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

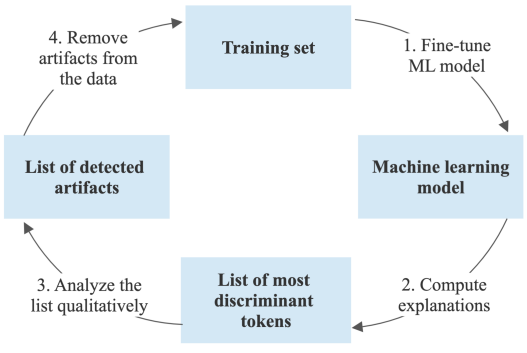

FIGURE 1 – TIPECS framework to remove artifacts from a dataset.

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.

Index