Multilingual Disinformation Detection for Digital Advertising

Zofia Trstanova, Nadir El Manouzi, Maryline Chen, Andre L V da Cunha, Sergei Ivanov

Multilingual Disinformation Detection for Digital Advertising

Žofia Trsťanová [†] **Nadir El Manouzi** [†] **Maryline Chen** [†] **Andre L. V. da Cunha** [†] **Sergei Ivanov** [†]

Abstract

In today's world, the presence of online disinformation and propaganda is more widespread than ever. Independent publishers are funded mostly via digital advertising, which is unfortunately also the case for those publishing disinformation content. The question of how to remove such publishers from advertising inventory has long been ignored, despite the negative impact on the open internet. In this work, we make the first step towards quickly detecting and red-flagging websites that potentially manipulate the public with disinformation. We build a machine learning model based on multilingual text embeddings that first determines whether the page mentions a topic of interest, then estimates the likelihood of the content being malicious, creating a shortlist of publishers that will be reviewed by human experts. Our system empowers internal teams to proactively, rather than defensively, blacklist unsafe content, thus protecting the reputation of the advertisement provider.

1. Introduction

In recent years, traditional news outlets such as newspapers, magazines, and television are being replaced as main sources of information in favor of social media, podcasts, messaging applications, and online websites (Pew Research Center, 2022; 2021a; 2018; Newman et al., 2021). This shift to digital news consumption has been accompanied by an online environment where news pieces from serious institutions increasingly compete for the public's attention against disinformation, fake news, and propagandistic content. As such, the prevalence of untruthful and misleading online content has become a source of concern for governments and other official institutions. Recent events where online disinformation played a crucial role include the COVID19 pandemic (Khan et al., 2022; Endo et al., 2022) and the 2016 US elections (Zhou et al., 2020a; 2019; Wang et al., 2018; Farajtabar et al., 2017). It also led to a significant economic impact (Rapoza, 2017). As companies shift their advertisement budget to online campaigns (Pew Research Center, 2021b), the preponderance of online disinformation has become a source of concern in the ad tech industry, where advertisers[1] and providers do not want to associate themselves with or help fund such publishers[2].

In this work, we propose a pipeline whereby ad providers can detect the presence of undesired content among their inventory and proactively block it. As there are billions of websites, the first step in our solution is to filter the web pages on a topic that can be harmful to the ad providers. We then apply a classifier that predicts the probability of each publisher being malicious and uses it as a ranking score. Finally, trained practitioners inspect the top-ranked publishers to decide if they should be blocklisted or not. Our procedure is intended to assist internal teams in assessing the quality of the publisher's content on a sensitive topic proactively, rather than in a post-hoc manner.

The remainder of this paper is organized as follows: Section 2 presents related work; in Section 3 we state the main problem and describe our methodology and define our solution, the REDD model; in Section 4 we perform experiments to address the following questions: Why did we choose fine-tuned embeddings for topic projection? Is the topic projection step necessary to train the disinformation ranking model? What should be used as the input to that model: embeddings or text? How does the multilingual setting impact the performance? In the final section, we highlight the main conclusions and discuss the next steps.

2. Related Work

Previous literature on disinformation detection focused mostly on social media content. Among existing methods, *propagation-based* and *content-based* are two commonly used techniques. Propagation-based methods analyze the

[†]Criteo AI Lab, Paris, France. Correspondence to: Žofia Trsťanová <zofia.trstanova@gmail.com>, Nadir El Manouzi <n.elmanouzi@criteo.com>, Maryline Chen <ma.chen@criteo.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]An **advertiser** is a company who pays for the possibility of showing ads for its products on the internet.

[2]A **publisher** is a website that receives revenue by accepting to display ads on its pages.

way the content circulates in the social media platform: who produced it, who spread it, and how producers relate to each other (Zhou et al., 2020a; Zhou & Zafarani, 2018; Wu et al., 2015; Castillo et al., 2011). Content-based methods attempt to identify disinformation by analyzing its textual and image contents and can generally be grouped as combinatorial (Pérez-Rosas et al., 2017; Shi & Weninger, 2016; Ciampaglia et al., 2015) or neural (Zhou et al., 2020b; Wang et al., 2018). Combinatorial methods explicitly represent textual aspects known to be indicative of suspicious content and usually rely on extensive feature engineering (Endo et al., 2022; Horák et al., 2021; Zhou et al., 2020a; Wang, 2017). Neural methods, on the other hand, learn representations with neural networks based on the raw text or image content. Architectures explored in the literature include LSTM, GRU, Bidirectional GRU (Endo et al., 2022), CNN, Bidirectional LSTM (Wang, 2017), Text-CNN (Zhou et al., 2020b; Wang et al., 2018), and BERT (Vorakitphan et al., 2022). Following trends in other Natural Language Processing tasks, neural networks typically outperform the combination of feature engineering and standard classification algorithms, while being easier to train.

Alternatively, some works have focused on detecting propaganda in news articles. Vorakitphan et al. 2022 employ a BERT-based model for spam identification and use RoBERTa-based model to extract sentence embeddings, which are combined with a series of hand-crafted features (Vorakitphan et al., 2021) and fed into a Bi-LSTM model. Blaschke et al. 2020 adopt a similar approach, combining BERT embeddings and hand-crafted features and feeding them into Bi-LSTM, multi-layer perception, and linear models, but using a more reduced set of features. Dao et al. 2020 use an LSTM model on top of GloVe embeddings for span identification, and another LSTM model over BERT embeddings for classification. Da San Martino et al. 2020 perform a literature review on computational propaganda detection, confirming that the most performant models are BERT-based. A work similar to ours is that of Chang et al. 2021, which selected a list of Twitter accounts with potential propaganda, obtained tweets from them, and manually labeled them according to 21 labels comprising 18 propaganda techniques. The authors then trained a BERT model for multi-class sentence classification. In our work, we adopt a similar, but simpler output: a binary classification, without any further categorization or detection of relevant text spans, and we focus on multi-language publisher content, which is different in nature from user-generated social media content. Also, we use specific embeddings, which were fine-tuned for a multi-label classification task.

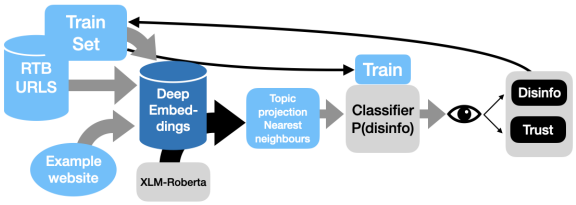


Figure 1. Disinformation prediction through content embeddings and human-in-the-loop approach.

3. Embedding-Based Disinformation Detection

Our main task is to rank domains with respect to their probability of containing disinformation on a certain topic of interest. Our solution has three steps: first, we filter a dataset of publisher pages, keeping only those related to the topic of interest; next, we apply a classifier to predict the probability of the content being disinformation; finally, we aggregate the page-level scores at domain-level, and a human reviewer manually inspects the most suspicious domains to confirm the label. In order to filter our dataset of pages on a particular topic, we rely on an embedding-based representation of web pages. The disinformation classifier is trained in a supervised manner. In order to rank the domains, we compute an average page score per domain. After human revision, the newly identified domains can be added to the training set and the model, refined (see Figure 1).

3.1. Topic Detector

The first step in our disinformation detection pipeline is filtering on a specific topic. We developed a zero-shot topic classifier that leverages similarity scoring between multilingual page embeddings. These are extracted from an XLM-RoBERTa model (Conneau et al., 2019) that had previously been fine-tuned for an unrelated multi-label text classification task. The topic classifier is topic-agnostic and does not require labeled data for our topic classification task. We determine whether a page talks about the topic by applying a manually defined threshold on the cosine similarity between the topic embedding and the page embedding.

In this section, we explain how we built the embeddings (for a given topic and an input web page), how to select the minimum threshold on the similarity scores, and finally how these two pieces can be applied together to build the topic classifier.

3.1.1. Embedding Prediction

We use the base version of XLM-RoBERTa (henceforth XLM-R), a multilingual language model based on the Transformer architecture (Vaswani et al., 2017), to compute the embedding by running the model inference on a part of the

text crawled from the web page (see Figure 5). We leverage a pre-trained XLM-R model that we have fine-tuned on a multi-label classification task for categorizing publisher web pages (see 3.2.2 for more information about the task and datasets). The quantized version[3] of this fine-tuned model is running at scale in our production environment. Since the model has been trained on annotated datasets with text extracted from web pages, it has learned some domain-specific knowledge of the content. The text is the concatenation of the content in the HTML tags [4]. This input is tokenized and truncated at 96 tokens. We extract the page embedding by taking the hidden state of the `<cls>` token in the last layer of the encoder, which gives us an embedding of dimension 768. We further reduce the dimension to 100 by Gaussian random projection, which does not impact the quality of the embedding (see section 4.1).

3.1.2. Topic Projection

To filter the pages on a particular topic, we start by selecting a few pages to serve as examples, getting their corresponding embeddings and computing the average to get the topic embedding. We then compute the similarity score of the page with respect to the topic via the cosine similarity between the page and topic embeddings. Finally, we rank all pages in an unlabeled dataset by decreasing similarity score and, through manual inspection, select a threshold (see Figure 6; note that, due to the heterogeneity of the embedding space, this threshold is topic-specific, so this process must be repeated for each new topic). The manual inspection process starts with a list of eligible thresholds, which is used to create buckets of similarity scores. For each bucket, the user visualizes a sample of web pages that belong to it. The optimal threshold is selected based on the relevance of the web pages sampled in the corresponding bucket: the selected bucket is the farthest one from 1 such that the majority of the pages in the bucket are relevant to the topic.

3.2. Disinformation Classification Model

Although topic filtering reduces the number of pages to be reviewed from billions to thousands, this is still a very high number to be reviewed by a human team within a reasonable time. To further narrow the search, we propose to train a disinformation classifier and rank the pages according to the classifier's output score. The reviewers can then select the top domains for inspection.

[3]We used the dynamic quantization provided by PyTorch. This method, which represents the model weights and activations as integer rather than floating point, has proven to be very useful for running deep models at scale: the model inference time decreases significantly at the cost of a small drop in performance.

[4]Tags are ordered by decreasing order of importance: title > description > h1 > h2 > ... > h6 > p.

3.2.1. Repurposed Embeddings for Disinformation Detection (REDD)

In order to predict the probability of disinformation, we introduce the Repurposed Embeddings for Disinformation Detection (REDD) model. This model takes as input the page embeddings and feeds them to a simple three-layer nonlinear classifier with scaled exponential linear units (SELUs). We train the model for several epochs on the topic-filtered dataset using a binary cross-entropy loss function and predict the scores per page.

Note that there are, in fact, many other options for the architecture or the classifier itself. Preliminary experimentation suggested that there is not much difference among these setups, since the classification task seems to be picked up easily, so we did not pursue this direction (other options would be to also compare against kNN, SVM, etc). In Section 4.3 we compare the nonlinear REDD with a one layer MLP.

3.2.2. Dataset

In order to train REDD, we use 60 domains already blocked for spreading disinformation, obtained from external and internal providers. Pages (URLs) from these domains serve as the disinformation examples and are associated with the positive label. On the other side, we manually select renowned media with a good reputation in fact-checking to get web pages with a negative label. The train dataset size is 17,473 samples and the test set size is 1,925. The sets contain 53% of positive (disinformation) examples. Both datasets have already been projected on a particular topic. The articles are in various languages, with some languages prevailing over others. This happens because disinformation in the topic of interest might be more prevalent in some languages than in others[5]. This causes our dataset to be skewed by language, making the multilingual generalization challenging: there is a risk that the model might simply become a language detector, which we aim to prevent (see Section 4.3).

4. Experiments

We first evaluate the quality of fine-tuned embeddings, which will be used for the topic projection and as the input to the REDD model. We compare it with embeddings extracted from the general-purpose NLP model. In the next experiment, we demonstrate the importance of the topic projection step by comparing against a model trained on a dataset that has not been filtered on the selected topic. Finally, we compare the performance of our REDD model, which takes embeddings as input, against a classifier trained

[5]This problem of modeling the sources rather than the content has already been encountered in the literature. See, e.g., Da San Martino et al. 2020.

directly on the text and discuss its impact on the multilingual task.

4.1. Repurposed Topic Embeddings

In order to evaluate the quality of the embeddings, we employ a metric that measures how web pages with similar content are close in the embedding space. The goal of these experiments is to compare the performance of our embeddings with embeddings coming from pre-trained open-source models and also study the impact of dimensionality reduction on the quality of the embeddings.

The datasets for the embedding quality evaluation consist of web pages and their associated categories. These categories have been manually selected by humans for web pages in 13 languages and belong to the IAB content taxonomy[6]. Details on the 13 datasets, their language and number of web pages are shown in Table 4.

The **Probability of Same Categories** score (pSameCat) measures the average probability that the nearest neighbors of a web page share its categories. The metric assumes that web pages that have their embeddings close in the vector space should share related topics (represented by their web page categories). In our case, for each web page, we search for the 5 nearest neighbors and compute the proportion that shares at least one category with the page. The metric for the dataset is the average of these individual scores.

For our 13 datasets, we evaluated the pSameCat scores of the embeddings extracted from four models. We also evaluate this metric for these embeddings after a dimension reduction to 100 dimensions using Gaussian random projection. The first model is our XLM-R model, fine-tuned for the multi-label classification task on our internal annotated data. The three other models are open-source pre-trained multilingual Transformer-based models: mBERT (multilingual BERT; Devlin et al. 2018), distilmBERT (a distilled version of mBERT; Sanh et al. 2019), and XLM-R (not fined-tuned on our data).

The pSameCat scores for all datasets and model embeddings are depicted in Table 1. For both raw and dimension-reduced embeddings, the fine-tuned XLM-R model significantly outperforms all the pre-trained models across all languages. This shows that fine-tuning moved web pages that share similar content and topics closer to each other in the embedding space. Moreover, after we apply dimensionality reduction on these embeddings, their quality remains high with only a small drop of 1.3% in average pSameCat score, a good signal compression. We chose therefore to use the dimension-reduced embeddings in our system.

[6]We use the v2.2 version of https://iabtechlab.com/standards/content-taxonomy.

4.2. The Necessity of the Topic Projection Step

As we mentioned before, our method consists of three steps: topic projection, prediction of the probability of disinformation, and human review. In order to justify the necessity of topic projection to build the disinformation ranking model, we trained an end-to-end model on the unfiltered dataset, i.e. on the dataset without the topic projection. We compare this setup against training the model on the topic-filtered dataset, described in Section 3.2.2. Without the topic projection, the model achieves an AUC-ROC on the (topic-projected) test set of 0.65 compared to 0.955 (see Table 2). This shows that the topic filtering step is crucial to be able to capture the particular disinformation content.

We are aware that having to build a specific model per topic bears the disadvantage of not generalizing to other topics. However, this is not an issue for our particular application, since our production pipelines allow for such design, and the review process is intended to be iterative.

4.3. Embeddings versus Text as the Model Input

In this experiment, we compare training the classifier described in Section 3.2.1 on the raw text instead of the embeddings. The training and test datasets are described in Section 3.2.2: they contain the text of these articles and the fine-tuned embeddings from XLM-R. The results are reported in Table 2. Surprisingly, the XLM-R model trained on textual content achieved an AUC-ROC of 1 on the test set. After examining the predictions, we found that the model learned to detect a particular language instead of any disinformation topic. That is: it predicts high scores if a page is in one particular language, in which the disinformation topic is more prevalent, while consistently assigning low scores to any page in any other language. We show the predicted scores per category and language in Table 3. This issue is expected for our disinformation task, and finding a workaround is not trivial[7]. Note that our dataset has a strong language bias, so the evaluation of this issue is particularly hard: in the language where disinformation occurs most often, there are almost no negative examples.

However, for the REDD setup, where we take the fine-tuned embeddings as the input, the model focuses on the content instead of the language and is able to separate the disinformation content. We believe that the reason behind this is that the embeddings are already in a multilingual space, hence the training for disinformation prediction is language-agnostic. When using 100-dimensional (fixed) embeddings as the input to the nonlinear classifier, we obtain an AUC-ROC of 0.955, but the predicted probabilities are more equally distributed across the languages, see Table 3

[7]One could try translating the whole training set in a single language or ensure that it contains a broad mixture of languages.

Table 1. pSameCat × 100, for datasets in several languages. Higher is better.

Model embeddings	Dim	EN	FR	DE	JA	IT	ES	PT	TR	NL	AR	RU	KO	ZH	Avg
pre-trained distilmBERT	768	43.2	49.3	46.7	36.9	49.3	49.7	45.5	47.0	47.0	48.5	47.1	40.6	57.6	46.8
pre-trained mBERT	768	29.9	36.9	32.7	32.2	36.3	39.2	34.8	43.4	32.9	42.3	37.2	41.7	48.8	37.6
pre-trained XLM-R	768	23.0	32.3	30.6	29.9	27.9	32.9	27.9	39.8	32.4	37.4	35.0	40.5	44.5	33.4
fine-tuned XLM-R	768	**75.4**	**75.6**	**75.2**	**59.8**	**70.9**	**73.7**	**69.8**	**57.0**	**74.1**	**61.4**	**73.0**	**67.0**	**75.9**	**69.9**
pre-trained distilmBERT	100	36.6	43.8	41.2	33.0	44.4	44.5	40.2	44.3	42.4	44.6	41.5	36.0	53.5	42.0
pre-trained mBERT	100	23.6	29.6	26.1	25.8	28.3	31.8	26.7	39.6	26.9	37.8	28.7	35.8	40.8	30.9
pre-trained XLM-R	100	16.6	26.4	22.7	23.3	22.4	27.7	22.6	33.9	26.1	32.9	27.5	34.6	36.2	27.1
fine-tuned XLM-R	100	**74.7**	**74.7**	**74.5**	**58.9**	**70.0**	**72.8**	**68.9**	**56.1**	**72.6**	**60.6**	**72.3**	**65.9**	**75.1**	**69.0**

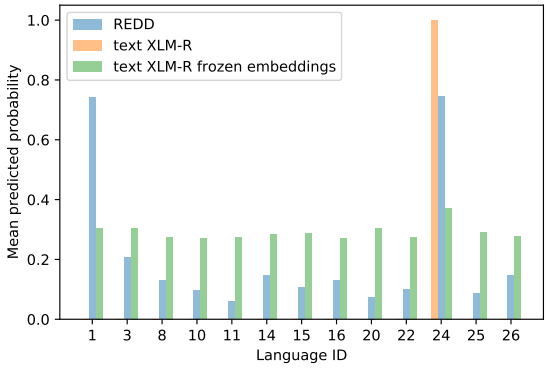
Table 2. Comparison of embeddings (REDD) versus text as input for the binary classifier.

Model	AUC-ROC	Loss	Precision@50	Precision@500
REDD linear	0.868	0.405	0.96	0.752
REDD non-linear	**0.955**	**0.251**	**0.98**	**0.922**
REDD no topic filter	0.65	2.431	0.64	0.57
Text fully trained	1	0.493	1	1
Text embeddings frozen	1	0.675	0.98	0.998

Table 3. Comparison of three models on language ID 24 versus the other languages.

		Trustworthy		Disinformation	
		24	*not 24*	*24*	*not 24*
REDD	mean	0.731	0.115	0.834	0.514
	std	0.278	0.196	0.194	0.422
Text XLM-R Fully Trained	mean	0.999	0.005	0.999	0.333
	std	5e-08	0.061	3e-08	0.516
Text XLM-R Embeddings Frozen	mean	0.355	0.277	0.367	0.298
	std	0.009	0.011	0.01	0.024

Figure 2. Distribution of predicted disinformation scores across various languages. The full-text model only learns to distinguish one language from others (orange and green), while the embedding-based model actually trains for the disinformation task.



and Figure 2.

The main results are presented in Table 2, where we summarize the comparison of the following models: *REDD linear*: a linear model taking the embeddings as input; *REDD non-linear*: a model with three non-linear (SELU) layers, taking the embeddings as input; *REDD no topic filter*: model trained on the unfiltered dataset and evaluated on the filtered test set; *text fully trained*: XLM-R with a classification head, fine-tuned end-to-end on the text of the pages; and *text embeddinds frozen*: XLM-R with a classification head, fine-tuned on the text of the pages, but with the embeddings kept frozen during the training. We observe that the non-linear version of REDD outperforms the linear version and that the model with topic filtering overperforms the model without one. The seemingly good performance of the text-based model over the embedding-based model is actually due to the former only learning to detect one particular language.

4.4. Human Review

The human review is done on the domain level: the domain score is computed as the average score of the pages in the domain. Ranking the domains by the highest scores helps to reduce the number of pages that need to be reviewed. We evaluated REDD on a set of articles prefiltered on the topic of interest. The predicted disinformation scores were aggregated to obtain a score per domain. This corresponded to 4000 ranked domains, out of which we sent the top 300 for human review. Out of these 300, 178 were flagged as suspected of spreading propaganda, and 26 were directly blocklisted from the publisher inventory[8]. In Figure 3, we show the distribution of the ranking of the blocked domains: these domains are more likely to be ranked higher by REDD among the top 300 reviewed domains. The evaluated precision at k with $k = 40$ equals 0.282, significantly above the random baseline at $26/300 = 0.086$.

Even though the number of blocked publishers might seem low, these were newly discovered domains, previously uncaught by external providers. This shows that the flexibility of our approach allows for adapting to industry-specific requirements. Finally, since our model works across many languages, we have identified disinformation domains on the same topic across various countries and languages.

[8]Our internal review process has several stages, where the whole content of the publisher is considered by the expert team.

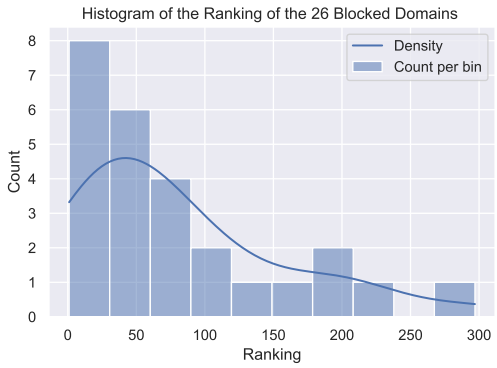


Figure 3. The distribution of the ranking position of the 26 identified disinformation domains among the 300 submitted for human review: the finally blocked domains are ranked higher by REDD.

5. Discussion and Future Work

In this work, we tackled the problem of disinformation identification by building REDD, a tool that red-flags domains that potentially spread disinformation. Even though the obtained probability predictions are not reliable for an automated decision, the tool allowed us to identify several domains not previously blocklisted. It can also support different languages and topics of interest, potentially preventing the spread of disinformation across different industries.

We have shown that filtering the training data on the topic of interest is necessary to obtain satisfactory results. Moreover, we demonstrated that leveraging fine-tuned embeddings helps the multilingual model focus on the disinformation task instead of on the language.

Our approach can be considered as the first step in improving publisher content analysis for digital advertisement in a multilingual setting. In the future, we can leverage orthogonal sources of information such as images, employ adversarial training (which has been suggested in the literature to improve performance: see Wang et al. 2018), or factorize a user-based disinformation graph to obtain embeddings and concatenate them with the content-based embeddings.

6. Acknowledgements

We would like to thank Béranger Dumont, Nicolas Pennequin, Thibault Becker, Kamila Jańczyk and François Zolezzi for their useful comments and contributions to this project.

References

Blaschke, V., Korniyenko, M., and Tureski, S. Cyberwalle at semeval-2020 task 11: An analysis of feature engineering for ensemble models for propaganda detection, 2020. URL https://arxiv.org/abs/2008.09859. (Cited on 2)

Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pp. 675–684, 2011. (Cited on 2)

Chang, R.-C., Lai, C.-M., Chang, K.-L., and Lin, C.-H. Dataset of propaganda techniques of the state-sponsored information operation of the people's republic of china, 2021. URL https://arxiv.org/abs/2106.07544. (Cited on 2)

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193, 2015. (Cited on 2)

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale, 2019. URL https://arxiv.org/abs/1911.02116. (Cited on 2)

Da San Martino, G., Cresci, S., Barron-Cedeno, A., Yu, S., Di Pietro, R., and Nakov, P. A survey on computational propaganda detection. 2020. doi: 10.48550/ARXIV.2007.08024. URL https://arxiv.org/abs/2007.08024. (Cited on 2, 3)

Dao, J., Wang, J., and Zhang, X. Ynu-hpcc at semeval-2020 task 11: Lstm network for detection of propaganda techniques in news articles, 2020. URL https://arxiv.org/abs/2008.10166. (Cited on 2)

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805. (Cited on 4)

Endo, P. T., Santos, G. L., de Lima Xavier, M. E., Nascimento Campos, G. R., de Lima, L. C., Silva, I., Egli, A., and Lynn, T. Illusion of truth: Analysing and classifying covid-19 fake news in brazilian portuguese language. *Big Data and Cognitive Computing*, 6(2), 2022. ISSN 2504-2289. doi: 10.3390/bdcc6020036. URL https://www.mdpi.com/2504-2289/6/2/36. (Cited on 1, 2)

Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., and Zha, H. Fake news mitigation via point process based intervention, 2017. URL https://arxiv.org/abs/1703.07823. (Cited on 1)

Horák, A., Baisa, V., and Herman, O. Technological approaches to detecting online disinformation and manipulation. 2021. doi: 10.48550/arXiv.2108.11669. URL https://arxiv.org/abs/2108.11669. (Cited on 2)

Khan, S., Hakak, S., Deepa, N., Prabadevi, B., Dev, K., and Trelova, S. Detecting covid-19-related fake news using feature extraction. *Frontiers in Public Health*, 9, 2022. ISSN 2296-2565. doi: 10.3389/fpubh.2021.788074. URL https://www.frontiersin.org/article/10.3389/fpubh.2021.788074. (Cited on 1)

Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C. T., and Nielsen, R. K. Reuters institute digital news report 2021, 10th edition, 2021. (Cited on 1)

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017. (Cited on 2)

Pew Research Center. Social media outpaces print newspapers in the u.s. as a news source, 2018. https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/, accessed 2022-05-17. (Cited on 1)

Pew Research Center. News on twitter: Consumed by most users and trusted by many, 2021a. https://www.pewresearch.org/journalism/2021/11/15/news-on-twitter-consumed-by-most-users-and-trusted-by-many/, accessed 2022-05-17. (Cited on 1)

Pew Research Center. Digital news fact sheet, 2021b. https://www.pewresearch.org/journalism/fact-sheet/digital-news/, accessed 2022-05-17. (Cited on 1)

Pew Research Center. Nearly a quarter of americans get news from podcasts, 2022. https://www.pewresearch.org/fact-tank/2022/02/15/nearly-a-quarter-of-americans-get-news-from-podcasts/, accessed 2022-05-17. (Cited on 1)

Rapoza, K. Can 'fake news' impact the stock market?, 2017. https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/, accessed 2022-05-17. (Cited on 1)

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019. URL https://arxiv.org/abs/1910.01108. (Cited on 4)

Shi, B. and Weninger, T. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-based systems*, 104:123–133, 2016. (Cited on 2)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762. (Cited on 2)

Vorakitphan, V., Cabrio, E., and Villata, S. "Don't discuss": Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification. In *RANLP 2021 - Recent Advances in Natural Language Processing*, Varna / Virtual, Bulgaria, September 2021. URL https://hal.archives-ouvertes.fr/hal-03314797. (Cited on 2)

Vorakitphan, V., Cabrio, E., and Villata, S. Protect: A pipeline for propaganda detection and classification. In *CLiC-it 2021- Italian Conference on Computational Linguistics*, Milan, Italy, January 2022. URL https://hal.archives-ouvertes.fr/hal-03417019. (Cited on 2)

Wang, W. Y. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL https://aclanthology.org/P17-2067. (Cited on 2)

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery*, KDD '18, pp. 849–857, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219903. URL https://doi.org/10.1145/3219819.3219903. (Cited on 1, 2, 6)

Wu, K., Yang, S., and Zhu, K. Q. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pp. 651–662. IEEE, 2015. (Cited on 2)

Zhou, X. and Zafarani, R. Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2, 2018. (Cited on 2)

Zhou, X., Zafarani, R., Shu, K., and Liu, H. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pp. 836–837, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359405.

doi: 10.1145/3289600.3291382. URL https://doi.org/10.1145/3289600.3291382. (Cited on 1)

Zhou, X., Jain, A., Phoha, V. V., and Zafarani, R. Fake news early detection: A theory-driven model. *Digital Threats*, 1(2), jun 2020a. ISSN 2692-1626. doi: 10.1145/3377478. URL https://doi.org/10.1145/3377478. (Cited on 1, 2)

Zhou, X., Wu, J., and Zafarani, R. Safe: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 354–367. Springer, 2020b. doi: 10.48550/ARXIV.2003.04981. URL https://arxiv.org/abs/2003.04981. (Cited on 2)

A. Model architectures

A.1. Three step reduction of number of pages for review

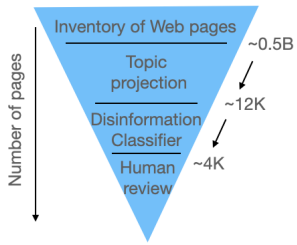


Figure 4. Topic projection and ranking of the pages reduces the number of pages that has to be reviewed by human reviewer.

A.2. Embedding prediction schema

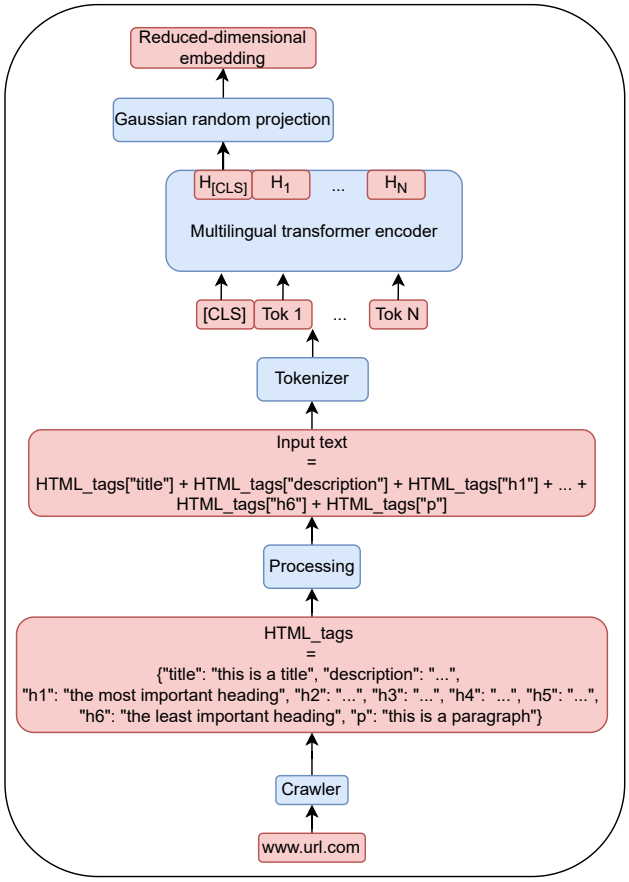


Figure 5. A detailed embedding prediction schema.

A.3. Topic classification schema

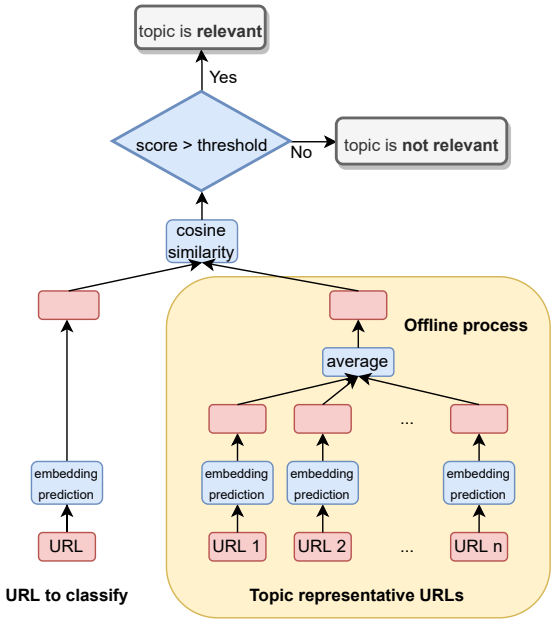


Figure 6. The topic classification schema.

B. Additional details

Table 4. Datasets with human-annotated web page categories. These datasets were used to evaluate the quality of the embeddings using the pSameCat metric. Some examples of categories are: News and Politics > Politics, Technology and Computing > Consumer Electronics > Smartphones, and Travel > Travel Type > Air Travel.

Language		Size
EN	English	53,290
FR	French	7,131
DE	German	5,276
JA	Japanese	7,772
IT	Italian	11,569
ES	Spanish	11,773
PT	Portuguese	9,081
TR	Turkish	11,409
NL	Dutch	6,921
AR	Arabic	6,302
RU	Russian	7,985
KO	Korean	7,304
ZH	Chinese	8,710