allocine	97.07 (97.49)	94.31 (93.77)
InfOpinion	95.60 (96.36)	96.20 (97.10)

Table 1: Classification accuracy on the Allocine and the InfOpinion datasets

3.2 Discriminant words

We continue our study by applying the LRP explanation method [7] to obtain word-level explanations for each text in the dataset. We then compute the mean relevance across all appearances of each word to get a list of the most discriminant words according to the models (independent of the context surrounding these words). In order to limit noise, we only consider words appearing at least 100 times in the datasets. We finally highlight the difference between the two models by ranking the words according to difference in mean relevance between the frozen and fine-tuned models. Below are the most different words for the

Allocine dataset (movie reviews)[1], translated into English:

- **More important for fine-tuned**: bouleversant (upsetting), régal (threat), adoré (liked), émouvant (touching), plat (flat), bijou (jewel), magnifique (magnificent), merveille (wonder), poignant (touching), éviter (to avoid)
- **More important for frozen**: regardez (watch), vieilli (aged), revu (already seen), penser (to think), Besson, sauce, proche (close), mourir (die), visuellement (visually), d' (of)

Interestingly, and despite model interpretation is admittedly harder to evaluate on quantitative bases, we observe that the improved performances of the fine-tuned model are not based on data artifacts but that, instead, its top words well reflect the considered task (i.e., they are connected to the appreciation of the movies). This is in contrast with the frozen approach, where the list of most relevant words rather relate to the movie style (which, without additional context, is not directly related to its appreciation). This suggests that despite similar performances, the frozen model is not based on the same easily interpretable features as the fine-tuned one. We next try to give a complementary viewpoint to this observation by analyzing the embedding spaces of the two models.

3.3 Text embeddings visualisation

The next figures show projections of the text internal embeddings for the frozen and fine-tuned models on both datasets, computed with multi-scale t-SNE [3].

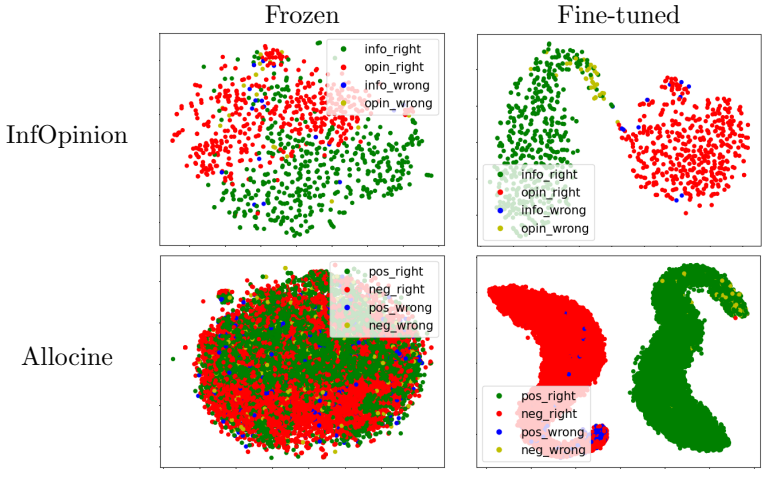


Table 2: t-SNE visualisations of the model's embedding space on the Allocine and the InfOpinion datasets

[1] Similar trends hold with the InfOpinion dataset, not displayed due to space limits.

These results suggest that the decision boundaries (in two dimensions) are clearer and that clusters with distinct labels are more separated using the fine-tuned internal text representations than the frozen ones, supporting the previous results of [8]. In addition to pushing clusters with different labels far away from each other, the fine-tuning also concentrates the information across fewer dimensions. Through a PCA of the internal text embeddings, we also computed the % of retained variance with respect to the number of Principal Components (PC). It turns out that the 1^{st} PC captures more than 95% of the variance for the fine-tuned internal text embeddings, whereas more than 200 PCs are necessary to preserve the same amount of variance in the frozen case.

4 Conclusion and further works

This works highlights the specificities of the fine-tuned vs. frozen models on two different data sets and classification tasks. While both methods can reach similar accuracies and none of the methods overfits the training data in our case studies, fine-tuned models appear to be more interpretable and to concentrate the information of their text embeddings in fewer dimensions, despite being more complex in terms of number of updated parameters compared to frozen ones.

These promising results for the fine-tuned approach naturally raise the question of the possible loss of generality they could come up with. As an opening experiment towards discussing this question, we finally performed a cross fine-tuning experiment, where a model is first fine-tuned on a task, and its text embeddings are then used to perform another task. A similar experiment was presented in [8] but focused on different classification tasks with the same dataset. We extend it to different tasks based on different datasets. The table below shows the accuracy obtained through this process:

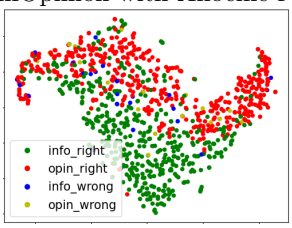
	F-t Allocine	F-t InfOpinion	Frozen
Allocine	97.07	94.32	94.31
InfOpinion	95.50	95.60	96.20

Table 3: Classification accuracy on the cross-task finetuning experiment

It is noteworthy that, even when using a different dataset, fine-tuning reaches very competitive accuracies, suggesting that it can preserve most of the relevant information of the frozen internal text embeddings. The figures below then depict the internal text embeddings of each dataset using cross-task fine-tuning:

This last figure is more contrasted since it suggests that only a part of the "information concentration" observed in Section 3.3 is maintained in a cross-task setting. This is also confirmed by the number of PC needed to capture 95% of the variance, which is worth 37 for the left figure (InfOpinion with Allocine fine-tuning) and 89 for the right one (Allocine with InfOpinion fine-tuning).

InfOpinion with Allocine f-t



Allocine with InfOpinion f-t

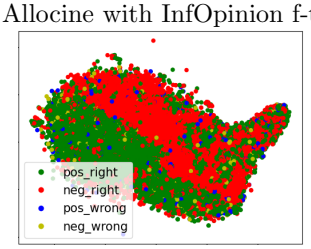


Table 4: t-SNE visualisations of the model's embedding space with cross-task finetuning

How much this last observation is due to the good/bad generalization potential of fine-tuned models or to the differences/similarities of the datasets and tasks considered in our experiments is an interesting open question. As a first step towards answering it, it would be interesting to quantify whether performance drops accumulate after multiple cross-trainings, and whether the intermediate "information concentration" of these last figures translates into reduced interpretability, for example using the techniques used in Section 3.2. Finally, because we focused on a case study on Camembert for French language, it's hard to know whether our findings generalize to other models or languages. This question is let for further works.

References

[1] R. Thomas McCoy et al., Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference, in CoRR, abs/1902.01007, 2019, arXiv

[2] A. Merchant et al., What Happens To BERT Embeddings During Fine-tuning?, in Proceedings of the Third BlackboxNLP Workshop (EMNLP), p 33–44, November 2020, published by ACL

[3] C. de Bodt et al., Fast Multiscale Neighbor Embedding, in IEEE Trans. Neural Netw. Learn. Syst., p 1-15, 2020

[4] O. Kovaleva et al., Revealing the Dark Secrets of BERT, in Proceedings of the 2019 EMNLP-IJCNLP, p 4364–4373, November 2019, published by ACL

[5] M. E. Peters et al., To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks, in Proceedings of the 4th Workshop on Representation Learning for NLP, p 7–14, August 2019, published by ACL

[6] J. Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in Proceedings of the 2019 NAACL-HLT , Volume 1 p 4171–4186, June 2019, published by ACL

[7] H. Chefer et al., Transformer Interpretability Beyond Attention Visualization, in CVPR 2021, p 782–791, published by Computer Vision Foundation / IEEE

[8] Y. Zhou and V. Srikumar, A Closer Look at How Fine-tuning Changes BERT, in Proceedings of the 60th Annual Meeting of the ACL, Volume 1 p 1046–1061, May 2022, published by ACL

[9] N. F. Liu et al., Linguistic Knowledge and Transferability of Contextual Representations, in Proceedings of the 2019 NAACL-HLT, Volume 1 p 1073–1094, 2019, published by ACL

[10] S. Gururangan et al., Annotation Artifacts in Natural Language Inference Data, in Proceedings of the 2018 NAACL-HLT, Volume 2 p 107–112, June 2018, published by ACL

[11] A. Vaswani et al., Attention is All you Need, in NEURIPS 2017, p 5998–6008, 2017

[12] Y.Liu et al., RoBERTa: A Robustly Optimized BERT Pretraining Approach, in CoRR 2019, arXiv

[13] Z. Lan et al., ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, in ICLR 2020, April 2020, published by OpenReview.net

[14] L. Martin et al., CamemBERT: a Tasty French Language Model, in Proceedings of the 58th Annual Meeting of the ACL, p 7203–7219, July 2020, published by ACL