Toxicity-Associated News Classification: The Impact of Metadata and Content Features

Paula Fortuna[1], Luís B. Cruz[2], Rodrigo Maia[1,3], Vanessa Cortez[1] and Sérgio Nunes[1,2]
[1]INESC TEC
[2]FEUP and [3]FCUP, University of Porto
R. Dr. Roberto Frias s/n, Porto, Portugal
paula.fortuna@upf.edu, sergio.nunes@fe.up.pt

Abstract

In this work, we study toxicity-associated news, which we define as news with a high percentage of toxic comments. Little research exists in this topic. We address two open questions: (i) Is the comments' toxicity related to the textual content of the news pieces?, and (ii) Are there other contextual factors of the news (i.e. metadata) interfering with the presence of toxicity in its comments? To answer, we annotate 1,995,560 Twitter messages for toxicity, which are replies to 29,726 news pieces from 25 newspapers from the UK and USA. We experiment with content and metadata features and use both classical machine learning and BERT classifiers. We found that metadata features have the best performance when used to train a GDBT classifier (F1 = 0.723). This was the case even when comparing with BERT. Additionally, we contribute to future studies of toxicity-associated news by providing an annotated dataset to the community. With this resource, it is possible to further investigate the effect of other content and metadata-based features to identify toxicity-associated news.

1 Introduction

Journalism has evolved in the last years regarding both news production and news consumption. The internet drastically changed the media ecosystem and social networks platforms currently have a determinant role in news production — e.g., social networks serve as platforms for collecting information or for reporting last-minute events. Furthermore, online reactions (e.g., comments, likes, shares) serve as indicators of popularity and influence the writing and presentation of news. While paper and television declined as a means for news consumption, social networks not only increased but also changed the way people interact with news media. Social networks are permanent forum debates for news (Orellana-Rodriguez and Keane 2018) as they facilitate the communication between the public and news providers. However, online participation is not always civilized. A 2014 Pew Report notes that 73% of adult internet users have seen someone being harassed online, and 40% have personally experienced it (Duggan 2014). In this context, researchers have turned their efforts to the identification of the various forms of negative online communication. Nonetheless, little research has been conducted in understanding the link between news and the content of the comments associated to them, in particular when this content is toxic.

In this work, we aim at studying *toxicity-associated news*, which we define as news with a high percentage of associated toxic comments. We review previous work on this topic for English (Magu, Hossain, and Kautz 2018; Zannettou et al. 2020), German (Daxenberger et al. 2018) and Turkish (Berk and Filatova 2019). To the best of our knowledge, our work is the first that automatically classifies toxicity-associated news of an extensive set of renewed newspapers from the UK and USA, using both content and metadata features. Previous works using classification focus exclusively on content-based approaches. Furthermore, we also innovate by applying BERT classifiers to this problem.

In the next section, we review the related work on toxicity-associated news. Section 3 outlines the procedure for the dataset collection and annotation. Sections 4 and 5 describe our classical machine learning and BERT experiments, respectively. Section 6 presents and discusses the results of the experiments and, finally, Section 7 draws some conclusions and outlines possible future work.

2 Related work

Previous studies focus on triggers for toxicity in online conversations. For instance, Almerekhi et al. (2020) define toxicity triggers in online discussions as a non-toxic comment that lead to toxic replies. They analyse Reddit conversations and conclude that triggers of toxicity contain identifiable features and that is possible to detect them with a ROC-AUC score of 0.87. In our work, we hypothesize that if comments can serve as triggers for toxicity in online discussions, so does the news' content and context.

With the goal to investigate the work done in the area of toxicity-associated news classification, we start by describing and distinguishing the variety of terms often addressed when studying this phenomenon.

2.1 Concepts

The relation between news articles and toxic comments has been studied in the past. However, authors have been

using slightly different terminology: Magu, Hossain, and Kautz (2018) look at *uncivil speech provocation news*, Daxenberger et al. (2018) also address the phenomena of *incivility*, while Berk and Filatova (2019) frame the problem as *incendiary news*. Regarding online incivility, Anderson et al. (2014) "*[consider] online incivility to be a manner of offensive discussion that impedes the democratic ideal of deliberation, ranging from unrelated, rude critiques and name-calling to outrageous claims and incensed discussion, which is also known as flaming*" and Magu, Hossain, and Kautz use the definition of incivility by Anderson et al.. On the other hand, *incendiary news* correspond to articles that ignite hatred (Berk and Filatova 2019). In contrast to social media posts, incendiary news articles often do not contain any explicit slurs and insults but incite hate.

Instead of the aforementioned definitions, we opt to use the *toxicity-associated news* concept. Regarding toxicity, we use the definition adopted by the *Perspective API*[1]: any communication having violent, rude, and disrespectful behavior, that can make use of personal attacks, harassment, or bullying to cause any participant to leave the conversation (Georgakopoulos et al. 2018). We do so as we classify our comments by using this API. Furthermore, we prefer to be careful when implying a causal relationship between news and toxicity in its comments, as there is no clear evidence that it is the essence of the news article that causes such behavior. We opted then for 'association' instead of 'provocation' or 'incendiary' news.

2.2 Previous Approaches Related with Toxicity-Associated News Classification

In this section, we present the few studies that relate news with toxic comments. We divide previous approaches into the ones that use news classification and the ones that do not.

Previous classification works have focused on using text mining over content to study news associated with toxic comments. One example is the work done by Magu, Hossain, and Kautz (2018). In order to achieve their goals, the authors build a dataset of 15,000 English political news articles from the conservative news website Breitbart and the liberal website Politico. Then, to classify news as provocative of incivility, the authors also gather readers' comments on those news and classify them for incivility with a model trained with Wikipedia data. Each article is measured for how much it provokes uncivil speech. This is done by computing the percentage of uncivil comments answering to the article. Finally, news articles are classified as provocative of incivility: an article with a frequency of uncivil comments above the median is labeled as uncivil speech provoking. The authors use logistic regression to predict the uncivil speech classes of the news articles from the two sources. They use term frequency-inverse document frequency (TF-IDF) of bigrams from the news content. The news classification model achieved 0.64 accuracy, with 0.62 precision and 0.66 recall.

[1]https://www.perspectiveapi.com

Another example focusing on the news text content is the work done by Berk and Filatova (2019) concerning incendiary news articles. Differently from the previous work, the authors manually annotated news articles. They adopt the annotation guidelines of a project from a non-governmental organization. The participants of this project manually annotated 1,036 incendiary articles from various Turkish news media sources. To gather non-incendiary news articles, the authors resort to the Turkish version of the BBC and CNN newspapers, collecting 1,038 BBC articles and 948 CNN articles. To classify news as incendiary, the authors experiment using a linear Support Vector Machine (SVM), Naive Bayes and a feed-forward neural network with word embeddings, plus word and characters n-grams as features. Results show that the feed-forward neural network using word unigrams performs best, reaching an F1 score of 0.95 for cross-corpus classification.

Although not using classification, we found two relevant studies to our problem. Daxenberger et al. (2018) study incivility in comments in German newspapers. They found a news content feature that is associated with incivility in its comments. In this study, hard news (e.g., politics or economics) attracted significantly more uncivil comments than soft news (e.g., arts or lifestyle). Also, Zannettou et al. (2020) conduct a relevant work by performing a large-scale quantitative analysis of 125M comments posted on 412K English news articles. They analyze the collected articles and their comments using temporal analysis, user-based analysis, and linguistic analysis, to shed light on what elements attract toxic[2] comments on news articles. They found statistically significant differences in the comments' toxicity depending on the venue the news was published on, or the event the news was referring to. Also, articles that attracted a substantial number of toxic comments have different linguistic characteristics when compared to the remaining articles.

From the previous literature, we consider our work more similar to Zannettou et al.. The main distinction between both is that we make our data available and consider metadata features (e.g publication time, country, number of replies), while not all those features were considered in the past.

2.3 Toxicity Automatic Detection

In our work, we need to classify news pieces but their comments, as well. In recent years, the number of works addressing the second problem has grown steadily. Two surveys summarize previous research and conclude that the approaches rely frequently on Machine Learning techniques (Schmidt and Wiegand 2017; Fortuna and Nunes 2018). Some online services also follow this approach, such as the *Perspective API*. Their classifier uses Convolutional Neural Networks (CNNs) trained with GloVe word embeddings (Pennington, Socher, and Manning 2014) fine-tuned during training on data from online sources such as Wikipedia and The New York Times. At the OffensEval-2019 shared evaluation task, the Conversation-AI team ap-

[2]The authors use the term hate speech, but classify the data also using the toxicity classifier of the *Perspective API*.

plied the *Perspective API* as a baseline system for toxicity detection and, without any additional training on the contest data, obtained a very competitive result (12th out of 103 submissions, with an F1 score of 0.79) (Pavlopoulos et al. 2019).

2.4 Open Questions

The work we develop is distinct from the studies described in this section. To the best of our knowledge, this is the first study to automatically classify toxicity-associated news from an extensive set of renowned news outlets from the UK and USA using two types of features. We label those as content-based and metadata-based features. By applying those we aim to answer to the following questions:

- Is the comments' toxicity related with the textual content (*content-based*) of the news pieces?
- Or are there other contextual factors (*metadata-based*) of the news associated with the presence of toxicity in its comments?

As there is no public available dataset for this task, we collect and annotate a dataset.

3 Dataset

To study the relation between mainstream published news and the reactions to them, we capture news and comments in a social media platform (Fortuna et al. 2021). We choose Twitter because this platform is used by most news publishers to spread their articles. In the majority of the cases, publishers only include the news piece headline and the respective URL due to the 240 character limit imposed by the platform. Hence, we also gather the entire news piece from the publisher's website.

3.1 Dataset Collection

Communication Sources The choice of communication sources presented in this study is conducted following some guidelines. To this end, we begin to construct a universe composed of communication sources from the United Kingdom and the United States of America. One of the criteria we use was to select the entities cited in Reuters Institute's Digital News Report 2017 (Newman et al. 2017) as an initial focus point and build upon this list. With respect to the United Kingdom, the following sources are selected: "BBC News", "The Guardian", "Daily Mail", "Huffington Post", "Sky News", "The Telegraph", "The Independent", "Mirror", "Yahoo! News", "The Sun", "ITV News" and "Times". Viral content focused "The Lad Bible" and third-party content aggregator "MSN News" are not selected. We add "Metro" for being the most popular newspaper in the region and "Daily Express" for appearing in a research (Matthew 2017) conducted by the YouGov institution that tried to understand how readers see the political direction of communication agencies in the country. Regarding the United States of America, we include: "Yahoo! News", "Huffington Post", "CNN.com", "The New York Times", "BuzzFeed News", "The Washington Post", "NBC / MSNBC News", "ABC News", "BBC News", "USA Today" and "CBS News". Again, news content aggregator "MSN News" is left out. Exceptionally, "Fox News" is removed from the list after a Twitter boycott in which the news outlet did not post on social media, therefore making it impossible for us to gather the data according to our strategy. As an addition to the USA newspaper agencies list, "Time", "The Wall Street Journal" and "NPR News" are selected for their relevance and social importance, and also due to a strong social network presence reflected on the number of followers and posts on Twitter.

As a relevant fraction of the journalistic news presented in the selected websites is from news agencies, we acknowledged that this data was also important for the research and decided to include "Reuters" and "Associated Press (AP)" in the list, two of the most important global news agencies in the world that use English as a primary language.

Data Crawling The data extraction process took place between 2018-12-27 and 2019-01-14 (14 days). To achieve the goals previously described for this task, we collect the posts from each news media Twitter account and its associated comments (i.e. replies) using the Twitter Stream API. We also scrap each news piece from the publisher's websites to extract the complete news piece body. We exclude three sources due to problems in this step of the process. News articles from "The New York Times" and "CBS News" had no comments linked to them. Additionally, news from "Time" had all their titles stored as "Time", without the specific news item title. In the end, this collection includes 29,726 newspapers' tweets (including news title and body obtained from the newspaper website), and 1,995,560 reply tweets.

3.2 Comments Classification

As our work aims to understand how news articles are associated with toxicity in its comments, we first need to classify comments. For this, we use the *Perspective API* (Jigsaw 2020). The classification of comments from the USA took place between 2020-02-17 and 2020-03-17 (29 days) and the classification of UK comments took place between 2020-03-17 and 2020-03-30 (13 days). ´

Having finished the comment classification process, we start by analyzing the results obtained. We consider comments to be toxic if the score of the *toxicity* attribute returned by the *Perspective API* is greater or equal to 0.5. Considering this metric, the dataset used in this study is composed of 425,392 (21.3%) comments signaled as toxic and 1,570,168 (78.7%) signaled as non-toxic.

3.3 Labeling News as Toxicity-associated

Having classified all comments, we focus on news articles classification. For each article, we determine its toxicity-association. This metric is the percentage of toxic comments answering to a news article. For converting this metric to a binary class, we compute the median toxicity for all news articles and conclude that a "typical" news article has 11.1% of toxic reactions to it on Twitter. We label news as toxicity-associated if the percentage of toxicity in its comments is equal or higher than the median for all the news. This procedure is similar to the approach described

by Magu, Hossain, and Kautz (2018). We obtain 14,221 toxicity-associated news pieces and 15,505 non-associated.

Following the mentioned news labelling metric, we examine it by country and newspaper in Figure 1. We conclude that 11 out of 25 newspapers have a percentage of toxic associated news above 50%. For the UK, those are "Sky News" (67.1%), "ITV-news" (60.6%) and "BBC" (58.4%). For the USA, the newspapers are "The Boston Globe" (74%), "Huffpost" (72.1%), "The Washington Post" (69.5%), "ABC" (69.5%), "CNN" (68.2%), "USA Today" (60.3%) and "NPR" (52.7%).

4 Classical Machine Learning Classification Experiments

In these experiments, we use the collected dataset and extract features to classify news as toxicity-associated or not. We use 10-fold cross-validation (Chollet 2017), combined with holdout validation, in which we divide part of the data for cross-validation and parameter tuning with grid search and the other part for testing. We use stratified shuffle split (Pedregosa et al. 2011), assuring an equal distribution of newspapers by fold. We evaluate our model using a macro F1 score. We use 75% of news articles for training and 25% for testing.

We divided the features into metadata-based and content-based features and describe them in detail in the following sections.

4.1 Metadata features

Publication Time We extract the Twitter date of publication and transform it into categorical data. We divide the day into three time periods: morning (8 AM to 11:59 AM), afternoon (12 PM to 7:59 PM), and night (8 PM to 7:59 AM). Since time has a sequential nature, we encode it as an ordinal variable.

Country We use the newspaper country as a feature. Our data is composed of 12,214 articles from the United States of America and 17,753 articles features from the United Kingdom. We use one-hot encoding for this categorical feature.

Newspaper We consider the newspaper of the news articles, which corresponds to 12 distinct USA newspapers, and 13 distinct UK newspapers. We represent the newspaper with one-hot encoding.

Number of Twitter Comments This feature represents the total number of Twitter comments made to a news article. We encountered outliers and, for solving this issue, we apply a three standard deviations cut-off technique (Ilyas and Chu 2019). With this procedure, we remove 241 news articles from our data. Additionally, we found that the data is very positively skewed. With the intent of smoothing this distribution, we perform a logarithmic transformation.

4.2 Content Features

News Topic Category We develop a news topic classifier using spaCy's TextCategorizer module (Honnibal and Montani b). It is based on an ensemble of a bag-of-word (BoW) model with a Convolutional Neural Network to classify text. As a means to achieve a model capable of predicting a news article topic based on its headline, we train the TextCategorizer with the "News Classification Dataset" (Misra 2018). This dataset contains 202,372 news articles and respective topics. Since there is a total of 30 topics in this dataset, and some have a very reduced number of examples, we start by considering the ten topics with most examples. With those, we conduct a first experiment (*Start Categories*) and with the results of it (cf. Table 1), we decide to drop the classes with a weak performance, namely "Comedy" (F1 = 0.49) and "Black Voices" (F1 = 0.52). In the final experiment (*Final Categories*), we can see the performance of the model for the final eight topics (cf. Table 1). With this set of topics, the *Final Categories* model reached an accuracy of 0.80. Additionally, to the eight classes we add a new topic ("Other") to cover the cases where the predicted class has a probability score lower than 0.3. The category model is then applied to our news dataset and a topic is assigned to every news. Also for this feature, we use one-hot encoding.

Title Entities Aiming to extract entities present in news titles, we make use of the named entities recognition module provided by spaCy (Honnibal and Montani a). As news articles may have more than one type of entity simultaneously in their titles, we use one-hot encoding.

Body Entities We also extract relevant entities from the news body using the same spaCy's named entity recognition module. We choose to use: Person, NORP (nationalities or religious or political groups), FAC (buildings, airports, highways, bridges, etc.), ORG (companies, agencies, institutions, etc.), GPE (countries, cities, states), LOC (non-GPE locations, mountain ranges, bodies of water), Product (objects, vehicles, foods, etc), Event (named hurricanes, battles, wars, sports events, etc), Work of Art (titles of books, songs, etc), Law (named documents made into laws), and Language (any named language). We opt to leave out Numerical and Ordinal entities as these would have less interest in our analysis. As having all news body entities represented would result in a high dimensionality problem, we only represent the 1,000 most common news body entity terms using BoW.

Title Keywords With the intent of tokenizing and preprocessing news titles to extract keywords, we use the Rapid Automatic Keyword Extraction (RAKE) algorithm using the NLTK library[3] (Rose et al. 2010). We include news title keywords as features using BoW and selected the most frequent 1,000 keywords to limit the dataset dimensionality.

[3] https://csurfer.github.io/rake-nltk/_build/html/_modules/rake_nltk/rake.html

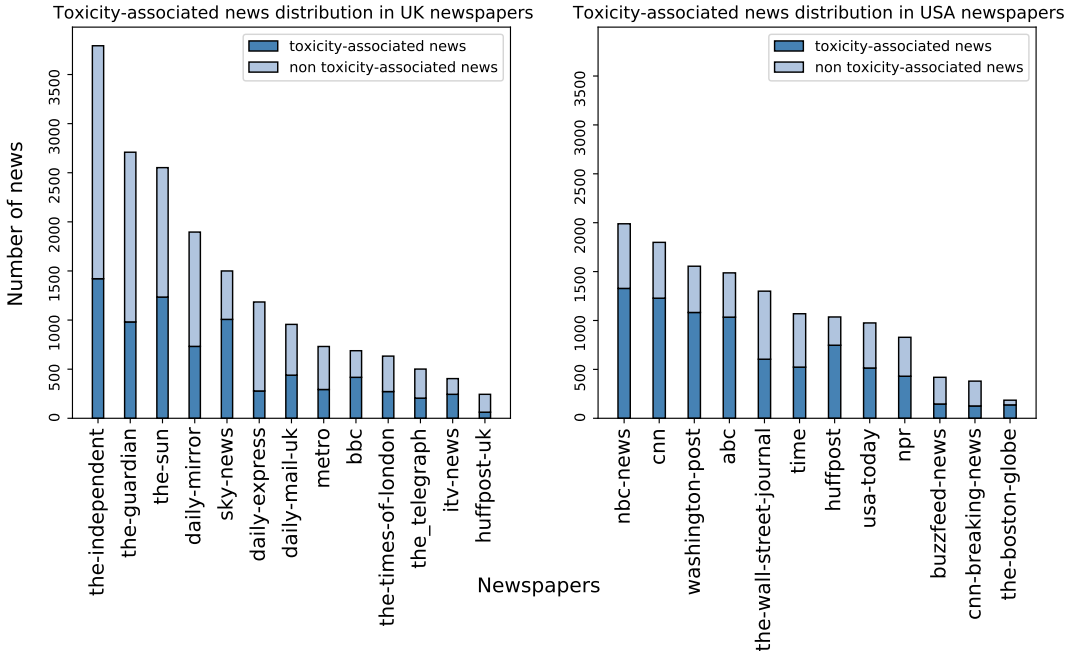

Figure 1: Toxicity-associated news distribution in newspapers.

Table 1: News topic classifier experiments results with different categories classes as target (bold values indicate categories chosen for the next experiment).

News	**Start Categories**			**Final Categories**		
Category	Precision	Recall	F1	Precision	Recall	F1
Black Voices	0.64	0.44	0.52	-	-	-
Business	0.59	0.53	**0.56**	0.67	0.49	**0.56**
Comedy	0.65	0.39	0.49	-	-	-
Entertainment	0.72	0.81	**0.76**	0.77	0.85	**0.76**
Healthy Living	0.68	0.69	**0.68**	0.66	0.70	**0.68**
Parents	0.67	0.67	**0.67**	0.77	0.63	**0.69**
Politics	0.82	0.90	**0.86**	0.86	0.89	**0.88**
Queer Voices	0.81	0.64	**0.72**	0.79	0.65	**0.71**
Sports	0.78	0.71	**0.75**	0.73	0.72	**0.72**
The Worldpost	0.74	0.69	**0.71**	0.68	0.76	**0.72**

4.3 Algorithms

With the intent of comparing various classifiers to label toxic-associated news accurately, we choose four different classifiers from the *Scikit-learn* library (Pedregosa et al. 2011): SVM with linear kernel using default parameters; Logistic Regression with a maximum number of iterations set to 1,000 and remaining parameters as defaults (Pedregosa et al. b); Random Forest with the number of estimators set to 50 and remaining parameters as defaults (Pedregosa et al. c); and GBDT with a learning rate of 1.0 with remaining parameters as defaults (Pedregosa et al. a). This decision is based on an analysis of the standard use of machine learning classifiers in the area of news classification (Kaur and Bajaj 2016; Helmstetter and Paulheim 2018; Suleymanov et al. 2018; Dadgar, Araghi, and Farahani 2016). Additionally, we choose to include GBDT since, although being used with variations, this algorithm has gained popularity in machine learning competitions, providing solutions with promising scores, and greater efficiency (Sandulescu and Chiru 2016; Volkovs, Yu, and Poutanen 2017).

5 BERT Experiments

We use $\text{BERT}_{\text{LARGE}}$ (L=24, H=1024, A=16) with 340M parameters in total. We opt for this model as it outperformed $\text{BERT}_{\text{BASE}}$ across all tasks in the work of Devlin et al. (2019). The input sentences are first processed with the BERT basic tokenizer to perform punctuation splitting, lower casing, and invalid character removal. The maximum sequence length is defined as 80, with shorter sequences padded and longer sequences truncated to this length. We use a batch size of 32 and fine-tune for 3 epochs over the data of all datasets. The dropout probability is set to 0.1 for

all layers and the Adam optimizer is used with a learning rate of 2e-5. We applied BERT both to the title and to the body of the news article.

6 Results and Discussion

6.1 Classical Machine Learning Classification Results

Regarding the test results of our experiments, we start by analyzing the impact of each feature in the classical classification setup, and this for every model (cf. Table 2). No matter the model, the number of comments to a news article is the feature that allows the best performance. The best score was achieved when using GBDT (F1 = 0.680). This result is consistent with the good performance in previous literature for this algorithm (Sandulescu and Chiru 2016; Volkovs, Yu, and Poutanen 2017).

Some of the features achieved a very similar F1 with different algorithms, namely title keywords (Random Forest, F1 = 0.591), newspaper (Random Forest, F1 = 0.580), and topic category (SVM, F1 = 0.570). Finally, we note that the time, country, title, and body entities achieved the worst performance with F1 scores worse than 0.53 for all the algorithms. To better explore the interference of the different features we tried three different configurations for the classification task: using only content features, only metadata features, or both sets (content and metadata) (cf. Table 3 and features groups in sections 4.1 and 4.2).

Metadata features achieved better performance than content features with all the algorithms. The best result was achieved with metadata features and again with the GBDT algorithm (F1 = 0.723). The result of combining both metadata and content features does not improve this result. Neither when using GBDT (F1 = 0.722) where the result is similar, neither when using any other algorithm. Results also show that the metadata features achieved better performance than when compared to use any individual feature.

6.2 BERT Results

Regarding the test results of the BERT experiments (cf. Table 4), we conclude that using the title (F1 = 0.647) it is better than using the body (F1 = 0.499) for classification. We believe this is because the news title and comments were collected on Twitter while the news body is from the news website. Hence, probably Twitter's readers do not open the link for reading the complete news piece.

When comparing with the classical machine learning experiments, using the title with BERT allows us to reach slightly better results (F1 = 0.647) than when using content features alone or grouped (best result with content features and GBDT with F1 = 0.635).

On the other hand, BERT performance could not overcome the best results with the classical machine learning experiments. This means that using BERT was worse than using metadata features with GBDT (F1 = 0.723). This result indicates that the context of the news (e.g., number of comments of a news, newspaper, time, country) was more important than the text in itself for the appearance of toxicity in our experiments. From the metadata set of features, the number of comments in the news was the most important variable. This indicates that people publishing toxicity are willing to do it in hot topics, or that hot topics are more prone to induce toxicity.

6.3 Comparison with State of the Art Results

Despite not using the same dataset, we compare these results with previous studies. In the work of Magu, Hossain, and Kautz (2018), when using TF-IDF and logistic regression, the model achieved an F1 score of 0.639[4]. This result is similar to the best result we achieve when using only classical machine learning text mining features (F1 = 0.635 with content and GBDT). We could only overcome this result with the use of metadata features. Berk and Filatova (2019) also use a text mining approach and report an F1 score of 0.95 for classification. In this case, the results are much better than the ones we achieve. We believe this may be due to a confounding effect in the experiment of that study. In the selection of news, the authors mention that incendiary news are manually annotated from various Turkish news media while non-incendiary news come from the Turkish version of the BBC and CNN newspapers. Hence, the classifier may be detecting the original journal-style instead of the target variable.

Based on Daxenberger et al.'s (2018) conclusion that some news topics receive more uncivil comments, we would expect that this feature would work better with our models. This was not the case and, in the future, we can experiment to improve our procedure for topic extraction.

When comparing our results with Zannettou et al. (2020), we think that it is consistent regarding metadata features mentioned as relevant by these authors. However, the authors point out the importance of certain events and the linguistic characteristics of news to attract toxicity. In our study, we did not find good results when applying content features. This inconsistency leaves open if a different set of text mining features would allow better results with our dataset, or if newsreaders' reactions in Twitter are less related with news content (this because only the title of the news is available in the platform and the entire content it is only accessed if the reader opens the provided URL for the news).

7 Conclusions

We addressed two open questions related to toxicity-associated news: 1. Is the comments' toxicity related with the textual content of the news pieces? and 2. Are there other contextual factors of the news (i.e. metadata) interfering with the presence of toxicity in its comments? We believe that the conducted experiments allowed us to bring insight to the open discussion. By using classical machine learning, we found that metadata features have the best performance when used together with a GDBT classifier. This was the case even when comparing with state of the art text classifiers, such as BERT.

[4]We computed this score based on the authors' reported precision (0.62) and recall (0.66).

Table 2: Individual features results with the test data.

Group	Features	Algorithm	Precision	Recall	Macro-F1
Metadata Features	Publication time	SVM	0.000	0.000	0.000
		Log. Regression	0.510	0.423	0.462
		Random Forest	0.510	0.423	0.462
		GBDT	0.510	0.423	0.462
	Country	SVM	0.603	0.509	0.522
		Log. Regression	0.603	0.509	0.522
		Random Forest	0.603	0.509	0.522
		GBDT	0.603	0.509	0.522
	Newspaper	SVM	0.560	0.309	0.398
		Log. Regression	0.551	0.343	0.423
		Random Forest	0.648	0.525	0.580
		GBDT	0.650	0.504	0.568
	Number of Twitter Comments	SVM	0.679	0.570	0.620
		Log. Regression	0.677	0.608	0.641
		Random Forest	0.673	0.664	0.668
		GBDT	0.664	0.699	**0.680**
Content Features	News Topic Category	SVM	0.533	0.611	0.570
		Log. Regression	0.600	0.420	0.494
		Random Forest	0.597	0.439	0.506
		GBDT	0.597	0.439	0.506
	Title Entities	SVM	0.546	0.121	0.197
		Log. Regression	0.359	0.532	0.429
		Random Forest	0.529	0.334	0.407
		GBDT	0.532	0.359	0.429
	Body Entities	SVM	0.563	0.412	0.476
		Log. Regression	0.549	0.472	0.508
		Random Forest	0.537	0.506	0.521
		GBDT	0.570	0.427	0.488
	Title keywords	SVM	0.638	0.462	0.536
		Log. Regression	0.618	0.518	0.564
		Random Forest	0.612	0.571	0.591
		GBDT	0.651	0.409	0.502

Table 3: Performance on the test set of models trained with different features configurations: using only content features, only metadata features, or both.

Features	Model	Precision	Recall	Macro-F1
Content	SVM	0.643	0.585	0.613
	Log. Regression	0.640	0.621	0.630
	Random Forest	0.640	0.574	0.606
	GBDT	0.650	0.620	0.635
Metadata	SVM	0.680	0.717	0.698
	Log. Regression	0.697	0.660	0.678
	Random Forest	0.701	0.664	0.682
	GBDT	0.695	0.753	**0.723**
Content + metadata	SVM	0.700	0.700	0.698
	Log. Regression	0.703	0.703	0.698
	Random Forest	0.713	0.713	0.706
	GBDT	0.704	0.704	**0.722**

Table 4: Performance on the test set of BERT models.

Method	Precision	Recall	Macro-F1
BERT title	0.648	0.648	**0.647**
BERT body	0.499	0.499	0.499

According to our results, it seems clear the importance of contextual features such as metadata (e.g., date, country, newspaper, news number of comments). However, we do not aim at disregarding previous literature results and findings. The conclusions here may differ from the previous studies because we do not use the same data: while previous works focus on viral news providers (Zannettou et al. 2020), we focus on mainstream newspapers; other focus on German (Daxenberger et al. 2018), and Turkish (Berk and Filatova 2019) while we address the English context; we collect data from Twitter while previous works focus on newspapers' comments sections, or only news information (Berk and Filatova 2019).

Furthermore, our work helps the study of toxicity-associated news by providing a public annotated dataset. The lack of datasets it is a critical issue that makes difficult the comparison of results across studies. With this resource we can further investigate some results, namely by using a more extensive set of content-based and metadata-based features, and also other contextual features apart from the metadata. Another interesting research line that it is also possible with this dataset is understanding if the content of previous comments in a thread influences the appearance of toxicity in the scope of online conversations, and in a recent work (Almerekhi et al. 2020).

References

Almerekhi, H.; Kwak, H.; Salminen, J.; and Jansen, B. J. 2020. Are these comments triggering? predicting triggers of toxicity in online discussions. In Huang, Y.; King, I.; Liu, T.; and van Steen, M., eds., *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, 3033–3040. ACM / IW3C2.

Anderson, A. A.; Brossard, D.; Scheufele, D. A.; Xenos, M. A.; and Ladwig, P. 2014. The "nasty effect:" online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication* 19(3):373–387.

Berk, E., and Filatova, E. 2019. Incendiary news detection. In *Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference*, FLAIRS-32, 161–166. Sarasota, USA: AAAI Press.

Chollet, F. 2017. *Deep learning with python*. USA: Manning Publications Co., 1st edition.

Dadgar, S. M. H.; Araghi, M. S.; and Farahani, M. M. 2016. A novel text mining approach based on tf-idf and support vector machine for news classification. In *Proceedings of 2nd IEEE International Conference on Engineering and Technology*, ICETECH 2016, 112–116. Washington, USA: IEEE Computer Society.

Daxenberger, J.; Ziegele, M.; Gurevych, I.; and Quiring, O. 2018. Automatically detecting incivility in online discussions of news media. In *Proceedings - IEEE 14th International Conference on eScience, e-Science 2018*, 318–319. Washington, USA: IEEE Computer Society.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Duggan, M. 2014. Online harassment. Technical report, Pew Research Center, Washington, USA. Available at https://radimrehurek.com/gensim/summarization/keywords.html.

Fortuna, P., and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Survey* 51(4).

Fortuna, P.; Cruz, L. B.; Maia, R.; Cortez, V.; and Nunes, S. 2021. Toxicity Associated News Dataset. INESC TEC Research Data Repository. Available at https://doi.org/10.25747/BEVW-D436. Dataset.

Georgakopoulos, S. V.; Tasoulis, S. K.; Vrahatis, A. G.; and Plagianakos, V. P. 2018. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, SETN '18. New York, USA: Association for Computing Machinery.

Helmstetter, S., and Paulheim, H. 2018. Weakly supervised learning for fake news detection on twitter. In Brandes, U.; Reddy, C.; and Tagarelli, A., eds., *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM 2018, 274–277. IEEE Computer Society.

Honnibal, M., and Montani, I. spacy named entities description. Available at https://spacy.io/api/annotation#named-entities. Online (accessed March 20, 2020).

Honnibal, M., and Montani, I. spacy textcategorizer model. Available at https://spacy.io/api/textcategorizer. Online (accessed April 06, 2020).

Ilyas, I. F., and Chu, X. 2019. *Data Cleaning*. New York, USA: ACM, 1st edition.

Jigsaw. 2020. Perspective api. Available at https://github.com/conversationai/perspectiveapi, accessed last time in May 2020.

Kaur, G., and Bajaj, K. 2016. News classification and its techniques: A review. *Journal of Computer Engineering of the International Organization Of Scientific Research* 18(1):22–26.

Magu, R.; Hossain, N.; and Kautz, H. 2018. Analyzing uncivil speech provocation and implicit topics in online political news. *arXiv preprint arXiv:1807.10882*.

Matthew, S. 2017. How left or right-wing are the uk's newspapers? Available at https://yougov.co.uk/topics/politics/articles-reports/2017/03/07/how-left-or-right-wing-are-uks-newspapers.

Misra, R. 2018. News category dataset. https://www.researchgate.net/publication/332141218_News_

Category_Dataset. Data retrieved from the Huffpost News, https://www.huffpost.com. Online (accessed April 6, 2020).

Newman, N.; Fletcher, R.; Kalogeropoulos, A.; and Nielsen, R. 2017. *Reuters Institute Digital News Report 2017*, volume 2017. Reuters Institute for the Study of Journalism.

Orellana-Rodriguez, C., and Keane, M. T. 2018. Modeling and predicting news consumption on twitter. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, 321–329. New York, NY, USA: Association for Computing Machinery.

Pavlopoulos, J.; Thain, N.; Dixon, L.; and Androutsopoulos, I. 2019. ConvAI at SemEval-2019 task 6: Offensive language identification and categorization with perspective and BERT. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 571–576. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. Gradient boosting for classification. Available at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html. Online (accessed, 2020).

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. Scikit-learn logistic regression classifier. Available at https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. Online (accessed, 2020).

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. Scikit-learn random forest classifier. Available at https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Online (accessed, 2020).

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(85):2825–2830.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing EMNLP*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Rose, S.; Engel, D.; Cramer, N.; and Cowley, W. 2010. *Automatic Keyword Extraction from Individual Documents*. John Wiley Sons, Ltd.

Sandulescu, V., and Chiru, M. 2016. Predicting the future relevance of research institutions - the winning solution of the KDD cup 2016. Available at http://arxiv.org/abs/1609.02728.

Schmidt, A., and Wiegand, M. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*.

Suleymanov, U.; Rustamov, S.; Zulfugarov, M.; Orujov, O.; Musayev, N.; and Alizade, A. 2018. Empirical Study of Online News Classification Using Machine Learning Approaches. In *IEEE 12th International Conference on Application of Information and Communication Technologies*, 1–6. Washington, USA: IEEE Computer Society.

Volkovs, M.; Yu, G. W.; and Poutanen, T. 2017. Content-based neighbor models for cold start in recommender systems. In *Proceedings of the Recommender Systems Challenge 2017*, RecSys Challenge '17. New York, NY, USA: Association for Computing Machinery.

Zannettou, S.; Elsherief, M.; Belding, E.; Nilizadeh, S.; and Stringhini, G. 2020. Measuring and characterizing hate speech on news websites. In *12th ACM Conference on Web Science*, WebSci '20, 125—134. New York, NY, USA: Association for Computing Machinery.