EUDETECTOR: Leveraging Language Model to Identify EU-Related News

Koustav Rudra
L3S Research Center, Leibniz University Hannover
Germany
rudra@l3s.de

Danny Tran
Leibniz University Hannover
Germany
danny.tran@stud.uni-hannover.de

Miroslav Shaltev*
Leibniz Centre for Tropical Marine Research (ZMT), Bremen
Germany
miroslav.shaltev@leibniz-zmt.de

ABSTRACT

News media reflects the present state of a country or region to its audiences. Media outlets of a region post different kinds of news for their local and global audiences. In this paper, we focus on Europe (precisely EU) and propose a method to identify news that has an impact on Europe from any aspect such as financial, business, crime, politics, etc. Predicting the location of the news is itself a challenging task. Most of the approaches restrict themselves towards named entities or handcrafted features. In this paper, we try to overcome that limitation i.e., instead of focusing only on the named entities (Europe location, politicians etc.) and some handcrafted rules, we also explore the context of news articles with the help of pre-trained language model BERT. The auto-regressive language model based European news detector shows about 9-19% improvement in terms of F-score over baseline models. Interestingly, we observe that such models automatically capture named entities, their origin, etc; hence, no separate information is required. We also evaluate the role of such entities in the prediction and explore the tokens that BERT really looks at for deciding the news category. Entities such as person, location, organization turn out to be good rationale tokens for the prediction.

CCS CONCEPTS

• **Information systems → Clustering and classification**; • **Applied computing → Media arts**; **Document searching**.

ACM Reference Format:
Koustav Rudra, Danny Tran, and Miroslav Shaltev. 2021. EUDETECTOR: Leveraging Language Model to Identify EU-Related News. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3442442.3452324

1 INTRODUCTION

Thousands of news are posted and consumed by a large group of diverse people across the world. News media try to reflect the present state of a region or a country to its audiences. However not all news posted in a region are focused on it. To work with a particular example, not all the news published in European media are related to Europe[1]. From the subset of Europe related news, not all are of equal importance. Identifying the news that have impact on Europe is crucial to obtain a clear picture of the undergoing socioeconomic processes. For example, a better explanation and representation of Europe to the own citizens and the outer world might be helpful to understand and solve, or at least mitigate, migration related issues which Europe is currently facing.

The identification of news that have an impact on EU across any aspect such as political, economical, social is a timely and important task. The objective of this paper is to develop an automated EU-news detector tool for news media. The problem is different from social media due to unavailability of any other personal information such as user details of a tweet. Side by side, whether a news has any impact on EU is a very subjective decision and to the best of our knowledge, no such datasets are publicly available. We handle both the issues jointly in this paper. First, we annotate 5000 news articles into EU-NEWS and NONEU-NEWS categories (details are given in Section 3). Second, we explore the context of news and power of recent pre-trained language models to achieve superior performance for this task. Standard models typically exploit presence of name entities to determine the relevance of the news to a location, region or country. The crucial point is however, that mere presence of locations does not always reveal the relevance of the news. We believe, that it is necessary to unfold the context of the news, in order to judge its focus and relevance. Recent studies [16] also showed that pre-trained language models basically works as knowledge bases. To demonstrate this, we have developed state-of-the-art BERT [3] based EU news detector (EUDETECTOR) to identify the Europe-related news by exploiting not only the name entities, but also their context of occurrence.

Finally, we inspect input tokens to understand the decision making process of EUDETECTOR. We primarily focus on entities like persons, locations, organizations etc and measure their contribution towards the prediction in terms of comprehensiveness and sufficiency [4]. We observe that such tokens work as rationales in the decision making process of EUDETECTOR and BERT use its pre-trained knowledge to judge the relevance of such tokens towards EU in the given context. Hence, our proposed model not only performs well but also quite interpretable in nature. In addition, we also develop a real-time application service based on our model.

This paper is organized as follows. In Sec. 2 we give an overview of related work. The dataset and method part is described in Sec. 3. The experimental results are present in Sec. 4. Finally, we conclude in Sec. 5.

*The research was conducted while the author was affiliated to L3S Research Center.

[1]Through this paper we use Europe and European Union (EU) interchangeably, as most of the European countries are member of the EU.

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452324

2 RELATED WORK

To the best of our knowledge, none of the existing works focus on EU-specific news detection. However, lots of works tried to detect geolocation of information available in social media. We give a brief overview of such work in this section.

Most of the tasks tried to automatically identify geographic locations present in a text [18]. Typically, standard NLP tools such as Core-NLP [14], Spacy [9] are applied to identify the named entities. The Cliff-Clavin system [1] was developed to identify the location of geotagged web-pages. They found that simple count of the number of occurrences of a location in the text provides a good indication about the geolocation of that news source. In the Profile system [12] and Modercai [7], authors developed an SVM based classifier using named entities and semantic features extracted from the text. They used word2vec models to make the models language agnostic. Recent systems [6, 13] are also developed based on that idea to process multiple languages and large amount of real-time data.

Most of the prior works worked on the assumption that each document contains a single location. On the other hand, some works are explicitly dedicated to identify all possible locations. Chung et al [2] developed a rule based system to identify locations even though it is not explicitly mentioned in the text. Halterman et al [8] developed a CNN-LSTM based network to perform the event-location linking task. A context aware algorithm has been developed in [5] with improvements of street-level geotagging as well of geotagging, even if no place has been mentioned. Some of the recent works have focused on geolocation extraction from Twitter social media [10, 11, 17]. However, they considered user specific features such as user names, screen name, user geolocation tag etc. Huang and Carley [10] used a CNN on tweets to extract linguistic features, like user name, specified location, tweet content, that may be connected to a specific country or city and use those information to classify the tweets to the corresponding location. Rahimi et al [17] used IP addresses of the user to train GCN models.

In contrast, our work is focused on news media and don't have access to any user specific or private information such as IP address. Some of the prior works tried to explore word co-occurrences through word2vec model. None of them had a central focus on EU-news detection. In this paper, we take a step towards that explore the power of recent contextual language models such as BERT [3] to predict the EU-relatedness of a news article. All of the cited related work has in common that it can be used as a building block of a system to identify news of European origin or news pointing geographically to Europe. However none of these methods targets the question of relevance. side by side, deep learning based approaches mostly focus on the task aspect but did not illustrate the role of entities in the prediction. In this paper, we take a step towards that and illustrate the power of BERT to identify entities using its pre-trained knowledge dictionary. We also highlight the role of such entities as rationale tokens through well defined explanation metrics.

3 DATASET AND METHOD

We crawled a set of news articles from the web over a period of one month (November 2019). All of the news are written in English

EU-NEWS:
" Less than a year after the signing of the friendship and cooperation agreement between Bulgaria and North Macedonia we are seeing signs of strain in the relations between the two countries. On 9 June Deputy Prime Minister Krasimir Karakachanov expressed concern, saying that North Macedonia was only using Bulgaria to enter the EU and NATO. [...] On 11 June the Bulgarian side of the commission flatly rejected the Macedonian proposal to honour [...] Gotse Delchev, a historical figure shared by the two countries, on 7 October. [...]"
NONEU-NEWS:
" Joe Biden criticized Sen. Elizabeth Warren's (D-MA) brand of politics on Tuesday as "elitist" after the progressive firebrand suggested he was running in the wrong party's presidential primary. [...] Biden proceeded to argue such tactics were not conducive to getting "any-thing done" or building a party that was capable of beating President Donald Trump in 2020. [...] "

Table 1: Two sample news from two different categories.

language and also come from different regions, since a differentiation of the news articles according to their European relation is the goal of this paper. We are able to collect around 93K news articles. As mentioned in Section 1, our objective is to develop a model to detect EU news. However, to the best of our knowledge, no such public datasets are available for this analysis. Hence, we manually annotate this news into two classes: EU-NEWS and NONEU-NEWS. Due to large amount of news in the original crawled set, we consider first 5000 news for our manual annotation. Our annotation process is as follows:

Annotation setup: Two human annotators independently annotate 5000 news articles into EU-NEWS and NONEU-NEWS categories based on following criteria: (i). **EU-NEWS:** A news is considered as EU-NEWS, if the news covers important information about EU (any of its constituent countries) and has an impact on any important topic of the country such as health, politics, education etc, (ii). **NONEU-NEWS:** A news article that is not related to EU and therefore has minimal or no impact. It may have a few European entities but the main content will not be about EU.

There is no objective measure to define the impact of a news. In this paper, an impact is defined as the amount of relevance or influence of an event, that is described in an article, on EU. We obtain $\kappa = 0.77$, that shows the inter-annotator agreement is significantly high. Table 1 shows examples of both categories of news. The first passage belongs to EU-news since the content is about the relationship between Bulgaria and North Macedonia. Both countries are part of Europe and the whole text rarely takes a different country into consideration. The high impact is given in a sense that the future of Europe may depend on the relationship between the two countries. Therefore the described event in the text is relevant for EU and marked as EU-NEWS. Similarly, the second passage is about the election in the United States. The plot does not take Europe into consideration. In a broader sense, EU may not highly interested in the outcome of the election either, hence the relevance is rather low.

We observe another interesting pattern in the annotation of NONEU-NEWS. More than 67% such news contains atleast one EU entity. For example, a news articles discussed issues about new banking systems and gave reference to an Germany bank. It is more about the banking system itself, so what that is, how the system's infrastructure looks like, what modules the banking system consists of, etc. The Germany entities appear because they also use that

Method	Accuracy	EU-NEWS			NONEU-NEWS		
		Prec.	Recall	F-score	Prec.	Recall	F-score
EU-ENTITY($K = 1$)	0.42	0.25	0.99	0.40	0.99	0.28	0.43
EU-ENTITY($K = 3$)	0.56	0.30	0.95	0.46	0.98	0.47	0.63
EU-ENTITY($K = 6$)	0.82	0.53	0.70	0.60	0.92	0.85	0.89
EU-ENTITY($K = 10$)	0.85	0.66	0.45	0.54	0.88	0.94	0.91
EU-ENTITY($K = 20$)	0.83	0.78	0.16	0.26	0.83	0.99	0.90
BiLSTM-DETECTOR	0.86	0.58	0.73	0.64	0.94	0.89	0.92
EUDETECTOR	0.88	**0.61**	**0.83**	**0.71**	0.96	0.89	0.92

Table 2: Performance of baselines and proposed method for EU-NEWS detection.

system. However, it does not have any impact on Germany. Finally, we have 1662 and 3338 EU-NEWS and NONEU-NEWS respectively.

Model: We follow the BERT single sentence classification architecture for this task. Technically, this is realized by forming an input to BERT of the form $[[CLS], < text >, [SEP]]$ and padding each sequence in a mini-batch to the maximum length (typically 512 tokens) in the batch. The final hidden state corresponding to the $[CLS]$ token captures the high level representation of the entire text. Finally, this 768 dimensional CLS vector is fed to a single layer neural network whose output represents the probability that text is EU-related. Different layers of BERT are responsible for various tasks such as parsing, entity extraction, etc. Recent studies established that pre-trained language models may be used as knowledge bases [16]. Hence, our hypothesis is BERT can learn all the EU related entities from the raw text itself. We highlight this part in our experiment section.

4 EXPERIMENTAL RESULTS

In this section, we report the experimental results for our EU-news detection framework.

4.1 Experimental settings

Training Data Generation: We randomly sample positive and negative instances from the entire set to form training, validation, and test sets. EU-NEWS is distributed to the training, validation, and test in the ratio 60%, 20%, and 20% respectively. Training set contains equal amount of NONEU-NEWS as EU-NEWS. Validation set contains negative samples twice that of positive ones. The rest of the NONEU-NEWS are part of test set. The three sets are disjoint to each other. As the sets are disjoint to each other, it is necessary to sample multiple times or to create different folds to generate a robust model. Each fold contains disjoint train, validation, and test sets. There is $< 50\%$ overlap between the corresponding of sets of any two folds. This is helpful in robust model creation. Finally, we created 15 folds. The order of the instances in each fold will also be shuffled to prevent order effects.

Baselines: We consider a baseline (EU-ENTITY) similar to Cliff-Clavin system [1] i.e., based on the presence/absence of EU-entities. We apply SPACY to identify all the entities of types organizations (ORG), people (PERSON), and countries, cities (GPE). Next we use Wikidata query service to judge the EU-relatedness of these entities. If the text contains $\geq K$ EU-entities, the text is classified as EU-NEWS. This K plays a key role in detection. We also consider bidirectional LSTM version of our model as a baseline where initial embeddings are initialized with pre-trained glove vectors [15] and BiLSTM is used instead of BERT.

Evaluation Metric: We consider overall accuracy and precision, recall, f-score of positive class (EU-NEWS) as our metrics.

Training Details: We train and validate using consistent and common experimental design. The neural model is trained for a fixed number of iterations (20) using binary cross-entropy loss and the accuracy is computed over the validation set to choose the best model. We conduct our experiments on Nvidia 32GB V100 machine. Each fold is treated independently i.e., a separate model is trained and validated for each fold and that model is used to infer the labels for corresponding test set of that fold. Finally, the metrics are averaged over 15 different folds. In ours experiments, for fair comparisons, we use the parameters commonly used in the earlier works, i.e., sequence length of 512, learning rate of $1e-5$, Adam optimizer, and a batch size of 16.

4.2 Performance results

From Table 2, we observe several interesting trends. As expected, if we restrict K to higher values, precision of EU-NEWS gets increased but recall goes down, almost becomes close to 0 at $K = 20$. The baseline model provides good performance at $K = 6$ and performance for EU-NEWS starts dropping drastically after that. On the other hand, our proposed EUDETECTOR obtains better precision, recall, and f-score for EU-NEWS than any value of K. Side by side, the performance for NONEU-NEWS is also comparable to the baselines. This ensures the effectiveness of the context in EU news prediction. BiLSTM model based on the glove word embedding got F-score around 0.64. BERT's pre-trained memory helps to detect the EU-NEWS with a f-score of 0.71. Side by side, BERT based model achieves better performance with less number of iterations (20) than BiLSTM model (200). Prior studies showed that different layers of BERT are involved in different text learning task [19]. In this study, attention weight distribution reveals that BERT itself focuses on specific entities and their associations with the surrounding text. This observation fits in line with the previous findings i.e., pre-trained language models such as BERT can work as knowledgebases [16]. To understand the importance of EU entities in contextual models, we evaluate performance of EUDETECTOR over various representations of original test set.

4.3 Explainability of EUDETECTOR

In the last section, we observe that BERT based EUDETECTOR performs quite well in EU-related news detection. This knowledge

DATA-TYPE	Example
Original-text	Angela Merkel is Germany chancellor. She met US president Donald Trump in last year G7 summit.
TEST-COMP	* * is * chancellor. She met * president * * in last year G7 summit.
TEST-SUF	Angela Merkel * Germany * * * US * Donald Trump * * * * *

Table 3: Format of test data for three different setups (original, comprehensiveness, sufficiency) in EU-NEWS detection.

primarily comes from large corpus over which BERT is pre-trained and EUDETECTOR is able to leverage this knowledge efficiently in accomplishing this task. In order to validate this hypothesis, we try to understand the importance of such tokens in the context of EU news detection. Following the idea of DeYoung et al [4], we measure **comprehensiveness** and **sufficiency** of EUDETECTOR. To measure these metrics, we modify the input as follows:

(1) We apply Spacy to identify different entities present in a news such as 'PERSON', 'LOCATION', etc.
(2) We focus on following named entities 'PERSON', 'NORP', 'LOCATION', 'FAC', 'ORG', 'GPE' and update them to prepare two different variations of the **test data** as described below:
   - **TEST-COMP:** All the above mentioned entities are replaced by a wildcard marker. Here, we have used '*'. This is required to measure *comprehensiveness* that checks the drop in performance in the absence of explanations (decisive tokens in our case).
   - **TEST-SUF:** In this setup, we replace all but the above mentioned entities with the wildcard marker (*). This is required to measure *sufficiency* that checks the drop in performance in the absence of all other tokens except explanations (decisive tokens in our case). Table 3 shows examples of such variations.

Note that, only the test data gets changed but training data is same. Hence, model is trained only on original text but used for inference on three different types of test sets (original, TEST-COMP, TEST-SUF). We measure **comprehensiveness** and **sufficiency** as follows:

1. Comprehensiveness: It is measured as the difference in performance metric between original text and TEST-COMP. For example, comprehensiveness of f-score is $F-score$(Original-text) - $F-score$(TEST-COMP). Ideally, the drop in performance should be huge if EUDETECTOR relies on those entities for prediction.
2. Sufficiency: It is measured as the difference in performance metric between original text and TEST-SUF. For example, sufficiency of f-score is $F-score$(Original-text) - $F-score$(TEST-SUF). The performance gap should be small if model relies mostly on the entities.

The results are reported in Table 4. Accuracy and F-score of NONEU-NEWS remain unaffected due to changes in the test set. Comprehensiveness of EU-NEWS is 0.214 and it indicates that EUDETECTOR primarily focuses on the EU-related entities present in the text. On the other hand, sufficiency value is 0.114. This highlights that EU-entities are primary factor but surrounding context

Metric	Accuracy	F-score(EU)	F-score (NONEU)
Comprehensiveness	0.009	0.214	0
Sufficiency	0.007	0.116	0

Table 4: Comprehensiveness and Sufficiency of EU and NONEU news.

and interaction among those words also play a role in EU-NEWS detection. However, the suppression of EU-tokens leads to major drop in performance. In both the cases, precision of EU-NEWS increases a bit but recall drops to a great extent.
Effect of Normalization: In this part, we try to validate our hypothesis that BERT implicitly identifies related entities and explores their EU connection. First we identify such entities using Spacy and apply Wikidata query service to judge the EU-connections of these entities. Finally, all the EU and non-EU related entities in the text are replaced by the tag 'EU' and 'Non-EU' respectively. For example, the original text in Table 3 is modified as 'EU EU is EU chancellor. She met Non-EU president Non-EU Non-EU in last year G7 summit.'. We observe almost similar performance like original text. For example, F-score of EU-NEWS and NONEU-NEWS are 0.701 and 0.927 respectively. The drop is insignificant. It further validates that BERT uses its large pre-trained knowledge base [20] to identify EU-NEWS.

4.4 Web-service

Finally, we deploy a real-time web-service to detect the EU-news on fly. Along with the prediction, it also highlights all the locations mentioned in the text in the world map. The system is available and running at http://eudetector.l3s.uni-hannover.de. Detection of EU news has applications in many domains. It will be helpful in deciding news editorial policies, handling migration issues in EU, mitigating media bias about EU, etc.

5 CONCLUSION

Detection of EU news has applications in many domains. It will be helpful in deciding news editorial policies, handling migration issues in EU, mitigating media bias about EU, etc. In this paper, we focus on developing a news detection system to capture the EU-related news. We use BERT to not only to identify specific entities but also to capture their relationship and interaction with surrounding words and phrases. Side be side, we create an annotated corpus of 5000 EU news articles. The results demonstrate the implications of this simple yet powerful approach to detect any geo-specific news. Side by side, we have shown that entities like persons, locations, organizations work as rationale tokens for the prediction. However, surrounding context information that may work as rationales is also important [4]. In future, we have a plan to annotate and incorporate such rationale texts into model design to make the model more explainable. We also have a plan to extend this binary model to categorized one so that news can be classified in different levels based on their impact.

Acknowledgement: Funding for this project was in part provided by the European Union's Horizon 2020 research and innovation programme under grant agreement No 832921.

REFERENCES

[1] Rahul Bhargava, Ethan Zuckerman, and Luisa Beck. [n.d.]. CLIFF-CLAVIN: Determining Geographic Focus for News Articles [Extended Abstract].

[2] Jin-Woo Chung, Wonsuk Yang, Jinseon You, and Jong C. Park. [n.d.]. Inferring Implicit Event Locations from Context with Distributional Similarities. In *Proc. IJCAI.*

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018).

[4] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4443–4458.

[5] Md Sadek Ferdous, Soumyadeb Chowdhury, and Joemon M Jose. 2017. Geo-Tagging News Stories Using Contextual Modelling. *Int. J. Inf. Retr. Res.* 7, 4 (2017), 50–71.

[6] Aswin Krishna Gunasekaran, Maryam Bahojb Imani, Latifur Khan, Christan Earl Grant, Patrick T. Brandt, and Jennifer S. Holmes. 2018. SPERG: Scalable Political Event Report Geoparsing in Big Data. *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)* (2018), 187–192.

[7] Andrew Halterman. 2017. Mordecai: Full Text Geoparsing and Event Geocoding. *The Journal of Open Source Software* 2, 9 (2017).

[8] Andrew Halterman. 2019. Geolocating Political Events in Text. *ArXiv* abs/1905.12713 (2019).

[9] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[10] Binxuan Huang and Kathleen M. Carley. 2017. On Predicting Geolocation of Tweets using Convolutional Neural Networks. *CoRR* abs/1704.05146 (2017).

[11] Chieh-Yang Huang, Hanghang Tong, Jingrui He, and Ross Maciejewski. 2019. Location Prediction for Tweets. *Frontiers in Big Data* 2 (2019), 5.

[12] M. B. Imani, S. Chandra, S. Ma, L. Khan, and B. Thuraisingham. 2017. Focus location extraction from political news reports with bias correction. In *2017 IEEE International Conference on Big Data (Big Data)*. 1956–1964.

[13] Maryam Bahojb Imani, Latifur Khan, and Bhavani Thuraisingham. 2019. Where Did the Political News Event Happen? Primary Focus Location Extraction in Different Languages. *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)* (2019), 61–70.

[14] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. ACL*. 55–60.

[15] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[16] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?. In *Proc. EMNLP*.

[17] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. Semi-supervised User Geolocation via Graph Convolutional Networks. *CoRR* abs/1804.08049 (2018).

[18] Christos T. Rodosthenous and Loizos Michael. 2018. GeoMantis: Inferring the Geographic Focus of Text using Knowledge Bases. In *ICAART*.

[19] Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2019. How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations. In *Proc. CIKM*. 1823–1832.

[20] Jonas Wallat, Jaspreet Singh, and Avishek Anand. 2020. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. arXiv:cs.CL/2010.09313