ETH zürich

ETH Library

Team "DaDeFrNi" at CASE 2021 Task 1: Document and Sentence Classification for Protest Event Detection

Conference Paper

Author(s):
Re, Francesco Ignazio; Vegh, Daniel; Atzenhofer, Dennis; Stöhr, Niklas Werner

Publication date:
2021

Permanent link:
https://doi.org/10.3929/ethz-b-000508467

Rights / license:
Creative Commons Attribution 4.0 International

Originally published in:
https://doi.org/10.18653/v1/2021.case-1.22

Funding acknowledgement:
787478 - Nationalist State Transformation and Conflict (EC)

This page was generated automatically upon download from the ETH Zurich Research Collection.
For more information, please consult the Terms of use.

Team "DaDeFrNi" at CASE 2021 Task 1: Document and Sentence Classification for Protest Event Detection

Francesco Ignazio Re Dániel Végh Dennis Atzenhofer Niklas Stoehr
ETH Zurich, Switzerland
{franre, davegh, dennisa}@ethz.ch niklas.stoehr@inf.ethz.ch

Abstract

This paper accompanies our top-performing submission to the *CASE 2021* shared task, which is hosted at the workshop on *Challenges and Applications of Automated Extraction of Socio-political Events from Text*. Subtasks 1 and 2 of Task 1 concern the classification of newspaper articles and sentences into "conflict" versus "not conflict"-related in four different languages. Our model performs competitively in both subtasks (up to 0.8662 macro F1), obtaining the highest score of all contributions for subtask 1 on Hindi articles (0.7877 macro F1). We describe all experiments conducted with the XLM-RoBERTa (XLM-R) model and report results obtained in each binary classification task. We propose supplementing the original training data with additional data on political conflict events. In addition, we provide an analysis of unigram probability estimates and geospatial references contained within the original training corpus.

1 Introduction

Can natural language processing (NLP) be leveraged to extract information on socio-political events from text? This is an important question for Conflict and Peace Studies, as events like protests or armed conflicts are frequently reported in textual format, yet are costly to extract. The workshop on *Challenges and Applications of Automated Extraction of Socio-political Events from Text* (CASE 2021) aims at bringing together political scientists and NLP researchers to improve methods for automated event extraction[1]. As part of this workshop, a shared task is proposed to advance progress on various problems associated with reliable event detection (Hürriyetoğlu et al., 2021).

We combine the data provided by CASE 2021 with additional data sources to train a XLM-RoBERTa (XLM-R) model for subtasks 1 (document classification) and subtask 2 (sentence classification). Our model reaches competitive F1 scores ranging between 0.730 and 0.866 and is best-performing amongst all submissions for document classification in Hindi. Our exploratory analysis unveils relevant insights into the training data provided in the shared task. We find differences in the use of state versus non-state conflict actors based on conditional probabilities, and we identify an outlier in the English corpus via a Tf-Idf-weighted principal component analysis (PCA). Moreover, we conduct an analysis of the geospatial patterns in the underlying data. This report proceeds as follows: First, we briefly outline the datasets that we use. In sections 3 and 4 we elaborate on our model selection and on various conducted experiments. Finally, we report the results for subtasks 1 and 2. With these results in mind, section 6 delves into an exploratory analysis of the training data to better understand potential pitfalls.

2 Dataset

In order to train our model, we leverage the data provided by the organizers as well as additional data on political conflict events. In this section, we describe both of these datasets.

2.1 Dataset provided for the shared task

The data for the CASE 2021 shared task derives from the *Global Contention Dataset* (GLOCON Gold) (Hürriyetoglu et al., 2020), a manually annotated dataset containing news articles in various languages. The training data consists of texts in three different languages: English articles from India, China, and South Africa, Spanish articles from Argentina, and Portuguese ones from Brazil. For subtask 1, the texts are labelled on the document level, with a binary label indicating whether the document mentions a political conflict event or not.

[1] This workshop is a continuation of the shared tasks CLEF 2019 Lab Protest News (Hürriyetoğlu et al., 2019), and event sentence co-reference identification task at AESPEN at LREC 2020 (Hürriyetoğlu et al., 2020).

Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) , pages 171–178
August 5–6, 2021, ©2021 Association for Computational Linguistics

For subtask 2, these documents are broken down to individual sentences, again with a binary label indicating whether the particular sentence mentions a political conflict or not. Crucially, the training data does not contain texts in the Hindi language, while Hindi texts are contained within the testing set. With a limited amount of texts to learn from, we consider expanding the training data in multiple ways, which we elaborate on in the following.

2.2 Extension with conflict event datasets

In order to fine-tune our model, we aim to extend the training data. To do so, we rely on two strategies: supplementing with data from other sources and translating the original training data.

For conflict-related texts, we harness a dataset provided by the Europe Media Monitor (EMM) (Atkinson et al., 2017; Pierre et al., 2016). This allows us to not only add more English texts, but also provides more Spanish and Portuguese data instances. Specifically, we rely on the human annotated data of the EMM project[2], thus we can be confident that these texts are indeed conflict-related. In addition, we supplement the English training set with data from the *Armed Conflict Location & Event Data Project (ACLED)* (Raleigh et al., 2010).
In order to obtain more negative examples (sentences not mentioning an event) and to add texts in Spanish and Portuguese, we web-scrape various newspaper articles linked on Twitter[3]. To make sure that these articles do not pertain political conflicts, we select only articles that are featured in tweets mentioning words unrelated to conflict[4]. Our second strategy to increase the available information is to translate the original training data. Using the Google Translate API we translate each text into all languages relevant for the task. This also equips us with texts in Hindi to train our model on. Overall, these efforts enable us to increase the available training data substantially:

- T_0: dataset related to subtask 1 as provided in the shared task.
- T_{mix}: combined dataset of subtask 1 and 2.
- $T_{0_{\text{noNER}}}$ and $T_{mix_{\text{noNER}}}$: The previously defined datasets with named entities removed.
- T_1, T_2, T_3: This data includes the articles from the additional sources. The datasets are constructed in a way so that the ratio between positive and negative labels is the same as in T_0. T_1 does not contain any of additional data obtained through translation, while T_2 and T_3 contain all the additional data. The difference among the two is that T_2 undergoes pre-processing steps (removal of punctuation and tags), whereas T_3 is fed into the model without being manually pre-processed first.

3 Model selection

Informed model selection is crucial for competitively solving the task. We choose pre-trained *Transformer*-based (Vaswani et al., 2017) classification models due to their state-of-the-art performance in various tasks (Devlin et al., 2019; Valvoda et al., 2021). Given the fact that the provided dataset is multilingual, we face a crucial design decision: option (a): select a monolingual model e.g. BERT (Devlin et al., 2019), that is pre-trained on huge, unlabeled text corpora in English with the need to translate all the other languages in the dataset back to English, then fine-tune the model on that. Or option (b): choose a multilingual model e.g. a multilingual version of BERT(mBERT), XLM((Lample and Conneau, 2019) or XLM-Roberta (XLM-R)(Conneau et al., 2020), that handles multiple languages simultaneously and fine-tune the model on the original languages. We ultimately choose the XLM-R model to experiment with. Recent results suggest that multilingual models achieve better performance, especially for low-resource languages.

4 Experiments

To conduct our experiments we rely on implementations provided by the Huggingface library [5][6]. For experiment tracking we make use of Wandb library [7]. After several rounds of hyperparameter search, we select a batch size of 16, learning rate of 2e-5, weight decay of 0.01 and train for 4 epochs. We train models for each of the subtasks separately (T_0), then we experiment with combinations of datasets, mixing subtasks and languages (T_{0mix}).

[2] https://labs.emm4u.eu/events.html
[3] We use the Python library *Newspaper3k*
[4] Specifically, we filter for mentions of "fashion", "football", "art","festival","movie". Including news reports on sport events could be particularly useful, since they are often described with language that is reminiscent of conflict
[5] https://huggingface.co/
[6] We open-source our code at https://github.com/denieboy/ACL-IJCNLP_2021_workshop
[7] https://https://wandb.ai/site/

Task 1	Subtask-1					Subtask-2				
Dataset	**en**	**es**	**pr**	**hi**	**avg**	**en**	**es**	**pr**	**hi**	**avg**
T_0	0.8650	0.8023	0.7572	-	0.8082	0.8717	0.8560	0.8811	-	0.8609
T_{0mix}	0.8711	0.7702	0.7841	-	0.8085	0.8720	0.8217	0.8835	-	0.8591
T_{0noNER}	0.7788	0.8138	0.8430	-	0.8119	0.9007	0.8484	0.8667	-	0.8719
T_{0mix_noNER}	0.8616	0.8056	0.7630	-	0.8101	0.8679	0.8320	0.8565	-	0.8521
T_1	0.8547	0.8011	0.7935	0.8241	0.8183	0.8780	0.8098	0.8785	-	0.8554
T_2	0.9111	0.8718	0.8468	0.8386	0.8671	0.9348	0.8670	0.8896	-	0.8971
T_3	0.8860	0.8895	0.8704	0.8546	0.8751	0.9695	0.9305	0.8948	-	0.9316

Table 1: F1 macro scores for task 1 subtasks 1 and 2 obtained with models fine-tuned on different dataset

Task 1	Subtask-1					Subtask-2				
Dataset	**en**	**es**	**pr**	**hi**	**avg**	**en**	**es**	**pr**	**hi**	**avg**
submission	0.8069	0.7301	0.7722	0.7877	0.7742	0.7928	0.8517	0.8662	-	0.8369

Table 2: F1 macro scores on the final test set achieved by our best model

We achieve the best results when training on the combined dataset including all the languages.

We try different combinations of extensions (T_1-T_3), e.g. having a balanced dataset or keeping the original imbalance rate of the shared task data. Finding protest events in Hindi language is challenging. Therefore, we translate protest events from English sources. Additionally, we experiment with removing contextual information and basing our classification on linguistic patterns only. To this end, we remove all named entities from the dataset (T_{0noNER}-T_{0mix_noNER}). The results, surprisingly, reveal only a slight degradation compared to the original dataset and even a small increase in performance on subtask 2 on English text.

5 Results

In this subsection we present the results achieved by our XLM-R models fine-tuned on different datasets. Table 1 shows the F1-macro score achieved on the different train / validation splits. Generally, we find that increasing the amount of the training data yields better scores. In Table 2, we present an evaluation of our model on the test set, on which we achieve F1-macro scores up to .867.

6 Discussion

In this section we present and analyse the conflict event data corpus, performing a descriptive analysis on the dataset using unigram probabilities and geospatial coordinates.

6.1 Unigram probability estimation

We take a probabilistic perspective and model the relation between the content of each document and its associated label considering texts as bags-of-words. Examining the different datasets provided for subtask 1, we study the three corpora (English, Portuguese and Spanish) independently.

6.1.1 Conditional probability estimates

We treat the terms "unigram" and "word" interchangeably. Given a word w, we denote the probability $P(D = 1|w)$ as the probability that the word w comes from a document . Similarly, we define $P(w|D = 1)$ as the probability that a conflictual document contains the word w. We estimate $P(w|D)$ with $\hat{\pi}_{w|D}$ and $P(D|w)$ with $\hat{\pi}_{D|w}$. Hence, we have

$$\hat{\pi}_{w|D} = \frac{\sum_{d_1 \in \mathcal{D}_1} \mathbb{1}\{w \in d_1\}}{\sum_{j=1}^{|V|} \sum_{d \in \mathcal{D}_1} \mathbb{1}\{w_j \in d_1\}}$$

$$\hat{\pi}_{D|w} = \frac{\sum_{d_1 \in \mathcal{D}_1} \mathbb{1}\{w \in d_1\}}{\sum_{d \in \mathcal{D}} \mathbb{1}\{w \in d\}},$$

with $\mathcal{D}$ being the corpus of all documents in a language, and $\mathcal{D}_1$ the subset of all conflict-related documents in $\mathcal{D}$. $\hat{\pi}_{D|w}$ can also be thought as the accuracy computed on the documents containing w, while predicting all of them as conflict-related.

6.1.2 Discriminative information

In this subsection we compute the probability estimates previously introduced and present them graphically in Figure 1. In the right plot, the words are represented by $P(D = 1|w)$ on the x-axis and by $P(w|D = 1)$ on the y-axis. The words on

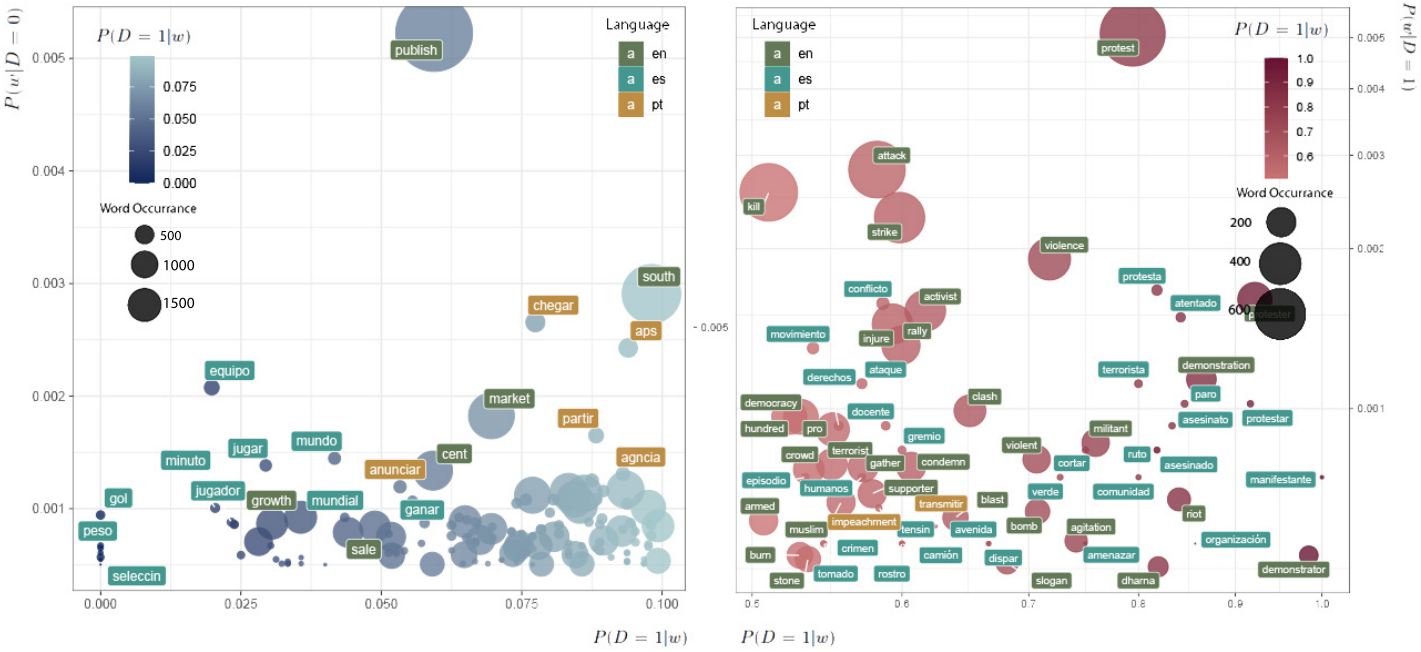

Figure 1: Sample of unigrams in the GLOCON Gold training corpora (English, Spanish, Portuguese); each circle represents a unigram, with circle size corrsponding to term frequency. For each corpus, we compute $P(D|w)$ and $P(w|D)$ as defined in Section 6.1.1. The left plot presents all unigram with low $P(D = 1|w)$ and with $P(w|D = 0) > 0.0005$. $P(w|D = 0)$ indicates how likely a unigram w is to occur in articles that are not conflict-related. Words like "growth" and "peso" contain much discriminate information - having very high $P(D = 0|w)$, but low $P(w|D = 1)$. The reverse logic applies for the right graph, displaying all the unigram with $P(w|D = 1) > 0.0005$.

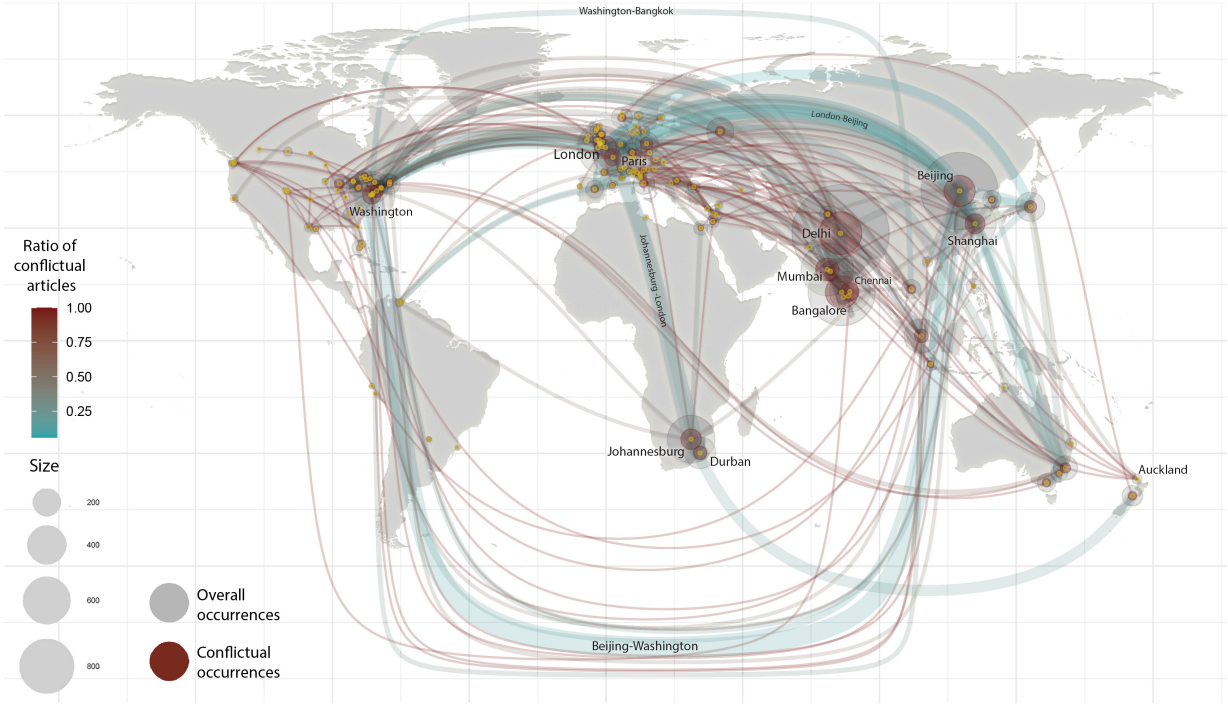

Figure 2: Undirected network of city co-mentions as introduced in Section 6.2.1; the nodes represent all cities present in the English GLOCON Gold training set. The size of the nodes corresponds to the number of occurrences that each city is mentioned. The edges are coloured according to the ratio of articles pertaining "conflict" versus "no conflict" that the cities share. The imbalanced ratio between both classes is well reflected in the map, with the light blue edges being the thickest. Edges related to conflict articles are more, but reveal lower weights.

the left plot have $P(D = 1|w)$ as the x-axis and $P(w|D = 0)$ y-axis. Indeed, a word would be a good classifier if both $P(w|D)$ and $P(D|w)$ were high. There are however no such words in our corpora. This finding reinforces our presumption that more general words contain less information relevant for our context-dependent task.

6.1.3 Result interpretation

This section summarises the information displayed in Figure 1. The right plot shows that, for words with high $P(w|D = 1)$, English ones seem to have higher $P(D = 1|w)$ if compared with Spanish and Portuguese. In fact, the Portuguese ones have $P(D = 1|w)$ not exceeding 0.7. The right plot also shows an interesting pattern with regard to conflict actors. Rather surprisingly, terms related to state-based conflict actors like `police`, `officer` or `military` do not seem to be the most useful words to identify conflict-related texts. In fact, in terms of conditional probabilities these are not very discriminatory terms for the classification (e.g. we obtain $P(D = 1|$ military $) = 0.31$, and accordingly $P(D = 0|$ military $) = 0.69$ for the English case, $P(D = 1|$ militar $) = 0.37$, and thus $P(D = 0|$ militar $) = 0.63$ for the Spanish case). On the other hand, non-state conflict actors are much more indicative of a text covering a conflict event. As seen in the graph, terms like `activist` or `protester` are highly suggestive for a conflict context. We also suspect that polarized sentiment could be a valuable indicator of conflict-related texts, because conflict-news contain negatively associated words - such as `kill`, `violence`, `terrorism` - but also terms that in certain contexts may have positive connotation, like `dharna` (peaceful protest), `democracy`, `pro`, `activist`, `supporter`. The existence of polarized sentiments among words with high $P(D = 1|w)$ could be indicative of the narrative style that is adopted for describing conflict events, with stories being usually reduced to oppressors-against-oppressed narratives.

6.2 Geospatial analysis

The analysis described in previous sections mainly focuses on words that appear with relatively high frequency in the corpus. Key contextual information of an article like place, time, actors etc. is usually very specific and thus likely to have lower frequencies. Nevertheless, contextual information plays a major role in detecting conflict events. Thus, we conduct an analysis on the geospatial entities of the English corpus provided by the shared task.

6.2.1 A geospatial undirected network

We construct an undirected network from entity co-mentions as displayed in Figure 2. The network can be seen as a symmetric matrix having as element in position (i, j) the number of times city i appears in an article where also city j is present. Nodes of the network represent the cities prevalent in the English corpus. If a document cites k cities, they will be represented in the network as a k-vertex clique. The network summarizes the relationship among the major locations involved in the events of the English set. The size of each node corresponds to the overall number of articles a city appears in. On an interpretative level, a conflictual edge does not imply that the two cities represent actors standing in conflict with each other. In fact, actors of different cities could as well be partaking in the same protest, hence sharing a common cause, rather than a divisive one. The most frequent cities cited are Indian cities such as Delhi, Bangalore, Chennai and Chinese ones like Beijing and Shanghai. In general, it is interesting to notice how the entire African continent is underrepresented if compared to others, South Africa being the only African state whose cities are mentioned (Braesemann et al., 2019; Stoehr et al., 2020).

6.3 Outlier detection with Tf-Idf

This section investigates the variability of the documents on a term-frequency level. Computing Tf-Idf embeddings for each corpus and reducing their dimensionality with PCA, we are able to detect few outliers. In particular, the document with ID 106495 in the English corpus is written in Afrikaans and not in English. A more detailed analysis can be found in the appendix.

7 Conclusion

In conclusion, the paper outlines two major contributions to the CASE 2021 shared task. Firstly, our XLM-RoBERTa model for classification Task 1.1 and Task 1.2 yields competitive results, especially for the Hindi subtask, where no training data was available. Secondly, we provide a descriptive analysis of idiosyncrasies contained with the provided text corpora. Our analysis qualitatively investigates geographical connotations in the corpora and possible outliers using word probability estimation.

Acknowledgments

Dennis Atzenhofer gratefully acknowledges financial support by the European Research Council (ERC Advanced Grant 787478).

References

Martin Atkinson, Jakub Piskorski, Hristo Tanev, and Vanni Zavarella. 2017. On the creation of a security-related event corpus. In *Proceedings of the Events and Stories in the News Workshop*, pages 59–65. Association for Computational Linguistics.

Fabian Braesemann, Niklas Stoehr, and Mark Graham. 2019. Global networks in collaborative programming. In *Taylor and Francis*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ali Hürriyetoğlu, Osman Mutlu, Farhana Ferdousi Liza, Erdem Yörük, Ritesh Kumar, and Shyam Ratan. 2021. Multilingual protest news detection - shared task 1, case 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Ali Hürriyetoglu, Erdem Yörük, Deniz Yüret, Osman Mutlu, Çagri Yoltar, Firat Durusan, and Burak Gürel. 2020. Cross-context news corpus for protest events related knowledge base construction. *CoRR*, abs/2008.00351.

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, Osman Mutlu, and Arda Akdemir. 2019. Overview of clef 2019 lab protestnews: Extracting protests from news in a cross-context setting. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 425–432, Cham. Springer International Publishing.

Ali Hürriyetoğlu, Vanni Zavarella, Hristo Tanev, Erdem Yörük, Ali Safaya, and Osman Mutlu. 2020. Automated extraction of socio-political events from news (AESPEN): Workshop and shared task report. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 1–6, Marseille, France. European Language Resources Association (ELRA).

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Soille Pierre, Burger Armin, Aseretto Dario, Syrris Vasileios, and Vasilev Veselin. 2016. Towards a JRC earth observation data and processing platform. In *Proceedings of the 2016 conference on Big Data from Space (BiDS'16)*. Publications Office.

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: An armed conflict location and event dataset: Special data feature. *Journal of Peace Research*, 47(5):651–660.

Niklas Stoehr, Fabian Braesemann, Michael Frommelt, and Shi Zhou. 2020. Mining the automotive industry: A network analysis of corporate positioning and technological trends. In *Complex Networks XI*, pages 297–308. Springer International Publishing.

Josef Valvoda, Tiago Pimentel, Niklas Stoehr, Ryan Cotterell, and Simone Teufel. 2021. What about the precedent: An information-theoretic analysis of common law. In *arXiv*, volume 2104.12133.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

A Appendix

A.1 Outlier detection with Tf-Idf

This section investigates the variability of the documents of the training corpus provided by the shared task. We try to qualitatively assess possible articles that differ significantly from the rest of the corpus.

A.1.1 Tf-Idf word representation

We produce a Tf-Idf word embedding representation of the corpus in order to gain a deeper understanding on the variability of the documents in terms of term-frequencies. Given a word w and a document d, tf-idf associates a score $\text{tf}(w, d) \cdot \text{idf}(w, \mathcal{D})$ to the word-document pair. The first term refers to how often a word occurs in a document, and the second one refers to how often a word occurs in the overall corpus.

A.1.2 Dimensionality reduction with PCA

After computing the Tf-Idf embeddings, we perform Principal Component Analyis to reduce the dimensionality of the problem. The principal components are calculated on the original Tf-Idf embedding matrix and on its normalized version, with zero mean and unit variance.The results are more interpretable on the normalized matrix, even though it disregards the idf-term of the embeddings. The analysis is carried on the three corpora independently. The representation displays most of the data points as cluttered into one dense cluster, with very few ones standing out. Among these, in the English dataset for example, the data point with ID 108218 is not in English but in Afrikaans. Another article that stands out is the one with ID 106495; it contains 16108 characters whereas the 0.99 quantile of the character length distribution per document is 6290. A graphical representation can be found in the appendix in Figure 3. In Portuguese and Spanish instead, the reason why some articles are isolated from the group is less evident and it is probably more related to the category of content that the articles talk about.

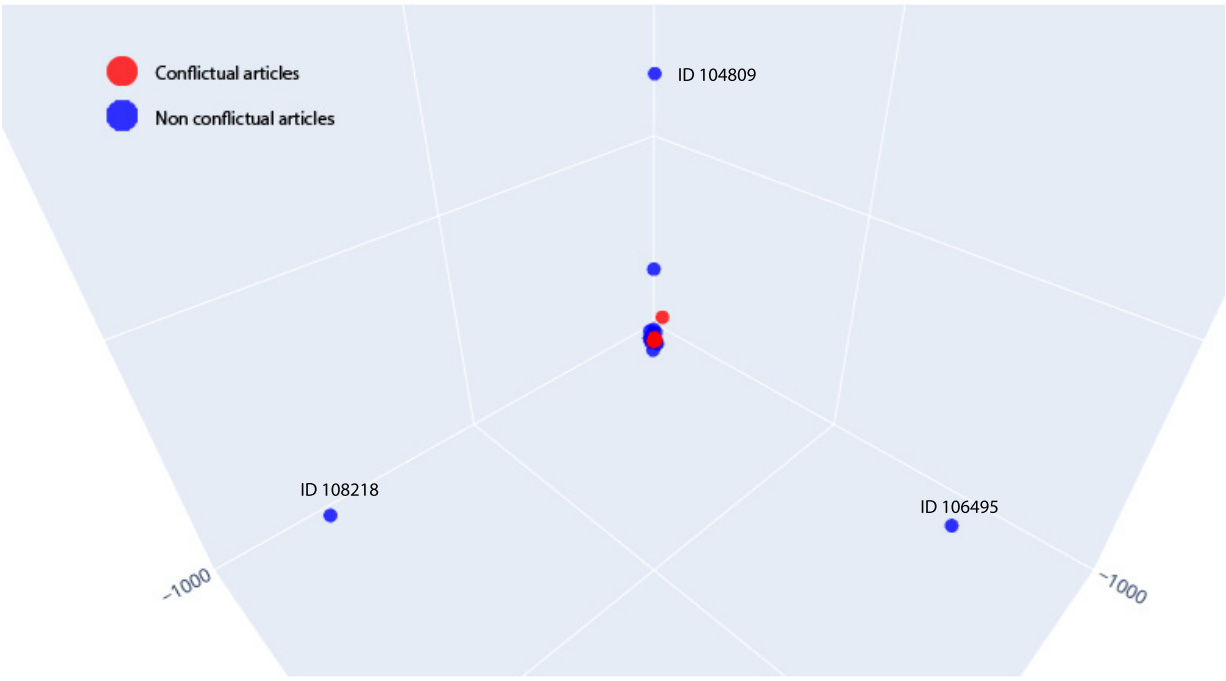

Figure 3: This figure shows the training English set with the first three principal components. Even if most of the data is concentrated in one dense cluster, there are a few points that can be very easily distinguished. They generally are either in a language different than English (ID 108218), or have other very rare characteristics, (ID 106495 having an extremely large character length).