Automatic Categorization of ESWD Weather Reports in French

Mija Pilkaite [1], **Davide Buscaldi** [2]

1. *LIX, Ecole Polytechnique*
   91120 Palaiseau, France
   mija.pilkaite@polytechnique.edu
2. *LIPN, CNRS UMR 7030*
   Université Sorbonne Paris Nord
   93430 Villetaneuse, France
   buscaldi@lipn.fr

RÉSUMÉ. *Dans le contexte des préoccupations liées au changement climatique, les phénomènes météorologiques violents représentent une problématique majeure en raison de leur incidence sur la société humaine. Les chercheurs ont collecté des données sur ces événements afin de mieux comprendre leur corrélation avec le changement climatique et d'améliorer notre capacité à les prédire et à nous y préparer. Le Laboratoire européen des tempêtes violentes (ESSL) a mis en place la base de données européenne sur les phénomènes météorologiques violents (ESWD), permettant au public de signaler des évenements de ce type. L'utilisation du traitement automatique de diverses sources médiatiques, telles que les actualités et les médias sociaux, a suscité un intérêt croissant afin d'identifier et de recenser les phénomènes météorologiques violents de manière plus précise et objective. Ce travail se concentre donc sur l'utilisation de différentes techniques pour extraire les informations pertinentes et rendre l'ESWD moins dépendant de l'intervention humaine.*

ABSTRACT. *Amidst the concerns for climate change, severe weather events represent an important issue because of their impact on human society. Researchers have been collecting data regarding these kinds of events to try to understand their relationship to climate change and improve our ability to predict and prepare for them. The European Severe Storms Laboratory has created the European Severe Weather Database (ESWD), which allows the public to report and share information about severe weather events. To address this issue, there has been growing interest in using automatic processing of various media sources, such as news and social media, to identify and survey severe weather events more accurately and objectively. Thus, this work focuses on using different techniques to extract the relevant information and make the ESWD less human-dependent.*

MOTS-CLÉS : *European Severe Weather Database, Classification, TAL*

KEYWORDS: *European Severe Weather Database, Classification, Natural Language Processing*

1. Introduction

As the advancements in Machine Learning have been fast-paced, we are looking for more adaptations and uses for it in the science world. Climate change has been also a burning topic for several years, thus this project provides an opportunity to connect these two fields to better explore the severe weather events in Europe. The analysis and prediction of severe weather events are of great importance, as extreme weather conditions can lead to significant economic and human losses (C. A. Doswell, Kay, 2005) and accurate categorization and extraction of severe weather events can help understand the frequency, intensity, and distribution of such disasters, as well as for planning and designing effective mitigation strategies (Jessica Mercer, Taranis, 2009). The European Severe Weather Database (ESWD) serves as a valuable resource for weather data, containing reports of severe weather events in Europe within most of the countries and a wide range of categories (Groenemeijer *et al*., 2009). The ESSL [1] (European Severe Storms Laboratory) started as an informal network of European scientists to advance research on severe convective storms and extreme weather events on a European level; today it is managing an actively growing database that could significantly benefit from automation, particularly in incorporating events sourced from news and social media. This idea is at the core of this preliminary study - assessing the viability of leveraging Natural Language Processing techniques to discern whether a news article references a severe weather event worthy of inclusion in ESWD, and in the affirmative case what would be the category of the event.

2. Data Description and Preprocessing

We obtained from ESSL a set of 22,317 entries from the ESWD, covering severe weather events that occurred in France from January 1, 2016, to December 31, 2022. For each event, 22 fields are present: among these a unique id, timestamp, latitude and longitude, location, region, meteorological data, number of victims and injured, the type of the event, an event description, and a reference (usually the title and the link to an on-line article). Since we wanted to predict the event type from text, we used the ‘EVENT TYPE’ column as target and the ‘REFERENCE’ column for the input text (the event description is often empty as it’s an optional short description added to the event). As it is shown in Figure 1, there are 8 different severe weather events and the dataset is unbalanced.

However, of the total 22,317 entries, only 12,867 contain a non-empty reference, reducing the set of available data. We also had to preprocess the data to remove empty or non-informative references, such as the following ones:

"Joel Feneule (on Facebook), 17 Nov 2022."
"Report via Kachelmannwetter.com, 4 Nov 2022."

1. https://www.essl.org/cms/

Finally, we used Spacy [2] language detection to keep only the texts in French. After this step, we obtained a set of 3,246 reports. More pre-processing was required to clean reports such the following one:

"Eyewitness report via Alpes 1 (on Facebook), 14 SEP 2022. ""Hautes-Alpes : un violent orage de grêle s'est abattu sur Gap"

in order to keep only the second part of the reference. This was done with hand-made rules (for instance, removing all text matching patterns such as "eyewitness report" or texts that are below a length threshold of 5 words) and the use of Spacy's language detection features.

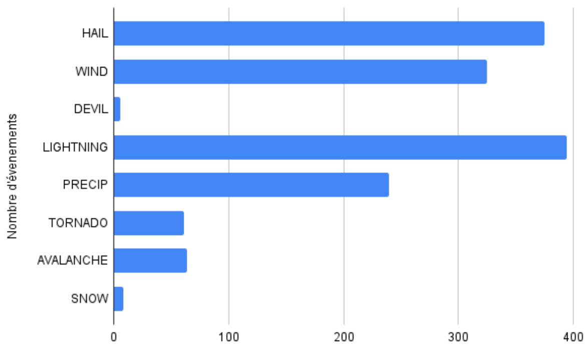

FIGURE 1. *Distribution of the 8 event types in the final dataset.*

At the end of this pre-processing step, we kept $1,472$ events distributed as seen in Figure 1. As it can be seen, the wind category is not so dominant as in the full one. The reason is that many of the wind reports come from meteorological station reports, therefore there is no meaningful text associated.

3. Text Classification Models

We classified the text focusing on various types of text representations and classification algorithms.

Text Representation We represented the text of the reports in three different ways: as Bag-of-Words with the frequency of the words, tf.idf (term frequency-inverse document frequency) vectors, and dense vectors obtained using Sentence-BERT. For both BoW and tf.idf we remove stopwords and keep the words that occur at least 2 times. In this way, reports are represented by sparse vectors of size $2,918$. Sentence-BERT (SBERT) is a sentence encoder based on Siamese architecture, which modifies the

2. https://spacy.io/

BERT (Bidirectional Encoder Representations from Transformers) architecture, showing state-of-the-art performance in various natural language processing tasks (Devlin *et al.*, 2018). SBERT can be used to encode each sentence into a dense vector of size 768.

Classification Methods We considered three widely used algorithms: logistic regression (LR), random forests (RF) and a fully connected neural network (FCNN). For the random forest classifier we employed hyperparameter tuning using GridSearch with 3-fold cross validation. The FCNN was set up with two hidden layers of 512 units each both with Dropout layer (dropout rate 0.2) and a Softmax output with categorical cross-entropy loss. We also set the batch size at 64 and 20 training epochs. We considered fine-tuning a BERT model, but we faced problems due to the imbalance of the dataset (the model was predicting always the same label).

4. Results and Analysis

We split the data into a random partition, using 80% of data for training and 20% for testing. The results across all representations and methods are shown in Table 1.

TABLEAU 1. *Accuracy obtained using the various text representation methods and classification methods.*

Text Representation	LR	RF	FCNN
BoW	82.9%	68.9%	82.6%
tf.idf	83.9%	68.7%	81.6%
SBERT	82.1%	67.8%	75.1%

From the results, it is evident that Sentence-BERT is not adequate to represent the data in this kind of task. Bag-of-Words is a representation that on average yields the best results across all methods. Tf.Idf is only better with the LR classifier. This result can be explained by the fact that the rare words (that are boosted by idf) are not important clues for understanding the category of the event. Random Forests performed poorly on the dataset. From the confusion matrix in Figure 2, it can be seen that this model is never able to predict the PRECIP class, being split between LIGHTNING and WIND. As it is evident from Figure 3, the discriminative power of single words tends to be weak and there is a bias towards the most represented classes.

We examined the best results to identify the situations that still posed problems for classification. First of all, we calculated the confusion matrix in Figure 4. From this matrix, it can be seen that PRECIP is the category less easily identified as it is often confused with WIND.

Let us look at some misclassified examples:

Actual: LIGHTNING Predicted: WIND Reference: ['Orages', 'départ', 'feu', 'maison', 'détruite', 'arbre', 'couché', 'tarn', 'Haute-Garonne', 'tarn-et-garonne', 'FRANCE', 'TV', 'INFO', '30', 'aug', '2022']

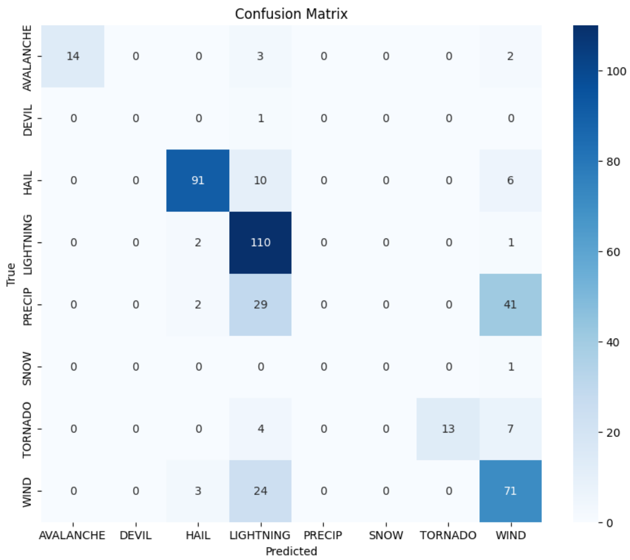

FIGURE 2. *Confusion matrix for tf.idf and Random Forests*

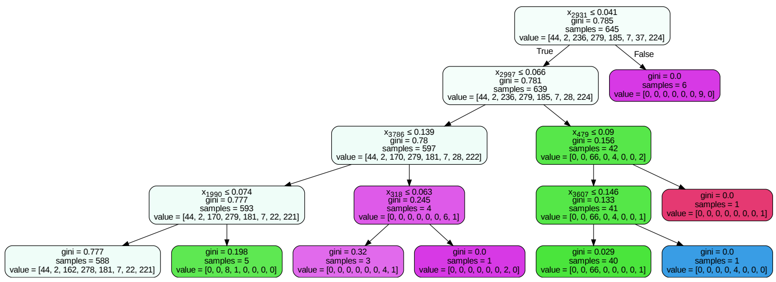

FIGURE 3. *A single Decision Tree from the RF model with tf.idf weights*

Actual: HAIL Predicted: WIND Reference: ['violent', 'orage', 'évacuation', 'Saint-Etienne', 'femme', 'prisonnier', 'voiture', 'Villars', 'Progrès', '01', 'Jul', '2019']

Actual: LIGHTNING Predicted: HAIL Reference: ['intempérie', 'orage', 'faire', 'gros', 'dégât', 'ouest-aveyron', 'vendredi', 'soir', 'ladepeche.fr', '29', 'june', '2020']

Most of the misclassifications stem from words that can be used for multiple events. For example, the word 'orage' for the model is typically associated with wind, but it can be also associated to lightning and hail as we can see in the first two examples.

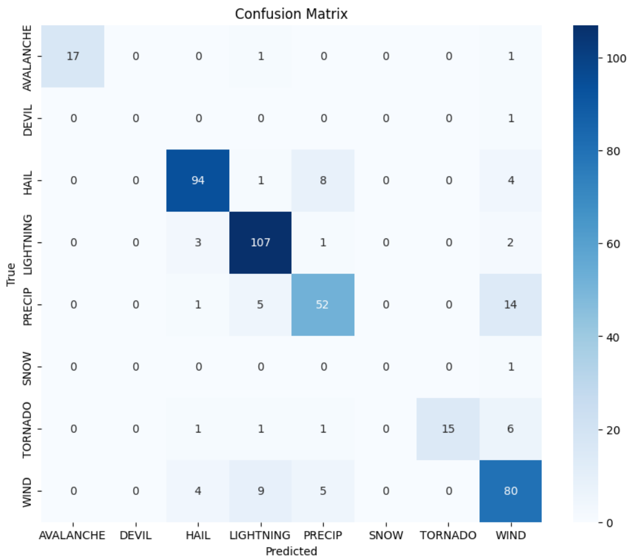

FIGURE 4. *Confusion matrix for tf.idf and Logistic Regression*

We then extracted from the LR model the most important features (words) for each target class. In Table 2, we sort the top 10 features by their weight magnitude in descending order.

TABLEAU 2. *Most important features for each event type.*

	AVALANCHE	DEVIL	HAIL	LIGHTNING	PRECIP	SNOW	TORNADO	WIND
1	avalanche	tourbillon	grêle	foudre	inondation	neige	tornade	ws
2	skieur	jul	grêlon	incendie	orages	15	2019	meteofrance
3	avalanch	apr	limousin	foudroyer	inonder	jan	oct	arbre
4	jan	jardinier	jun	kachelmannwetter	boue	000	facebook	mini
5	savoie	ferté	agriculteur	jun	pluie	souffert	nov	feb
6	2021	bernard	eyewitness	sep	eau	le	dec	tempête
7	dec	surprendre	report	maison	provence	lorrain	direct	vent
8	alpes	ouest	centre	nord	oise	alsace	keraunos	coup
9	mort	tornad	bilan	feu	orage	chute	youtube	report
10	feb	argelès	21	com	progrès	priver	azur	march

As we have seen through the process, there are various keywords that help our models identify which severe weather event the reference is talking about. However, this also leads us to another pressing question - how do we make that our models do not confuse references about not-weather events or about weather events that are not severe if they contain these keywords? This is explored in the final part of this work.

5. Binary classification of Severe Weather events

Besides classifying the event types, it is crucial to have a classifier that could distinguish between severe weather events and not or even other articles that contain similar keywords. For this reason, we scraped from the web 500 headlines containing some of the most important keywords seen in Table 2, and built a classifier with these headlines (negative class) and 500 randomly picked events from the ESWD database, labeled as positive samples.

With this setup and 10-fold cross-validation, we obtain for tf.idf and LR 98.9% accuracy, indicating that it is possible to effectively diversify the ESWD-worthy reports from general news related to weather.

Some examples of misclassified instances:

– Haute-Savoie glissement terrain bloque route Thyez dauphiner Libéré 15 July 2021 for suscriber only (True label: 1, Predicted label: 0)

– grêle sud-est orage l'Ouest (True label: 0, Predicted label: 1)

– météo France surprenante (True label: 0, Predicted label: 1)

– Mickael B. observatoire ciel Orageux Tornade Médoc 2018 2019 (True label: 1, Predicted label: 0)

As it can be seen, for the second example the model is probably correct as we didn't filter out those headlines referring to severe weather. In the other cases, the context is not big enough, which represents probably the main difficulty of this task.

6. Conclusions

In this work, we explored the classification of severe weather events using textual data from the European Severe Weather Database, using various classification models. Overall, most of the models tested achieved good performance with $\sim 80\%$ accuracy, indicating the textual data contains meaningful signals to distinguish between different types of severe weather events; although the imbalance of classes poses an obstacle to further improvement of these results. We have also seen that the data collected by volunteers is especially noisy and is not held to any standard, thus making any sort of processing quite a difficult task. We also developed a binary classifier to filter out false data that contains key weather-related terms but does not actually describe a severe weather event. This classifier achieved over 98% accuracy, indicating that it is possible to create a severe weather monitor for news and social media to enrich the ESWD with automatically collected information.

There are several promising avenues for future work based on this project. With a larger dataset, deep learning models may achieve even higher accuracy in classifying severe weather events and sub-categories. Contextualized word embedding models like BERT (Devlin *et al.*, 2018) and RoBERTa (Liu *et al.*, 2019) should also be explored as they have achieved state-of-the-art results on various text classification tasks.

Bibliographie

C. A. Doswell H. E. B., Kay M. P. (2005). Climatological estimates of daily local nontornadic severe thunderstorm probability for the united states. *Weather and Forecasting*, vol. 20, nº 4, p. 577-595.

Devlin J., Chang M.-W., Lee K., Toutanova K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Groenemeijer P., Kuehne M., Liang Z., Dotzek N. (2009). New capabilities of the european severe weather database. *In 5th European conference on severe storms, Landshut, Germany*, p. 311-312.

Jessica Mercer I. K., Taranis L. (2009). Framework for integrating indigenous and scientific knowledge for disaster risk reduction. *Climate Change Modeling, Mitigation, and Adaptation*, vol. 34, nº 1, p. 214-239.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D. *et al.* (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.