FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Prediction of toxicity-generating news using machine learning

Luís Braga da Cruz

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Sérgio Nunes

Cosupervisor: Paula Fortuna

July 22, 2020

Prediction of toxicity-generating news using machine learning

Luís Braga da Cruz

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Gabriel Torcato David, PhD

External Examiner: Pedro Rangel Henriques, PhD

Supervisor: Sérgio Sobral Nunes, PhD

July 22, 2020

Abstract

The increased popularity of social networks in today's society, in combination with the facilitated acquisition of devices that provide access to those platforms, developed the conditions for the increase of users in these platforms. This growth sparked the attention of businesses such as news media outlets, which started enforcing their presence in social networks to provide users with news articles widening their audience. The use of social networks provides users with a costless and quick interaction, making interactions more impersonal and contributing to an increase in aggressive communication. Therefore, such interactions must be identifiable so that social networks can combat the use of aggressive communication harmful to its environment. Thereby, in this thesis, we investigate the problem of detecting toxicity-generating news.

An initial objective was to make a review of the topic from a computer science perspective. We analysed the differences between toxicity and related concepts (hate, offensive and uncivil speech) and how they overlapped, complementing with our definition. Regarding past research, we could only find similar studies regarding the prediction of incendiary news, incivility and hate speech generating news. Despite the studies found, we concluded that, to our knowledge, there had been no studies on prediction of toxicity-generating news. With this conclusion, we decided to perform a review of studies on the area of news classification intending to review standard feature extraction techniques and machine learning algorithms used in the area. A second objective was the study of toxicity in comments and news present in our dataset. We used a dataset developed in the context of the *Stop PropagHate* project to predict hate speech generating news. The dataset contained news articles Twitter posts related to news media outlets from the USA, UK, Portugal and Brazil and respective user Twitter comments to those articles. We used the *Perspective API* to classify comments as toxic. From this classification, we concluded that the median toxicity in comments was 11.1%. With this metric, we considered news to be toxicity-generating if the mean toxicity of the comments towards that article was equal or higher than the median toxicity.

As a final objective, we intended to predict news as toxicity-generating and understand which features contributed for the prediction. For this goal, we extracted meta-data features and news-content features. We conducted experiments with training, validation and test phases. We performed several feature combination experiments and concluded that our best model represented a combination of meta-data features and news content features, reaching an F1 score of 0.74. Furthermore, analysing the feature correlation, we concluded that the model's performance was not resultant of a subset of features, but rather the combination of all. Regarding the features that most contributed to the classification of toxicity-generating news, we concluded that features such as the number of comments a news originated, and title keywords, were the most significant contributors to the classification. Moreover, we concluded that articles which titles contained keywords relative to highly debated social topics such as "racist", "gay" and "nazi" in conjunction to political entities such as "Trump" contributed to the identification of toxicity-generating news. All the mentioned points and objectives were met. We successfully developed a model reasonably capable of classifying toxicity-generating news, and understand factors behind the phenomena.

Resumo

O aumento da popularidade das redes sociais nos dias de hoje, em combinação com a fácil aquisição de dispositivos que permitem o acesso a essas plataformas, contribuíram para o aumento dos seus utilizadores. Este aumento chamou à atenção de negócios como por exemplo dos media, os quais começaram a reforçar a sua presença nestas plataformas com o objetivo de partilhar as suas notícias alargando o seu público. A utilização de redes sociais permite aos utilizadores uma fácil e rápida interação, o que consequentemente levou a um aumento na agressividade da comunicação. Deste modo, é importante a identificação de comunicação agressiva, para que estas plataformas consigam combater uma comunicação danosa ao seu ambiente. Assim sendo, nesta tese, investigamos o problema da deteção de notícias geradoras de toxicidade.

Como objetivo inicial, fizemos uma revisão geral do estado da arte no contexto da engenharia informática. Analisámos as diferenças entre toxicidade e conceitos relacionados (discurso ofensivo, incivil e de ódio) e como estes se sobrepõem, contribuindo com a nossa definição de toxicidade. Relativamente a estudos relacionados, apenas encontramos trabalhos relativos à deteção de notícias incendiárias, incivis e geradoras de discurso de ódio. Apesar dos trabalhos encontrados, verificámos que este trabalho é o primeiro a abordar a deteção de notícias geradoras de ódio. Com esta conclusão, decidimos expandir a nossa revisão do estado da arte à área da classificação do tópico de notícias. Esta decisão baseia-se na intenção de revisão das técnicas normativas de extração de recursos usadas na área de inteligência artificial e dos seus algoritmos de classificação. Um segundo objetivo foi o estudo da presença de toxicidade em comentários e notícias que constituem a coleção de dados utilizada. Para alcançar este objetivo, usamos dados recolhidos no âmbito do projeto *Stop PropagHate*, constituído para a deteção de notícias geradoras de discurso de ódio. Estes dados são relativos a *tweets* de notícias e respetivos *tweets* de comentários de jornais dos EUA, Reino Unido, Portugal e Brasil. Para a classificação de comentários como tóxicos foi usada a *Perspective API*, permitindo a extração da mediana da toxicidade presente nos comentários (11.1%). Com esta métrica, classificamos como notícias geradoras de ódio aquelas cujos comentários tivessem a sua media de toxicidade igual ou superior à mediana.

Como objetivo final, criamos um classificador capaz de identificar notícias geradoras de toxicidade. Investigamos dentro das características extraídas, quais contribuíram para a identificação. Com estes objetivos, recolhemos características baseadas no conteúdo dos dados e baseadas na meta-data. Conduzimos várias experiências combinando características, obtendo o melhor modelo uma performance F1 de 0.74. Analisando as características que contribuiram para a classificação, concluímos que a performance obtida não se deve a um subconjunto das características selecionadas, mas sim de todas. Concluímos que características do título e o número de comentários a uma notícia foram duas contribuidoras para a classificação. Adicionalmente, concluímos que termos referentes a temas controversos como “gay”, “racismo” e “nazi”, em conjunto com figuras politicas como “Trump”, contribuíram para a identificação de toxicidade em notícias. Todos os pontos e objetivos mencionados foram cumpridos. Desenvolvemos um modelo razoavelmente capaz de classificar notícias geradoras de toxicidade e entender os fatores por detrás dos fenomeno.

Acknowledgements

I would like to thank professor Sérgio Nunes and Paula Fortuna for all the help and guidance on my research. I would also like to express my gratitude to my parents, who helped move all stones found along the way, to my friends who accepted all my requests for help. To my grandparents for the inspiration to be perfectionist. For my ancestors, for the aspiration to set the bar high. Finally, to Beatriz for being there when strength weakened.

Luís Braga da Cruz

"Success is not final, failure is not fatal: it is the courage to continue that counts."

Winston Churchill

Contents

List of Figures

List of Tables

Abbreviations

API	Application Programming Interface
AUC	Area Under Curve
BoW	Bag of Words
CNN	Convolutional Neural Network
GBDT	Gradient Boosted Decision Trees
KNN	K-Nearest Neighbors
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
NN	Neural Network
PoS	Part-of-speech
RAKE	Rapid Automatic Keyword Extraction
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine

Chapter 1

Introduction

1.1 Problem Contextualization

The increasing use of social networks and online communication makes it easier to share one's opinion [73; 55]. This fact turns the attention of various industries towards these online platforms, for instance, as newspapers do. As an outcome of the interest of moving their business towards social networks, there has been a boost in recent years on the number of journals with an online presence, which in turn increments the number of online news available for social network users to read [25]. In this context, the interaction of users towards news pieces has greater importance since, in today's time, having users finite time to spend, with more resources competing for their attention, challenges arise for users and news media producers that compete for that attention.

The presence of news covering controversial themes can generate reactions in users that lead them to use comment sections to spread negative behaviour that is detrimental to the journalistic environment, where such comments are accessible to all users of such platforms. This way, toxic, offensive and uncivil speech may arise. Although these concepts overlap, it is essential to understand that they are distinct.

Hate speech is the public expression of encouragement and promotion of hatred, discrimination and hostility towards a particular group of people. Such expression contributes to an atmosphere of intolerance, which in turn instigates more attacks against such groups [36]. Offensive speech can be defined as any communication that might contain derogatory words towards a group, but it is used in a different qualitative way, that is, the context is considered. (e.g., *I'm tired of people saying I look like my brother calling me Deondre' like serious Succ My Ass fag asses*) [3]. Uncivil speech is the use of insulting language to attack a non-protected group or a non-protected group member [48].

Since such interactions are harmful to readers and the journalistic environment, it is necessary to take action to regulate this kind of negative behaviour. With this goal in mind, researchers are focusing their attention on the automatization of the identification of this kind of rudeness and harassment, since most of the moderation still requires manual review [78]. Our work will be focusing on this area as well. Moreover, we intend to study whether there are properties in news that make it more likely to attract negative interactions in comments. The identification of such environments may benefit newspapers so that their focus can be directed to detecting these phenomena in news.

1.2 Motivation and Objectives

As the focus of this thesis, we intend to first delve on what solutions have already been developed, making it possible to understand which approaches provide better results.

Characterizing the state of the art in this field is an intricate task. Firstly, the detection of toxic speech regards the field of Natural Language Processing (NPL) which converges with various fields (e.g., speech segmentation). Secondly, it is crucial to identify the differences between toxic speech and other concepts (e.g., fake news), that go hand-to-hand with incendiary speech and are extensively studied by researchers. Thirdly, the area our work engages in is also subject of study in other fields, such as Law and Media Studies. Despite that, when focusing on the Computer Science perspective, the amount of investigation in the area of prediction of toxicity-generating news is narrow.

The second objective of our work is to use our own dataset for the prediction of toxicity-generating news. Using our dataset is an essential aspect since, despite the existence of several datasets publicly available regarding the area of study, the definition of what those datasets are based on may differ from our work. Adding to this point, we mean to prepare our dataset to be used in possible future studies regarding the Portuguese language, contributing as a novelty in the area. The final goal of our thesis is to develop a model using classical machine learning algorithms, that can predict if a news article will generate toxic behaviour. We intend to use our dataset of English comments and news to conduct our research. Despite the use of features relevant to news articles, we will be using the comments relative to those news pieces to classify those news as toxicity-generating news.

1.3 Document Structure

Along with the Introduction, this thesis is composed of five other chapters. Chapter 2 provides a brief introduction behind the methods used in all the process of classifying toxicity-generating news. In Chapter 3, is presented an overview of the state of the art regarding the classification of toxicity-generating news. In this section, we mention the approaches taken by relevant studies. Chapter 4, details the dataset used in our work, describing its composition. This chapter also outlines the metrics used for the classification of comments and news as toxicity-generating. Chapter 5 presents the methodology, the features extraction and selection process and reports the results gathered by our models. Finally, Chapter 6 presents the conclusions of our experiments, discussing the limitations and difficulties we had along the way in conjunction with future work to be done.

Chapter 2

Text classification theoretical background

With the intent of providing an introduction on the methods and processes used in this thesis, in this chapter we describe a set of techniques often used by researchers to prepare and process text for the classification task of machine learning algorithms. We also present a brief explanation of standard feature extraction techniques, as well as machine learning algorithms used for classification problems. Along with the description of such algorithms, we provide an explanation of the performance metric widely used in our area of research.

2.1 Data preprocessing

In order to efficiently classify text, it is necessary to have a clean dataset in order to achieve better performance in machine learning. Classifying text poses a significant challenge not only for machines but also for people, due to the noise of data in conjunction with a need for world knowledge [58]. Furthermore, users tend to distort text using emoticons, symbols, hashtags, and numbers. Text may also include URL's and mentions. This way, it is of significant importance that researchers standardize the dataset using the following techniques so that machine learning models can achieve relevant results:

Lowercasing is defined as the technique of transforming a sentence to only lowercase characters. This technique is a common feature of text preprocessing in studies, since it removes ambiguities and varieties within the same word, such as "Jew" and "JeW" that otherwise could be considered distinct words.

Tokenization is the process of separating a string into various tokens such as words, discarding punctuation. Tokenization is a method widely used by researchers when working with classification tasks since in most cases, text information is present in data. Regularly punctuation may contain useful sentiment information (e.g., "!!!"), for this reason, some tokenizers consider some types of punctuation.

Lemmalization is the method of grouping the extensions of a root word, so that those can be analyzed as one item. Lemmalization depends on the detection of the meaning of the word in a sentence.

Stemming similarly to lemmalization, stemming is the reduction of the derivations of a root word. Differently from lemmalization, this technique does not consider the morphological analysis of the words. Thus, different words with the same root can represent two distinct items (e.g., "studies" converts to "studi" and "studying" to "study").

Stopword Removal is the process of removing words, such as "a" that frequently appear in text and do not add any contribution to the search task. Thus, this method is very used in NPL problems.

Hashtag Removal hashtags can be removed or normalized, becoming a sentence (e.g., "#refugeesnotwelcome" resulting in "refugees not welcome"). The normalization may interest researchers since most provide additional information and keywords.

Emoticon Removal the removal of emoticons varies with the purpose of the study. Emoticons can be useful for sentiment analysis due to their sentiment denotation. This way, emoticons are useful for research regarding harmful behaviour, since frequently such behaviours are expressed in an emotional manner. The use of information portrayed by emoticons is, therefore, subjective to the intentions of each particular study.

2.2 Text mining techniques

With the growing attention of the scientific community on the detection of offensive communication in text, studies regarding the topic are growing at a considerable pace [55]. The increase of studies leads to the increase of combinations between features and algorithms explored. In this section, we describe the most common feature extraction techniques used by researchers to extract useful information from text.

TF-IDF Term Frequency-Inverse Document Frequency is the value that measures the importance of a word to a document in a collection or corpus. The value increases proportionally to the number of times a word appears in the document. Words used very frequently do not influence the metric, due to adjustment by the offset of the number of documents in the corpus that contain the word [68].

N-grams N-grams compose a sequence of n words from a given sample of text, commonly represented as a list. This feature is commonplace in studies regarding text classification due to its text processing nature, making word and character n-grams an advantageous technique since it allows the classifier the ability to embody the meaning of an N size word or character [42].

Bag of Words is the representation of words occurring in a text with disregard of the order, semantic and syntactic meaning of each word present in the bag, but keeping its frequency [30].

Word embeddings is a recurrent feature present in deep learning approaches. It is a feature that maps words from the vocabulary to vectors, whose relative similarities correlate with semantic similarity. The following four paragraphs describe various word embedding techniques.

Word2Vec is one of the techniques used to produce word embeddings. Word2Vec receives a large text and generates a vector space, with each word assigned a corresponding vector. Word vectors are positioned in the vector space so that, words sharing similar contexts in the text are located in proximity to one another in the space. Word2Vec can have two model architectures them being: continuous bag-of-words (CBOW) or continuous skip-gram. In CBOW, the model predicts the current word, looking at the neighbouring words. In continuous skip-gram, the model uses the current word to make a prediction of the surrounding words. Such prediction is accomplished by assigning heavier heights to words in the proximity, and lighter weighs to distant context words. CBOW is faster, in contrast, skip-gram does a better job for words with lower frequency [51].

GloVe is one of the techniques also used to produce word embeddings. Same as Word2Vec, GloVe creates a vector space, with a vector for each word present in the provided text. This technique differentiates from the Word2Vec in that GloVe is a count-based model, meaning that, the vector space positioning of word vector is made by word-word co-occurrence statistics [64].

FastText Similarly, as other word embedding techniques, FastText creates a vector space representing words. Differences are on the word vector representations. Each vector represents the character n-grams of the word. (e.g., “matter” with n = 3, has a FastText representation of <ma, mat, att, tte, ter, er>) [13; 41].

BERT Bidirectional Encoder Representations from Transformers are different from other word embeddings in the basis that, unidirectional word embeddings only lookup words sequentially left-to-right or right-to-left. BERT makes use of the transformer encoder to read a complete series of words at once. Hence it is regarded as bidirectional, or non-directional. This characteristic allows the model to acquire the word context based on all of its surroundings, [23].

Sentiment Analysis Considering that harmful speech usually arises when users express their opinion, embracing the sentiment behind such messages in combination with other features may improve the classification process. Studies frequently approach sentiment identifying the polarity of the messages [60].

Part of Speech is a feature that categorizes each word regarding the grammatical classification (e.g., adjectives, determiners, verbs). Even though words often can contain more than one grammatical meaning (e.g., glue, can be classified as verb and noun), PoS is often used in combination with sentiment analysis and n-grams, which is helpful to acquire dependencies between words [6].

2.3 Classification algorithms

With the development of machine learning for the classification of text and news, many studies have used various algorithms for the classification task. In recent years, researchers have experimented using deep learning approaches for the classification of text, with reports of better performances compared with classical machine learning algorithms. This way, the following paragraphs describe various algorithms used by researchers in such classification processes.

SVM Support Vector Machine is an algorithm that can solve linear or regression problems. The algorithm uses a hyperplane which separates the data into classes. This algorithm is widely used in studies, particularly in classification problems with high dimensional data [17].

Naive Bayes is a simple probabilistic classifier applying Bayes' theorem with strong (naive) independence assumptions between the data's feature. A naive Bayes classifier considers each feature to contribute independently to the classification probability, regardless of any possible correlations between features. Bayes classifiers have a good advantage of requiring small training data to estimate the parameters necessary for classification [46].

Logistic Regression is an algorithm that uses a logistic function to model a binary dependent variable. The logistic function converts the independent variables (variables calculating the possibility of a value being labelled as one of the binary values) into a probability [61].

GBDT Gradient Boosting Decision Tree is an algorithm which consists of an ensemble model, typically a decision tree, containing weak prediction models. The learning procedure sequentially fits new models to provide a more accurate estimate of the response variable, allowing the model to learn from previous mistakes. The main idea behind this algorithm is to construct the new base-learners to have the highest relationship with the negative gradient of the loss function, linked with the entire ensemble [56].

2.3.1 Deep learning

Since recent studies report having better performance with approaches using deep learning, we describe standard deep learning algorithms used for the classification problems, the paragraphs describe different neural networks.

CNN Convolutional Neural Networks are a type of neural networks that are composed of an input layer, an output layer and a hidden layer that incorporates various convolutional layers. Such layers undertake operations called convolutions. These are linear operations involving the multiplication of a set of weights with the input [47].

RNN Recurrent Neural Networks are a type of artificial neural network that use internal memory to process sequences of data, unlike CNN's. Such processing allows forming a directed graph from a temporal sequence. The recurrent term referrers the two classes of networks, them being finite and infinite impulse. The finite impulse recurrent network is a directed acyclic graph that can be substituted by a feed-forward NN. On the other way, an infinite recurrent network can be substituted by a directed cyclic graph.

LSTM Long Short-Term Memory neural networks are one of the architectures of an RNN. They are composed of a memory cell, which contains information about dependencies for a time duration, and three regulators, input, output and forgotten, that control the flow of information [32].

GRU Gate Recurrent Units are a simpler version of an LSTM. As the name proposes, they use a set of gates to control the flow of information but lack the presence of a memory cell. The activation of the GRU at a specific time is a linear interpolation between the previous activation and the candidate activation [15].

Decision Trees is a decision making algorithm that considers characteristics of data, which are the branches of the tree, and output a decision represented as the leaves. Such algorithms have the advantage of being intuitive and straightforward to understand, as data normalization is not needed since the decision tree will have the same structure with or without data preprocessing, and perform well on large datasets.

2.4 Classification Assessment

Since the task of classifying toxicity may be a problem of binary classification, with a text having only two possibilities of being toxic or not, model performance may be scored using the F1 score. This measurement, also known as the F-score tests the accuracy considering the Precision and Recall. The Precision is the division of the true positive classifications by the all positive classifications. Recall consists of the division of true positives by the relevant elements. The relevant elements are the set of true positives and false negatives, in other words, all the values that should have been considered as positives. F-score varies from 0 to 1, a score of 1 representing perfect Precision and Recall and 0 otherwise.

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{2.1}$$

Considering situations where the target variable in a dataset has an imbalanced distribution, variations of the F1 score may be applied. Employing variations of the F1 score may be the case if researchers have the intention of crediting more value the minority class or not. This way, we describe two variations of the F1 score useful when having an imbalanced dataset.

F1-macro is a macro-average approach will compute the macro average of the Precision and Recall metric independently for each class of the target variable, finishing with the calculation of the F1 score of those metrics. This way, when using a macro-average approach, all classes are treated equally, turning this metric insensitive to the imbalance nature of the classes [50].

F1-micro is a micro-average approach that combines the contributions of all classes to compute the average metric. This way, a micro-average approach takes into consideration the imbalance of classes in dataset [50].

Chapter 3

Revision on toxicity prediction in news

Intending to investigate the work done in the area of detection of toxicity-generating news, we describe the variety of terms often addressed when studying negative communication and its results. With this intention, we further enumerate in the present chapter an overview of the current status of what has already been explored regarding the prediction of toxicity-generating news. This chapter provides an analysis of the difficulties and challenges reported by past studies in the area, enumerating opportunities we identified from the review of the state of the art.

3.1 Toxicity and related concepts

When studying toxic communication in news, terms and definitions of negative conversation differ. Authors define this kind of harmful messages as toxic, hate, offensive or uncivil speech. In order to correctly detect toxicity-generating news, we first need to understand the differences between these terms and how they overlap. With this goal in mind, in the current subsection, we present the definition of several terms used in relevant studies and describe what we consider to be toxicity.

The first issue to address is the subjective nature of defining each of the terms mentioned. This fact poses as much of a problem for machines as for humans. Despite the nonexistence of a formal and universal interpretation, hate speech is considered as *"any communication that disparages a person or a group on the basis of some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics"* [59]. There are no established directives on how the definitions of hate and offensive speech differ. Despite the absence, there is a consensual opinion that hate speech targets disadvantaged social groups [8]. Regarding social media policies, the YouTube code of conduct determines that a post is promoting hate when *"it encourages violence against individuals or groups based on the attributes such as age, disability, ethnicity, gender, nationality, race, immigration status, religion, or by dehumanizing individuals or groups by calling them subhuman, comparing them to animals, insects, pests, disease, or any other non-human entity"* [80]. Twitter classifies hateful conduct as the promotion of *"violence* *against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease"* [75]. A survey on hate speech detection [55] considers that hate speech is typically plotted as *"criticizing an individual or a group based on some common things it may be race, complexion, civilization, gender, sexual orientation, nationality, religion, or other trademarks"*. Another study [49] considers hate speech to be the denigration of a group of people. Supplementary work [28] considers hate speech as *"any expression that is abusive, insulting, intimidating, harassing, and/or incites to violence, hatred, or discrimination. It is directed against people based on their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth"*. Researchers [27] emphasize the differences between directed and generalized hate speech, focusing their attention on the target of hate speech rather than the distinction of its presence. Nevertheless, the authors define the term as any conversation that diminished an individual due to their characteristics.

The vast amount of opinions on what determines a discourse to be hateful has demonstrated a consensus on the fact that it is an attack to a minority, although demonstrating the classification to be subjective as well. The challenge to establish the boundaries between free speech and what is acceptable has proven to be thin and hard to find [14]. Research found [3] differentiates between hate and offensive speech. They define hate speech as the language used to express hatred directed to groups of individuals to depreciate, humiliate or insult their members. In counterpart, offensive speech is defined by the authors as the communication that might have deprecatory words towards a group. However, it is used in a different qualitative way, that is, the context is considered. Other research [38] separates offensive speech into three categories, namely: vulgar, pornographic or hateful, being the last one the offence towards one's race, religion, country. When studying offensive comments [22], these are described as any communication that aims to outrage one or more individuals. Considering that hate speech belongs to this type of comments, as well as bullying, profanity, and harassment.

"The definition of incivility has been long debated by communication scholars" [5]. The unanimous agreement on the definition of incivility poses a challenge since cultural background contributes to subjectivity [48]. Regarding online incivility, a work found [7] defines the term as *"a manner of offensive interaction that can range from aggressive commenting in threads, incensed discussion and rude critiques, to outrageous claims, hate speech and harassment"*. Other researchers [5] 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"*. Research on incivility [48] uses the definition of the term by another work [5] adding that, despite hate speech and incivility share similarities, the terms are not strictly depending. The affirmation meaning that uncivil speech is only considered as hate speech when attacking a minority or individual.

Despite on the one hand the vast amount of research on hate, offensive and uncivil speech, on the other hand, toxicity in news has been less explored compared with the prior terms, despite their similarities. When studying toxicity triggers in online discussions, a study [4] mentions that little research has been associated with it. The authors, although not establishing a definition, refer to toxic comments as any communication that hinders user interaction and opinion expression by using rudeness or harassment, culminating in a distasteful experience. Other researchers [29] consider toxic comments to be any communication having violent, rude and disrespectful behaviour, that can make use of personal attacks, harassment or bullying in order to cause any participant to leave the conversation. Although not having a definition to toxicity, another study [79] refers to the term as any negative behaviour that uses personal attacks and various forms of online harassment (e.g., name-calling, threats of violence, discrimination) to impact the health of the community the victims find themselves in. Additionally, in their study [48] researchers include the toxicity term in the guidelines provided to annotators when determining uncivil comments. The authors consider toxicity to be any communication that impels the reader to leave the conversation.

3.2 Our definition of Toxicity

Taking in to account the vast amount of definitions on each of the terms, and with the fact that many of them overlap, sharing similarities inter-terms and intra-terms, we take into account such similarities when providing our definition. We take into consideration various definitions [40; 29; 4] and consider toxic speech as any aggressive and rude communication that attacks a group or individual based on any of its characteristics (taking into consideration the context of the conversation). Being the attack an attempt to disrupt and harm the reader and its integrity, disrupt any logical and civilized discussion or to encourage the victim's abandonment of the conversation due to a harsh environment.

3.3 Previous Approaches and Results

Regarding what has already been done, the current section divides previous approaches into two subsections. The first, regarding work developed by researchers in the area of negative behaviour prediction in news, that has proven to be useful on the development of our work. The second subsection refers to studies associated to news classification that, although not directly related to our area of study, shares similarities to the work we developed, guiding us on standard practices related to news articles classification and feature extraction. All references to measures, feature extraction techniques and algorithms mentioned in this chapter, can be further understood in Chapter 2, where we explain each of the matters.

3.3.1 State of the art on news toxicity prediction

With the vast increase of social network usage, so does increase the human interactions in those online platforms, providing the means for an increase of rough communication. A 2014 Pew Report notes that 73% of adult internet users have seen someone harassed online, and 40% have personally experienced it [26]. Acting to combat such behaviours, researchers have turned their efforts to the identification of the various forms of negative communication enumerated in Section 3.1. Although the vast amount of research in areas such as hate speech classification, work done on the prediction of news that generate such negative behaviours is narrow, being this fact a significant motivation for our work. To our knowledge, the work described in this thesis is the first of its kind in the prediction of toxicity-generating news. We next describe studies related to the area of our work, that have guided us through the process.

A relevant work [11] used a feed-forward neural network to predict incendiary news articles. Instead of defining the term themselves, the authors adopted the annotation guidelines developed in the work of a non-governmental organization *Nefret Soylemi* (hate speech) directed by Hrant Dink. Participants in this project manually annotated Turkish news articles as incendiary. In their work, *Nefret Soylemi* annotators considered incendiary articles as those who ignite hatred. From this project, the authors collected 1,036 incendiary articles from various Turkish news media. In order to gather non-incendiary news articles of the same topics as the ones labelled as incendiary, the authors resorted 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 experimented using a linear Support Vector Machine (SVM), Naive Bayes and a feed-forward neural network with Turkish word embeddings, word and characters n-grams as features. Results from experiments with different features and models showed that the feed-forward neural network using word unigrams performs best, reaching an F1 score of 0.95 for cross-corpus classification.

One related work [48], used a model composed of a logistic regression classifier to predict whether a news article will provoke uncivil comments. The authors considered a comment to be uncivil if it contained toxicity, a personal attack or aggression, terms also presented in work done in [79]. In order to achieve their goals, the authors constructed a dataset of political news articles from the conservative news website Breitbart and liberal website Politico. The collection process resulted in 15,000 articles from both sources, containing the articles themselves, date, title, and reader comments. Additionally, the authors made further use of the publicly available talk page comment corpus. This dataset contains over 100,000 comments from Wikipedia users. The dataset is used to train their model on the prediction of uncivil comments by identifying one or more of the three previously mentioned terms. Following the prediction of uncivil comments, the authors labelled an article as uncivil speech provoking. An article with a higher frequency of uncivil comments than the sources median uncivil frequency is labelled as uncivil speech provoking as demonstrated in Formula 3.2. Such measures are also used in our work, since knowing the sources median incivility frequency allows the division of the data in a balanced manner.

$$W_A = \frac{1}{n}\sum_{i=1}^{n} I_{c_i} \tag{3.1}$$

$$U_A = \begin{cases} 1 & \text{if } W_A > \text{Median incivility} \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

The authors trained the logistic regression classifiers to predict the uncivil speech classes of the news articles from the two sources, with an 80-20 train-test split, using term frequency-inverse document frequency (TF-IDF) of bi-grams as features extracted from articles. Results show the model achieved 64% Accuracy, with 62% Precision and 66% Recall.

Another work done [21] uses a logistic regression classifier to predict news articles to be prone to generating incivility, providing more detail in their approach in [81]. The authors follow the incivility definition of other research [16] which defines the term as *"as features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topics"*. For the collection of data for their experiment, they use Facebook pages of German newspapers outlets to gather data, focusing on newspapers with most followers. The process resulted in a total of 27,728 news articles and 1,056,002 comments, which they further reduced to 10,170 comments. News articles were obtained from Tagesschau, ZDF Heute, N24, RTL Aktuell, Die Welt, Sueddeutsche Zeitung, BILD, Spiegel Online, and ZEIT Online. As features, the authors used the comments themselves, the total number of words present in the comments, number of likes, the class of incivility of the comments (none, scattered or predominantly/exclusively uncivil), word uni-grams in the comments, two to four n-grams of Part-of-Speech (PoS) tags and a sequence of repeated punctuation (e.g., "!!!"). A total of 4,500 features were collected for each comment and passed to a logistic regression classifier, in order to predict the class the article would fall into. The model was trained using 5-fold cross-validation, not mentioning the percentage partition in train and test data. Results show the model achieving an accuracy of 72.4%, with an F1-macro score of 0.44. In order to improve the poor performance, the same model was used, switching the classification to a binary label (civil or uncivil). With this modification, the model performance increased, achieving an accuracy of 78% and an macro-F1 score of 0.68.

A similar work [10] used an SVM and Convolutional Neural Network (CNN) to predict the news probability (high or low) of inciting hate speech. Intending to achieve his goal, the author constructed a dataset of Twitter news article posts and respective tweet comments as well as news articles extracted from news media websites and subsequent comments from the respective comment sections. The news tweets were extracted between 2018-12-27 and 2019-01-14 (14 days) collecting a set of 73,035 Tweets in which 72,342 had a news piece associated with 1,734,652 comments from Twitter users and 933,480 comments from comment sections of news media websites. In order to classify news as probable of generating hate speech, the author used the mean value for the hate percentage present in comments relative to the news. News articles with a percentage of hateful comments higher than the mean percentage (17%) were labelled as having a high probability of generating hate speech replies, and low probability otherwise. The author experimented with SVM and CNN. Regarding the SVM approach, the author trains the model with 10-fold cross-validation using a 75-25 train-test split, using GloVe word embeddings for feature extraction from news leads, titles and text. The best performing model for the SVM approach reached an average F1 score of 0.54, representing an ensemble of three SVM classifiers trained using different features extracted from news leads, the news titles and meta-information regarding the news article. Regarding the CNN approach, the best performing model reached an F1 score of 0.61, representing a CNN and Long Short-Term Memory (LSTM) sequentially arranged, concatenating features extracted from news titles.

The work developed in this thesis is distinct from the studies described in this section. Differently from studies researching incivility [48; 21], we study the phenomenon of predicting toxicity-generating news, using the *Perspective API* to classify toxic comments. We focus our work more in the classifications of news, rather than the classification of its comments. Differing from the study researching incendiary news [11], we are not interested in such predictions, but the prediction of toxicity-generating news, and what characteristics influence the generation of toxic behaviour in readers. Unlike the research of hate speech inciting news [10], we are not interested in the prediction of hate speech generating news, but to the prediction of toxicity-generating news. Nevertheless, these studies have influenced our work, enlightening us with definitions, metrics, algorithms, features, datasets and baselines.

3.3.2 Work in related areas of study

With the goal of surveying classification models, features extraction techniques and difficulties associated with the area of news classification, in this section we mention the studies that provided us with a good description on those maters, contributing with direction and ideas on our work.

One work [18] used an SVM to classify news into various groups based on textual features. The authors used a total of 5,070 articles from two BBC datasets and a 20Newsgroup dataset, belonging to five different classes (computer graphics, objects offered for sale, baseball, Christianity, and political texts on guns). The performance of the classification model developed, ranged from an F1 score of 0.91, reaching 0.97 for the baseball class in the 20Newsgroup dataset and 0.99 F1 score for the same class regarding the BBC datasets. In their study, the authors used the TF-IDF algorithm to extract features from news articles. With this feature in mind, they provide a useful description to why they opted to use SVM compared to a Naive Bayes classifier. The authors explain that although Naive Bayes is a simple and efficient classifier, it does not model texts well. In counterpart, SVM does well with high dimensional data, which is the case when using TF-IDF. Despite this fact, the authors mention that time complexity analysis and non-use of standard datasets create problems in this algorithm. Moreover, the authors provide a good description of the steps of text classification, from text preprocessing, feature extraction and classification.

Another approach [74] developed a classification model to classify Azerbaijani news articles from two separate datasets, composing a total of 200,000 previously labelled articles from various newspapers. In their work, they started by using a Bag of words (BoW) for the textual representation of news with Naive Bayes to test their model with 10-fold cross-validation with 90% of data for training and the remaining 10% for testing. In order to select the categories to which their model would label news, the authors used the K-Nearest Neighbours (KNN) in order to optimize the number of categories, which reduced 48 categories to 8. The final categories used were society, world news, current events, sports, economics, entertainment, arts and politics. Using those categories as labels, the model reached an average F1 score of all categories of 0.68. After a first experiment, the authors moved to remove stop-words, which in turn increased the model performance to an average F1 score of 0.71. The authors experimented the same approach with an SVM, motivated by the fact that this classifier works well with text classification due to its ability to handle large number of features efficiently and effectively. Using the SVM, the model reached an average F1 score of 0.79. Removing stop-words, the model reached the highest performance of average 0.84 F1 score. Finally, the authors tested their model with a multi-layer perceptron classifier, which reached an accuracy of 86.3%. Despite having good performance with the neural network, the authors point out one difficulty of text classification with this kind of classifiers, this being its problem of feature space dimension. Since neural networks have higher computation and time complexity, this fact led to the need for feature space reduction techniques.

A research [31] used machine learning algorithms to classify fake news. In their approach, the authors focused their attention on using a large scale dataset with a lower quality of its label annotations, rather than having a smaller dataset with high-quality annotations. In order to achieve a large dataset, the authors collected tweets from trustworthy and untrustworthy sources, labelling tweets according to its source, resulting in 401,414 labelled news tweets. The authors provided a good description of what features they extracted, by dividing their them into groups: user-level features, tweet-level features, text features, topic features and sentiment features. From user-level features, 53 features were extracted such as the number of followers, frequency of tweets and ratio of retweets. Tweet level features are comprised of information such as the number of retweets, weekday, time and word count, resulting in 69 features. For text features, the authors explored BoW using TF-IDF representations of words and Doc2Vec word embeddings trained on the corpus. Topic features were extracted using a Latent Dirichlet Allocation model as well as a Hierarchical Dirichlet Process model. As for sentiment features, the authors used SentiWordNet to extract the polarity of the tweets, as positive, negative or neutral. The authors compared their model performance with various classifiers, namely naive bayes, SVM, decision trees, neural networks, and ensemble methods such as random forest and XGBoost. When comparing results with distinct features and models, the authors concluded that providing the model with user information improved its performance. The best performance is reached using XGBoost with an F1 score of 0.77 when considering only tweet features and an F1 score of 0.90 when considering tweet and sources features.

Another study [43] reviewed the algorithms and process workflow for news category classification. The authors described the process steps of news classification such as news collection, preprocessing, tokenization, stop-word removal, word stemming feature selection and classification. The authors also presented a good description of various text mining feature selection techniques such as Boolean weighting, information gain, TF-IDF, class frequency thresh holding. Furthermore, the study presented a good description of several widely used classification algorithms, naive bayes, SVM, neural networks, decision trees and KNN, mentioning the advantages and disadvantages of such classifiers. In their review, the authors pointed out that naive bayes worked equally well on textual and numerical data, although previous cases reported lower accuracy due to incorrect parameters. The authors mention that, although the KNN is a simple, effective and non-parameterized algorithm, it requires a lot of classifications time, with the identification of an optimal k value of clusters being a difficulty.

A study [24] classifies Sri Lankan news articles into 12 distinct groups namely, war-terrorist-crime, economy-business, health, sports, development-government, politics, accident, entertain, disaster-climate, education, society and international. The authors used a BoW with the frequency of the words. In order to reduce data dimensionality and disregard overly frequent words, words with low and high frequencies were removed. The classification process was carried out using an SVM, applying 10-fold cross-validation with 90-10 train test split. The authors explained that they chose to use the mentioned classifier since it can handle well high dimensional data. The model developed reached an average F1 score of 0.78 for the classification of all twelve classes.

The articles described in this section, although not being directly inserted in our area of research, provided us with a general idea of primary techniques used in the classification of news, regardless of the topic. With the review of these studies, we could conclude that naive bayes and SVM are two significant algorithms used to classify news, shedding light on the challenges each classifier brings. Nevertheless, the analysis of these studies reported other machine learning algorithms, such as logistic regression and random forest that motivated their use on our research. Despite the superior results of neural networks reported in the articles mentioned in this review, we opted to use classical machine learning algorithms, since these methods provide a more manageable task of analysis of results. Furthermore, the articles mentioned provided a good description of the features used in the majority of news classification tasks. This description influenced us to try and use not only features from text present in news titles but also features containing meta-information, like newspaper country of origin, time of publication.

3.4 Difficulties in classifying toxicity

This section provides an overview of obstacles encountered during the task of classifying toxic speech or any related term previously presented. Since the classification task is complex and challenging, it becomes relevant to survey the difficulties pointed out by past studies. Therefore, the following list enumerates the complications encountered:

Since every language is different, and even within the language itself, there are differences, such variations pose adversity upon detecting toxicity or any other term efficiently. Although our research only addresses English news articles, our review of the state of the art includes not only the English language but also studies using German, Azerbaijani and Turkish. The variation in language has always posed a challenge to the field of NLP and will continue to be a challenge to classifiers, leading to the need for the development of solutions robust to such variances [54].

With social networks being multi-lingual platforms, a journalist's personality encloses its cultural environment and consequently, contributes to the newspaper cultural background. The newspaper cultural background poses a difficulty since the influence of the cultural context might expose different behaviours, leading to variations that difficult the classification task of machine learning algorithms [21].

The variation of the context on which news appear poses a difficulty, since news articles in a platform may move users to behave in a more toxic manner than readers that use others platforms to consume news articles. This way, the variation of the environment context between news from a classifiers train set and test set affects the result [48].

As different newspapers have their distinct language models some newspapers strife to a more formal and scientific language, others are more directed to fast and straightforward reading models. The variations on language models may pose a barrier to classifiers. When comparing their model with cross-corpus experiments, the variation resulted in a lower performance [11].

The overview of the state of the art of classification of toxicity and related concepts enabled the conclusion that text classification is a challenging task due to the high dimensions of the data involved. This fact also poses the problem of a higher time and computation complexity [18; 74].

3.5 Opportunities in toxicity prediction in news

The current subsection describes possible points of interest that may cover some accessed needs regarding the field of toxicity prediction.

With researchers having to construct datasets for their research, developing and establishing a public dataset that can be used for the investigation of a specific phenomenon may accelerate the progress in the area. This way, researchers may focus their time and efforts on the development of new innovative approaches, and compare their work with others using the same data.

The provision of guidelines for uniformization of the definition on terms such as hate, offensive or uncivil speech may aid investigators to convergence on a more homogeneous definition of such terms. This action would consequently transpose to the models, leading a more effective identification of the phenomenon.

Approaches investigating the detection of negative behaviour have not been restricted to the English language. The development of solutions for the detection of toxicity-generating news in other languages commonly used such as Portuguese, French and Spanish is also required.

3.6 Conclusions

Reviewing the state of the art regarding the classification of toxicity-generating news, we first clarified the differences between concepts that appear in a similar context as toxicity, trying to understand how the area of study has unfolded. Concerning the definition of toxicity, we concluded that it had been debated in several studies from various fields of research, from social studies to online behaviour classification in social platforms. The definition of other similar terms such as hate speech, incivility and offensive speech had also been extensively studied in related areas of research. Despite the existence of such studies, with the small amount of work in our area of research, it is crucial that we provided our definition of toxicity since the definitions described in past studies did not appear in the same context as our work. This way, we required a precise and transparent definition that applied to our model, which we provided in this chapter.

Additionally, we delivered a detailed description of past studies that research similar problems to the researched in this thesis. Since there is little to no work done regarding the prediction of toxicity-generating news, we divided the description of past studies into two subsections. The first, providing studies similar to our topic, prediction of toxicity-generating news. The second, related to studies that describe the process and techniques used in the area of news classification. We decided to add such studies since there was little investigation in our research topic, and because our work also focuses on news articles classification. This way, we could make an informed decision on the techniques and features widely used in the news classification paradigm. Table 3.1 provides an overview of the best performing models of the studies present in our review of the state of the art.

Regarding studies similar to our area of research, it could be concluded that, to our knowledge, no work had been done in predicting news from generating toxicity. This fact makes our work a novelty in the study of the phenomenon of prediction of toxicity-generating news. Regarding the studies presented in the second subsection, it could be concluded that as for the models used, SVM and naive bayes were a superior choice for researchers when classifying news, regardless the classification task in hand. It could also be seen that neural networks achieved better results compared to machine learning algorithms. Despite the report of better results, we decided that our work would make use of classical machine learning algorithms as it would facilitate the interpretation of results and features. With the review of news classification studies, it could also be concluded that the majority of studies in the area of news classification made use of text features, such as the TF-IDF presentation of text or BoW.

Finally, in this chapter, we also describe some difficulties and opportunities identified during the review. We detected that when using text features, a significant adversity was the sparse and vast dimension of the features of this nature. The high dimension of features will, in turn, affect the performance of the model chosen by researchers to the classification task, since some models handle better the problem of high dimension. Another difficulty is the variation of the context in which data is comprised. With the review made on past studies, it could be seen that not having a data well suited to the context of the research in hand, compromises the model's performance, since the data might not be related to the same context as the one being studied. Analysing what has already been done, we could conclude that the prediction of toxicity-generating news had not been previously studied. With this said, our study gives researchers the possibility of analysing the characteristics in news that motivate the propagation of such behaviours. Based on the conclusions mentioned in this section, we decided to focus on the analysis of the characteristics of news that will generate toxic behaviour. In the following chapter, we describe the collection of data and provide a description of the data used in our research.

Table 3.1: Data, year, research, features, algorithms and metrics of the papers presented in the review of the state of the art.

Paper	Year	Research	Dataset	Features	Algorithms	Precision	Recall	Accuracy	F1
Similar areas of study									
[48]	2018	Prediction of incivility-generating news	15,000 political news from Breitbart and Politico	TF-IDF and bigrams	Logistic regression	0.62	0.66	0.64	0.64
[11]	2019	Prediction of incendiary news	2,074 news from Turckish news media outlets	word uni-grams	Feed forward neural network	0.77	0.94	-	0.95
[21]	2018	Prediction of uncivil news	27,728 news from German news media outlets	news meta-info, word uni-grams, two to four n-grams POS tags, puntuation	Logistic regression	-	-	0.78	0.68
Related areas of study									
[18]	2016	Classification of news	5,074 news from two BBC datasets and 20Newsgroup dataset	TF-IDF	SVM	0.95	-	-	0.96
[74]	2018	Classification of news	200,000 news from various Azerbaijani newspapers	BoW	SVM	0.84	0.84	-	0.84
[31]	2018	Classification of fake news	401,414 news from various English newspapers	TF-IDF, BoW	XGBoost	-	-	-	0.90
[24]	2013	Classification of news	non speficied number of news from Sri Lankan newspapers	BoW	SVM	0.91	0.68	-	0.78

Chapter 4

Toxicity prediction dataset for news

With the review of past studies provided in Chapter 3, we could conclude that there has been no work done on the area of prediction of toxicity-generating news. Despite the existence of public datasets related to the area of classification of offensive speech and incivility, we decided not to use them. This decision was based on the fact that such datasets were collected in other contexts (e.g., politics forums) or to research other phenomenons (e.g., offensive communication on Wikipedia forums), thus not well suited to our specific work on detecting toxicity-generating news. This way, we tried to use a dataset as close to the phenomena we are researching. From this reasons, in the present chapter, we describe the data used in our work and the collection process, done in a prior phase of the *Stop PropagHate* [2] project in which our work is inserted in. Further along in this chapter, we provide a description of the classification process of toxic comments, the labelling of news articles accompanied by an analysis of the results from these processes.

4.1 Data collection

With the intent of constructing a dataset to be used in the identification of hate speech generating news on Twitter, the author [10] uses the Twitter Stream API to collect news tweets and respective tweet comments. This tool was used as a first stage to seek tweets by news media accounts and associated comments. In this stage, the author obtained the news article URL, allowing the extraction of the news body. In conjunction with the body, the author also gathered information about comments belonging to the comment section of the news media website. The process of extracting data occurred between 2018-12-27 and 2019-01-14, producing a total of 73,035 tweets in which 72,342 had a news piece associated, 3,026,270 comments from Twitter users and 933,480 comments from comment sections of news media websites. The author gathered information from English newspapers, namely from the United States and the United Kingdom. Portuguese newspapers were also considered, namely from Brazil and Portugal. In an effort to select relevant news outlets from the mentioned countries, the author used cited entities in Reuters Institute's

Table 4.1: English language newspapers chosen for news articles data extraction.

British (UK) newspapers	American (USA) newspapers
BBC News	ABC News
Daily Express	BBC News
Daily Mail	BuzzFeed News
Huffington Post	CBS News
ITV News	CNN.com
Metro	Huffington Post
Mirror	NBC/MSNBC News
Sky News	New York Times
The Guardian	NPR News
The Sun	The Wall Street Journal
The Telegraph	The Washington Post
Times	Time
	USA Today
	Yahoo! News

Digital News Report2017 [57]. The author additionally added news outlets that have a large presence and support on Twitter. Regarding English language outlets, Table 4.1 lists the final selected news outlets from the United States and United Kingdom to perform data extraction on. Regarding Portuguese news outlets, Table 4.2 provides a list of the final Portuguese language newspapers considered for the data collection task. We encountered some data-related issues not identified during the collection phase. We could examine that news articles from the "New York Times" and "CBS News" had no Twitter comments linked to them, despite the existence of comments from such newspapers. Furthermore, we could identify that news from "Time" newspaper had all their titles miss collected as 'Time'. With these problems, articles and respective comments were excluded from the data we use in our classification models, but we used them to analyse toxicity in news articles related to those newspapers.

With the interest of studying the relation between news articles and the responses from users, the author developed a strategy to extract comments from news articles of the previously mentioned news sources. This way, during the data collection period previously specified in this section, the author monitored the news sources specified, gathering news article tweets. From these news pieces, user replies were captured, representing comments to a related news article. Since Twitter has a limit of 240 character length, it is a regular habit from news sources to tweet the news title followed by the URL that leads to the full article. In order to gather more data, the author scraped news websites extracting information about the news body, metadata and comments. The author additionally extracted information about user profiles. We do not use this information in our work.

Concluded the data gathering phase, the author proceeded with data processing. Since data from social networks, especially in Twitter, are unstructured and informal, data cleaning is a crucial step in processing data. For this reason, the author performed data cleaning, removing URLs, mentions, hashtags, unicodes, symbols and emojis. With data cleaned, the author aggregated

Table 4.2: Portuguese language newspapers chosen for news articles data extraction.

Brazilian (BR) newspapers	Portuguese (PT) newspapers
BBC News Brasil	Correio da Manhã
Carta Capital	Diário de Notícias
El País Brasil	Expresso
Época	Jornal de Notícias
Folha de S. Paulo	Observador
IG Último Segundo	Público
Jornal O Globo	RTP Notícias
O Estado de S. Paulo	SIC Notícias
Portal G1	TSF
Portal IG	TVI Notícias
Portal R7	Visão
Revista Isto é	
Revista Piaui	
UOL Online	
Veja	
Yahoo! Notícias	

data into three groups: news, Twitter comments, and website comments. Fig. 4.1 presents the composition of each collection in more detail.

From the three collections represented in Fig. 4.1, in our work we only make use of the *Comments_tw* and *News* collections. This way, making our analysis of the distribution of news by country (Fig. 4.2) we can see that the majority (24,987) is composed of British articles, with 18,318 news articles relative to American newspapers. Portuguese news are evenly balanced between Portugal and Brazil having 11,168 and 10,054 news articles, respectively. Comparing the number of news from each newspaper, we could examine in Fig. 4.3 that the number of news per news sources is very unbalanced. We can see that there is a significant gap between the number of news from the "Independent" and "Daily Express" newspaper. From this same distribution, we can also examine that there is a large discrepancy on the number of news relative to newspapers of the same country of origin. In particular, we can see that "The Telegraph" is the newspaper with the lowest number of news, which allows verifying that news from UK sources have a very unbalanced distribution. This fact may pose a difficulty our model since research, as mentioned in Chapter 3 has proven that the context and culture of which news articles are inserted, represents a challenge to classifiers. This way, having data that has contextual and cultural variances with such a high discrepancy in its quantity, may hinder the classification performance.

Much like the distribution of news by newspapers, the distribution of comments by newspapers (Fig. 4.4) has a similar unbalanced property. From this distribution, we can see that the "CNN" as a large gap to the second newspaper with most news comments ("NBC news"). The significant variance in the number of comments could also pose as a difficulty to the classification task since toxic comments have increased weight in news articles with fewer comments.

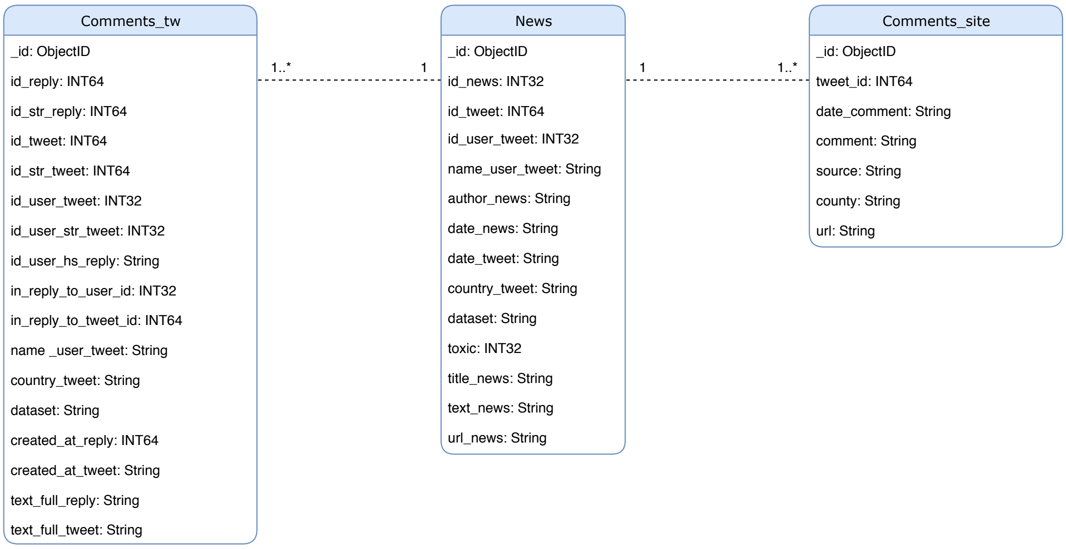

Figure 4.1: Dataset files description.

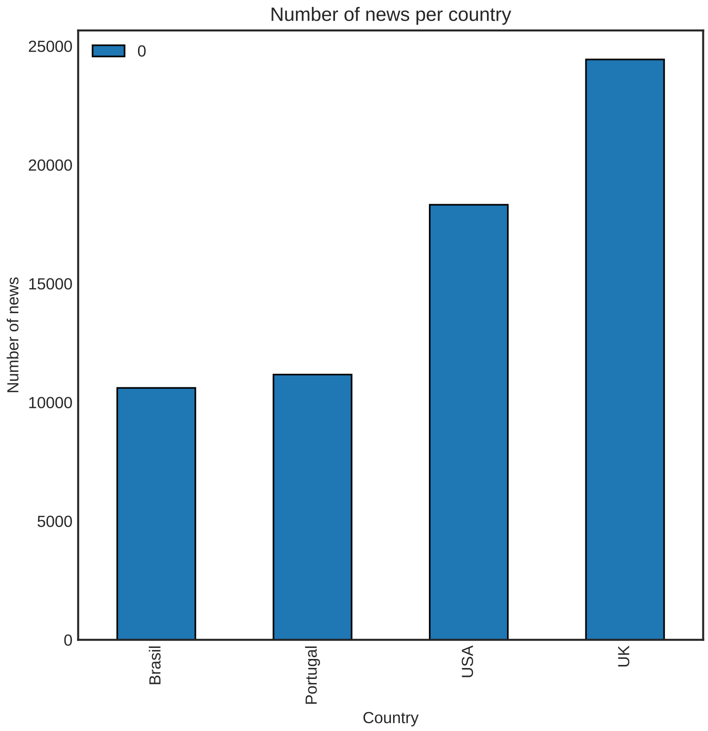

Figure 4.2: Distribution of news articles by news sources country.

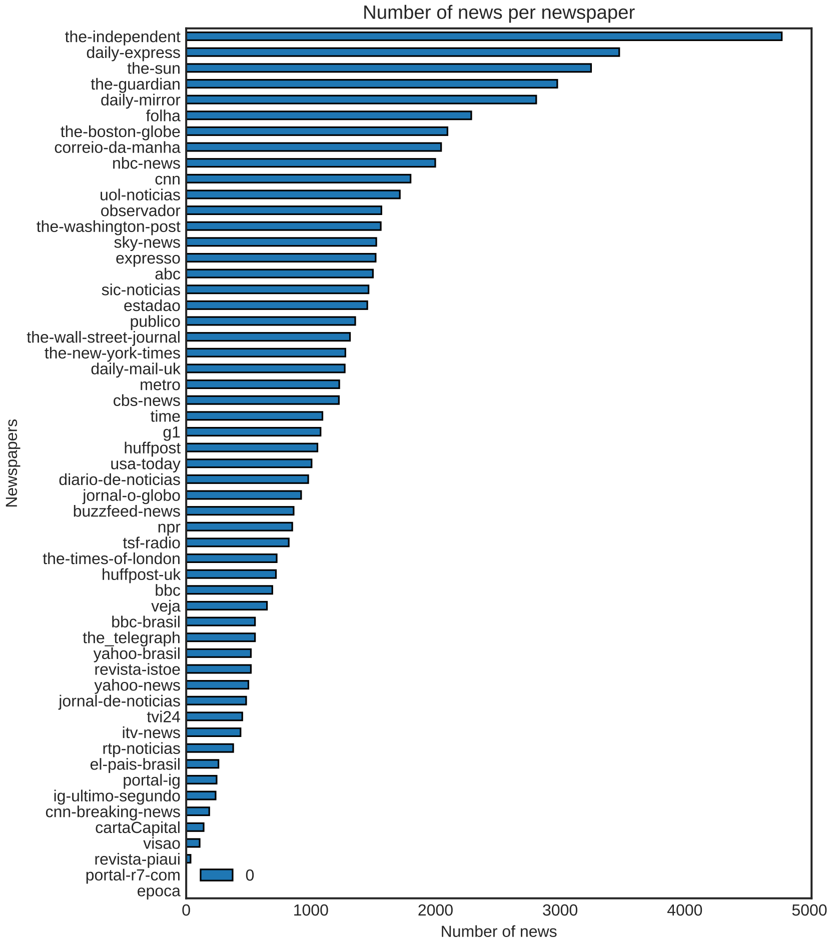

Figure 4.3: Number of news articles distribution by newspaper.

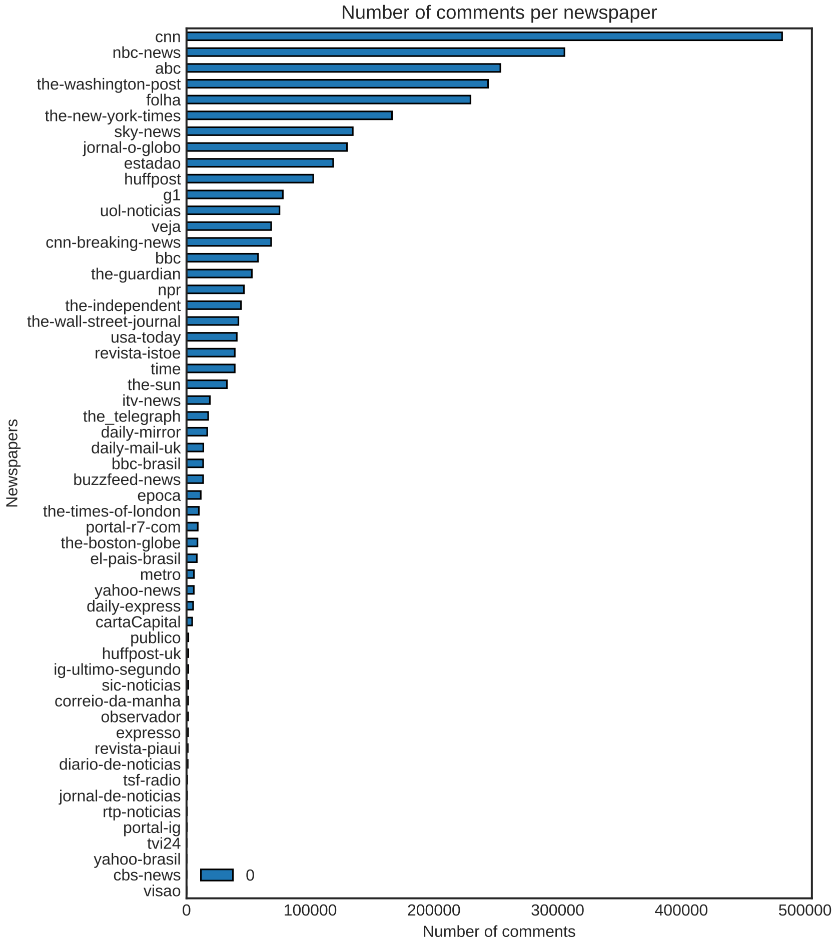

Figure 4.4: Number of comments distribution by newspaper.

4.2 Comment classification

As our work aims to understand how news articles provoke toxicity, we first need to classify comments related to those news articles. For this purpose, we make use of the *Perspective API* classifier. *Perspective API* [39] was developed by Jigsaw and Google's Counter Abuse Technology team in the Conversation-AI project [1]. The API uses numerous machine learning models to compute the score, ranging from 0 to 1, of the effect (e.g., toxic) a comment might have on a conversation. The model developed uses a CNN with GloVe word embeddings [64] using data from sources such as The New York Times.

We chose to use this classifier due to its recent popularity, being used at OffensEval-2019 by The Conversation-AI team [62] for the classification of toxic posts, having a generous F1 score of 0.79, without any need for additional training and tuning of the classifier. Furthermore, we chose to use *Perspective API* since we did not want the main focus of our work to be the classification of toxic comments, but the prediction of toxicity-generating news.

In order to identify various effects comments may have on discussions, *Perspective API* provides a diverse set of attributes/labels [40] that are suitable for several languages, namely Portuguese. For this reason, we selected attributes that were also supported for Portuguese, since we had the objective of classifying the comments from Brazil and Portugal as well. The following list provides the description provided by *Perspective API*, of the final selected attributes we use to classify comments on:

- **Toxicity**: "rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion".
- **Severe toxicity**: "very hateful, aggressive, disrespectful comment otherwise very likely to make a user leave a discussion or give up on sharing their perspective".
- **Identity attack**: "negative or hateful comment targeting someone because of their identity".
- **Insult**: "insulting, inflammatory, or negative comment towards a person or a group of people".
- **Profanity**: "swear words, curse words, or other obscene or profane language".
- **Threat**: "describes an intention to inflict pain, injury, or violence against an individual or group".

From the analysis of the number of comments per country (Fig. 4.5), we decided to start the classification of comments relative to news articles from USA sources since those represented the majority of English comments. We followed with the classification of British comments. Concluded the classification of all English comments, we proceeded towards the classification of comments from Brazilian news articles, finishing the classification task with comments from Portugal. The classification of American comments took place between 2020-02-17 and 2020-03-17 (29

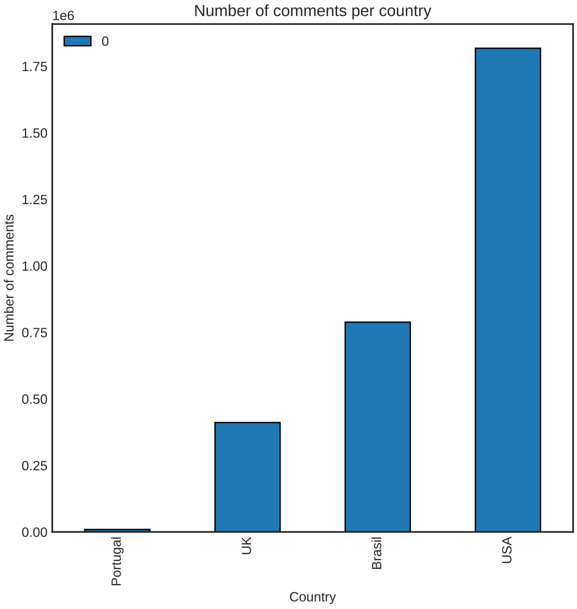

Figure 4.5: Number of comments distribution by country of origin.

days) and British comments between 2020-03-17 and 2020-03-30 (13 days). Brazilian comments were classified between 2020-03-31 and 2020-04-23 (23 days) with Portuguese comments being classified between 2020-04-24 and 2020-04-29 (5 days).

Having finished the comment classification process, we started by analysing the results from the API. We consider comments to be toxic if the score of the *Toxicity* attribute returned by *Perspective API* is greater or equal than 0.5. With this metric imposed, the comments used in this study are composed of 437,494 comments signalled as toxic and 1,625,100 signalled as non-toxic. From toxic comments, 352,824 are comments relative to USA news articles and 84,670 comments to news articles from UK newspapers. From non-toxic comments, 1,298,598 are comments relative to USA news articles and 326,502 comments to news articles from UK newspapers. Analysing the toxicity present in comments, we could see from Fig. 4.6 that news sources such as "The Sun" (20.8%), "The Huffpost" (20.1%) and "Buzzfeed" (19.8%) represent the three newspapers with a higher percentage of toxic comments. The presence of "The Sun" on the top and being "The Guadian" the newspaper with least toxic comments, suggests that the culture of the tabloid and the type of content produced can influence its consumers' responses and the behaviour they could have. Analysing the news toxicity histogram in Fig. 4.7, we can see that the majority of English news articles (14,237 articles) have less than 10% of their user comments signalled as toxic. Curiously we can also analyse that 780 news articles with the vast majority of comments (90%) being toxic. Finally, we calculated the median toxicity from all English news, 11.1%. From this value, we conclude that in median news articles have 11.1% of disrespectful or unreasonable comments. We also use the median comment toxicity to split our data in a balanced manner.

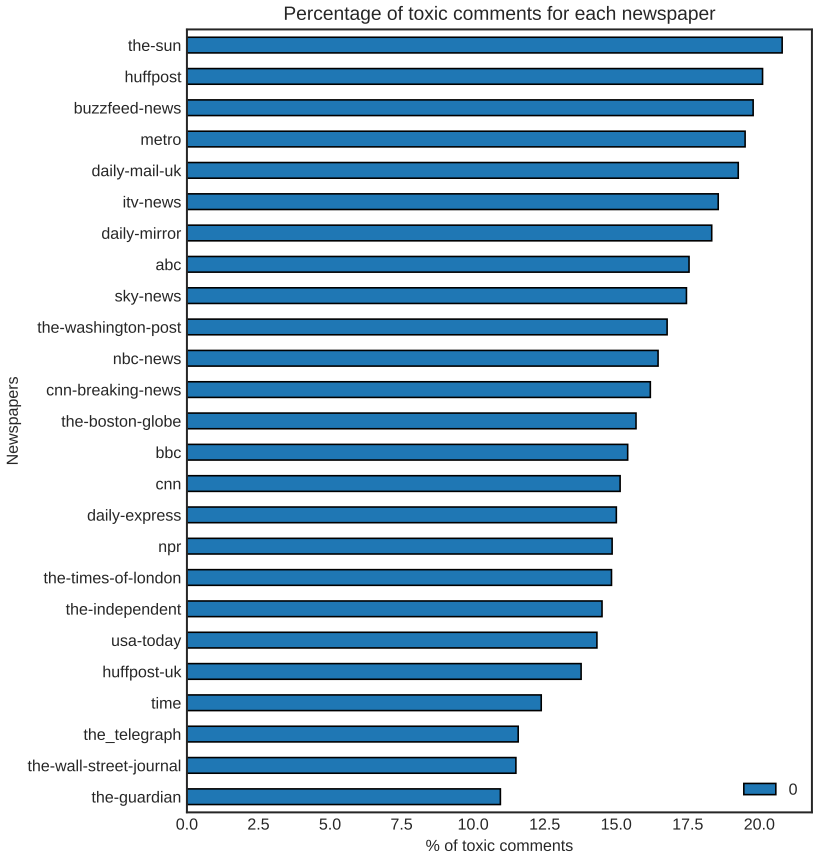

Figure 4.6: Percentage of toxic comments in Twitter reactions for each newspaper.

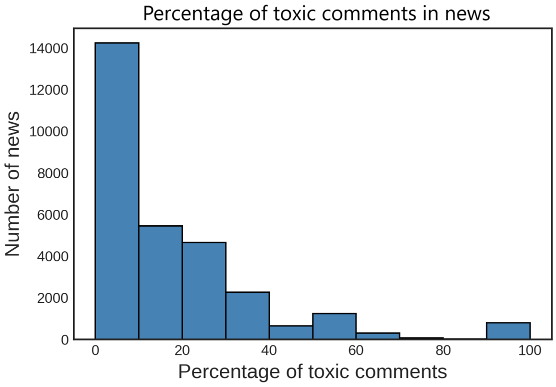

Figure 4.7: Percentage of toxic comment tweets relative to news articles.

4.3 Labeling toxic news

Having classified all comments relative to English news, and having the median comment toxicity (11.1%), we labelled news articles as toxicity-generating T_N, if the percentage of toxic labelled comments relative to that article P_N, is equal or higher than the median comment toxicity percentage, as shown in Equation 4.1. This classification metric was inspired by the work developed on the prediction of incivility-generating news [48].

$$T_N = \begin{cases} 1 & \text{if } P_N \geq \text{Median toxicity} \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

Following the mentioned news labelling metric, we could examine from Fig. 4.8 that almost half of the selected newspapers (11 newspapers) have a percentage of toxicity-generating news above 50% with “CNN Breaking News” having 74% of its news being toxic generating. Following this trend are the “Huffpost” (72.1%), “The Washington Post” (69.5%), “ABC” (69.5%), “CNN” (68.2%), “Sky News” (67.1%), “NBC News” (66.7%), “ITV” (60.6%), “USA Today” (60.3%) and “NPR” (52.7%). The plot of the toxicity-generating news in newspapers allows us to understand that the number of toxic impelling news is not directly proportional to the number of news a newspaper has. The “CNN Breaking News” is both the newspaper with the least number of news and the newspaper with the highest toxicity percentage.

On the other hand, “The Independent” is the newspaper with most news articles and a toxicity percentage of 33.1%. This fact allows us to conclude that the content, context and culture of the newspapers are contributors to the generation of toxicity in news. This might be due to the fact the distinct newspapers may create their content to their most valued customer type, redirecting valued news to the target audience preferences.

4.4 Conclusions

Following the review done on past studies in Chapter 3, we concluded that there were no datasets collected for the specific task of the identification of toxicity-generating news. For this reason, we determined to use the dataset constructed by [10] in the context of the *Stop PropagHate* project. The use of this dataset was a good fit for our study, since it collected news pieces from various newspapers of distinct countries, in addition to user comments relative to those news articles. We started by analysing and describing the data gathering process, understanding the methods used for news and comment extraction. Additionally, we described the outcome from the data collection phase, resulting in the data that would be used in our work. From the three collections gathered by the author, we decided only to make use of the news and comments extracted from Twitter. This decision was based on the fact that comments extracted from the news sources websites, would represent a variance relative to the platform to witch the comment was made. This change in platforms could insert variances, such as user behaviour that could hinder the performance of our model. From this decision, we analysed the composition of the comment and

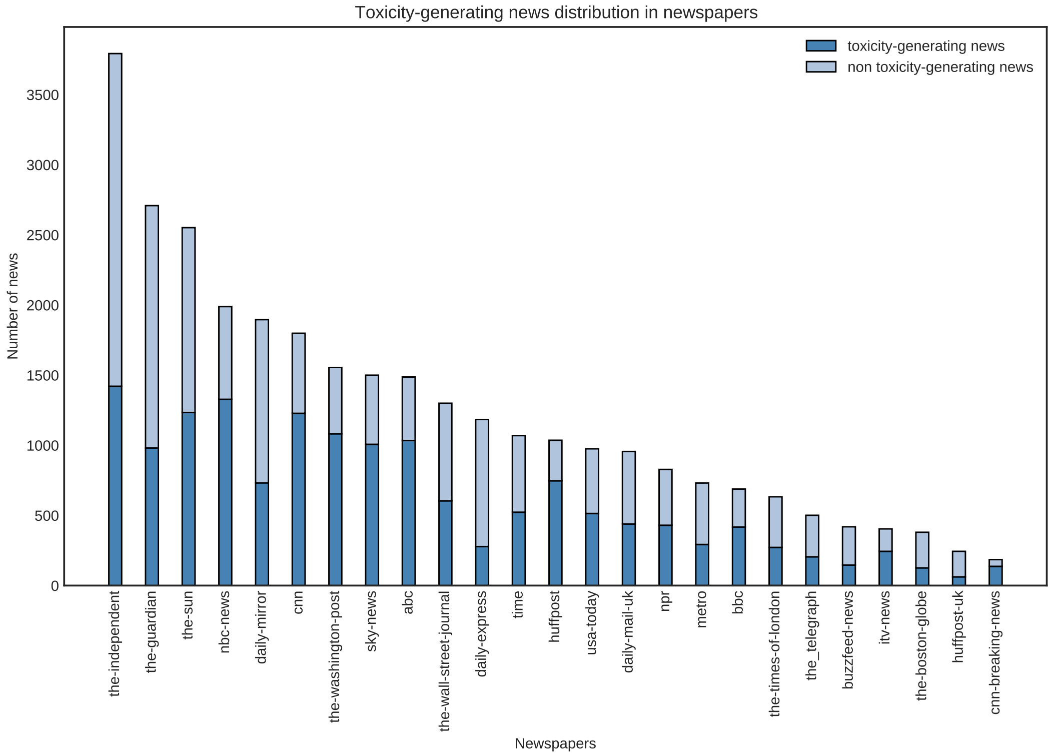

Figure 4.8: Toxicity-generating news distribution in newspapers.

news collection. We could conclude from Fig. 4.3 and 4.4 that the distribution of comments and news between newspapers was very unbalanced. We also concluded that the unbalanced nature of news and comments between newspapers might pose a challenge to the classification task. The different orientations and cultural backgrounds of newspapers may also pose a challenge to the classification process. Such variations can benefit classification models, making them robust to variations on the behaviour of newspapers and their consumers. This robustness is only possible when having a large and even amount of data for the model to learn, which is not the case in the dataset used in our work.

With the intent of focusing the classification task on the prediction of toxicity-generating news, we decided to perform the classification of comments using the *Perspective API*. From the attributes provided by *Perspective API*, we considered toxic comments those that had a *toxicity* attribute score greater or equal to 0.5. From this metric, we concluded that the median percentage of toxic comments was 11.1%. Furthermore, we could conclude that the extreme cases of “The Sun” as being the newspaper with the highest percentage of negative reactions (20.8%) and “The Guardian” the newspaper with the least (11.1%), suggested that the cultural variances and distinct orientation to which journalists draw news articles, contributed to the readers’ toxic behaviour.

Lastly, from the toxicity presence in comments, we described the news labelling process. We make use of the median toxicity percentage (11.1%) to classify news on toxicity-generating. News with a percentage of toxic comments higher than the median would have a positive label as to toxicity-generating. From this metric, we could conclude from Fig. 4.8 that having a larger number of news does not directly lead to a larger number of toxic comments. With the case of the "CNN Breaking News" being both the newspaper with the least number of news and highest toxicity percentage and "The Independent" the newspaper with most news articles, having a toxicity percentage of 33.1%. This fact indicates again that the newspaper personality and target audience posed as the main contributing factors to the growth of toxicity in news.

Chapter 5

Model performance and comparison

In the previous chapter, we described the data we use in our work, describing the approach taken for the classification of comments using the *Perspective API*. With the conclusion of the labelling process of news and comments, we focus on the development of a classifier able of identifying news as toxicity-generating. Furthermore, we want to understand which features contribute to the prediction of toxicity-generating news, and thus understand what characteristics are essential for the prediction of toxic behaviour. With this in mind, the present chapter details the set of features extracted and classification algorithms used. We also provide a description of the experiments conducted as well as the methods used in them, providing a description of the training, validation and testing steps used in experiments. We analyse and compare the results reported by different classifiers, interpreting which models perform better with a set of features. To conclude the chapter, we discuss the results obtained and the conclusions.

5.1 Methodology

In order to achieve the objective of this thesis, we performed several experiments so that we could compare the impact of distinct features in different models. We based the methodology in three particular phases, train, validation and test [12]. Each experiment varies with the features added to the models. A visualization of the pipeline constructed for the execution of experiments is presented in Fig. 5.1. A detailed description of each phase of the pipeline is presented in the following subsections.

5.1.1 Train and validation

With the intent of preventing the model from testing its performance on data that it has already seen, it is necessary to split data into two partitions, train and test. This way, the data used to train the model will not be used to assess its performance. To provide the model with data, we used the news articles dataset described in Chapter 4, composing a total of 29,726 news articles,

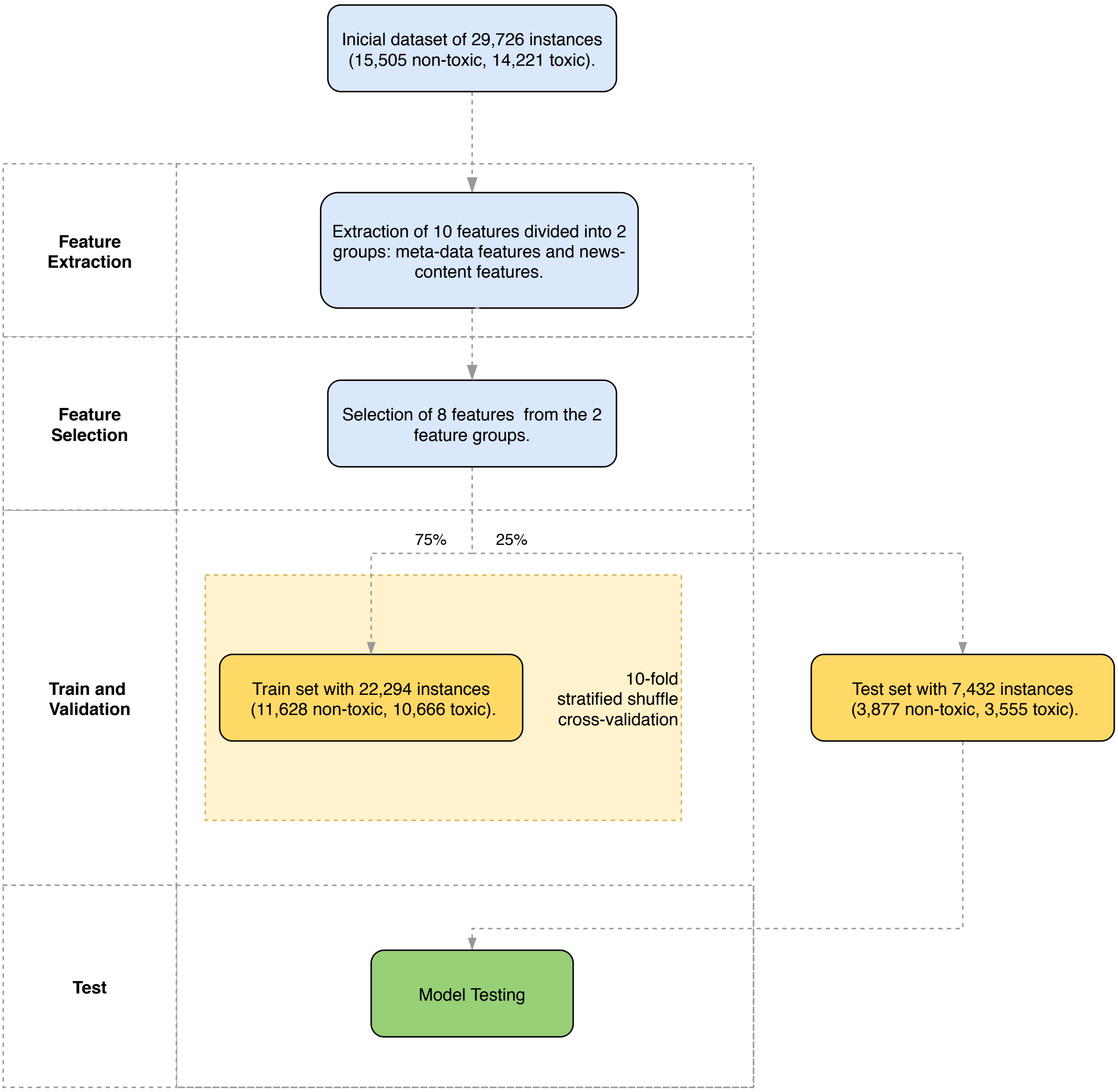
Inicial dataset of 29,726 instances (15,505 non-toxic, 14,221 toxic).
Feature Extraction
Extraction of 10 features divided into 2 groups: meta-data features and news-content features.
Feature Selection
Selection of 8 features from the 2 feature groups.
75%
25%
Train and Validation
Train set with 22,294 instances (11,628 non-toxic, 10,666 toxic).
10-fold stratified shuffle cross-validation
Test set with 7,432 instances (3,877 non-toxic, 3,555 toxic).
Test
Model Testing

Figure 5.1: Pipeline used for model comparison and performance evaluation.

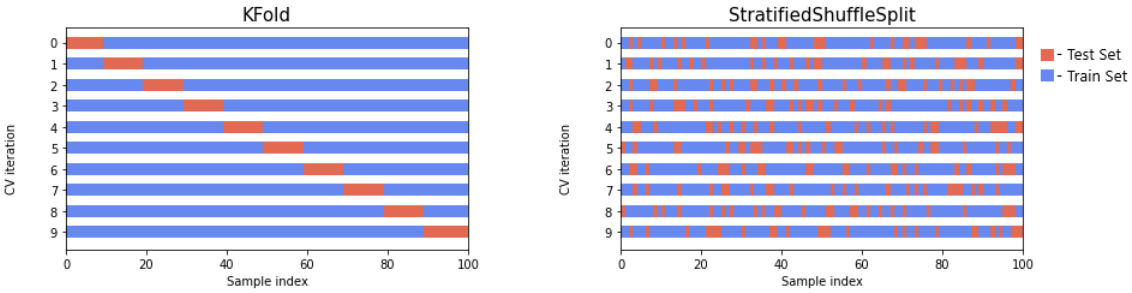

Figure 5.2: Cross-validation data split methods comparison using our data.

of which 15,505 are considered non-generators of toxicity and 14,221 toxicity-generating news. Even though the unbalanced nature of the datasets hinders the performance of machine learning models [9], since there are only 1,284 additional negative instances in the dataset, we considered our dataset to be balanced. Nevertheless, we additionally tested our model using a macro and micro variation of the F1 score to leave no room for doubts. With this said, we partition the data using a 75-25 percentage, with 75% of data for training (22,294), and 25% for testing (7,432).

Regarding the validation process, since our dataset is ordered by newspapers, using 10-fold cross-validation (2,972 per fold) meant that, in some folds, articles belonged to a single newspaper, this would hinder the model's capacity to learn. This way, we make use of the SciKit-learn 10-fold cross-validation stratified shuffle split method [63]. This method derives from the k-fold cross-validation but shuffles the data before dividing it into a 75-25 train test split, maintaining the percentage of the sample for each class. Fig. 5.2 provides a visual representation of data partitioning of two cross-validation methods, a classical 10-fold cross-validation, and a stratified shuffle split 10-fold cross-validation.

5.1.2 Test

In order to ascertain the models capacity to learn and predict on the data provided by the training set, each model was tested using 25% of the data, resulting in 7,432 news articles never seen by the classifier. To determine and understand the performance of the prediction, we make use of the Precision, Recall and F1 measures [65]. With the intent of further understanding the difficulties of the developed models, we use measures such as true positives, false positives, true negatives and false negatives. In the following section, we provide detailed information about the features extracted and a description of these features.

5.2 Feature extraction and selection

Having defined the methodology, we followed with the feature extraction and encoding that would later be used in the experiments with machine learning algorithms. The present section exhibits all the features extracted during the exploration phase. Further along in this chapter, we present the correlation analysis of the extracted features and interpret the relations between them. Lastly,

Table 5.1: Description of all features extracted from news articles.

Feature Group	Feature	Content
Meta-data	Newspaper	The newspaper publishing the news.
	Country	The country to which the newspapers article is established.
	Time of publication	The time of publication of the tweet containing the news (morning, afternoon, night).
	Number of comments	The number of twitter comments the article has originated.
	Number of comments bins	The bin to which the news article number of comments falls into.
News-content	Topic category	The category the news extracted from its title.
	News title entities	The entities present in the news title.
	News title entities frequency	The term frequency of the named entity present in news titles.
	Title keywords	The keywords from the news article title.
	Body entities	The entities from the news body.

we describe the feature selection process, indicating which features were chosen to be used in the classification of toxicity-generating news and which were dropped explaining the reason why.

5.2.1 Feature extraction

With the intent of providing our model with useful information, we extracted a set of ten features relative to news articles, to which we divided into two groups: meta-data features and news-content features. This set of features with the corresponding encoding resulted in a total of 2,093 features per article. Table 5.1 provides an overall review of all features extracted during this phase. We started the procedure by obtaining meta-data features from news articles, such as the newspaper, country, time of publication and the number of comment that the article has originated. As we experiment with such features, we further gathered news-content information, the entities present in news titles, as well as the topic category the news piece falls into. Finally, we extracted the keywords present in news titles and body entities. The following subsections provide a detailed description of each feature and its extraction process.

5.2.1.1 Article newspaper

As already mentioned in Section 4.1, our data is comprised of news articles from several sources from the USA and UK. With the intent of using this information, we extracted the newspaper of the news articles which reflected in a total 12,214 articles from 12 distinct newspapers from the USA, and a total of 17,753 articles from 13 distinct newspapers from the UK. When testing different models, the encoding of the newspaper feature was changed as well. This modification was done so that we could examine if one model would perform better with a particular encoding. With this said, we started by encoding the newspapers using a standard label encoding, representing each newspaper as an integer value. As label encodings can give the model an ordinal relationship where it does not exist, we also tested frequency encoding. Frequency encoding calculates the frequency that each newspaper appears in our data and uses it to label each newspaper with its frequency. Finally, we encoded the newspaper features using One-Hot Encoding (OHE) [52]. OHE creates an array where each element represents a possible category, the presence of a category is represented using a binary representation. Fig. 5.3 illustrates English based newspaper feature distribution. From this distribution, we can see that the newspaper feature is not balanced, and there are newspapers with more news than others. This fact proves that shuffling data in the validation set should be a crucial task for the better performance of models.

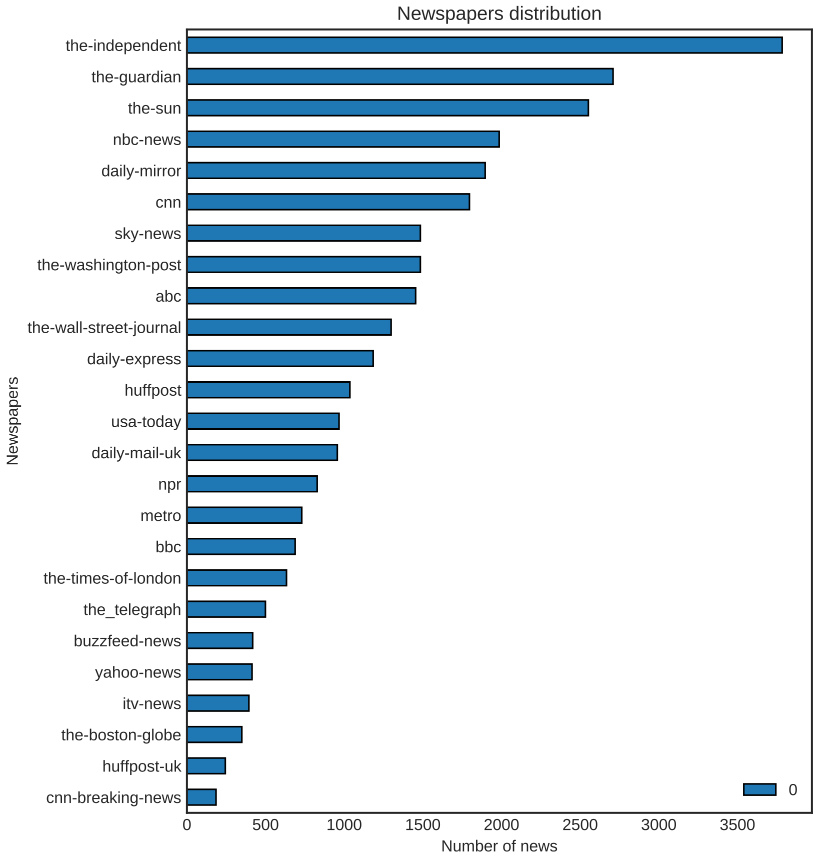

Figure 5.3: Newspaper feature distribution.

5.2.1.2 Article country

Similarly to the newspaper feature, the country of the newspaper source is used as a feature. Our data is composed of American and British articles, which resulted in a set of 12,214 articles features from the United States of America and 17,753 articles features from the United Kingdom as seen in Fig. 5.4. Regarding the encoding, similarly to the newspaper features, we experimented several types of encodings. We used label encoding, frequency encoding and OHE.

5.2.1.3 Article time of publication

Since our data has information about the hour of publication of the tweet containing the news article, we extracted the time of publication and transformed it into categorical data. The time was divided 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). Fig. 5.5 provides the distribution of the time of publication feature. Since time has an ordinal and sequential nature, we decided to use a label encoding.

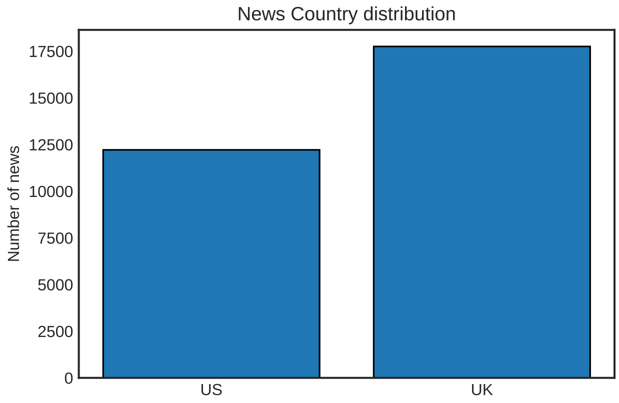

Figure 5.4: News articles country distribution.

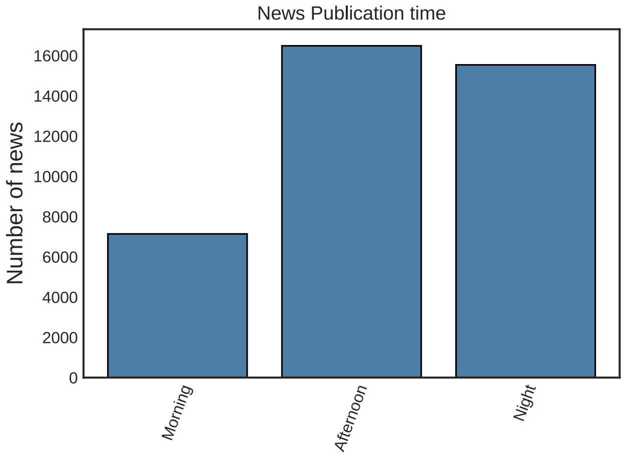

Figure 5.5: News articles tweets time of publication feature distribution.

5.2.1.4 Article number of comments

As our dataset contains information not only about news but also about Twitter comments relative to that news, we extracted two features from the number of comments: the total number of comments relative to that news article and the number of comments bins. The number of comments feature represents the total number of comments users made to a news article. The number of comments bins feature represents the uniform division of the number of comments distribution into five bins. This way, the number of comments bins feature represents the bin (1 to 5) to which the article's number of comments falls into. We encoded the number of comments bins using frequency encoding. Fig. 5.9 exhibits the distribution of the number of comments into five bins. Plotting the number of comments feature in Fig. 5.6, we can see that there are some outliers. Since the identification of outliers is a complex task [45; 44], we try to use standard methods of treating outliers. With this task in mind, we calculated the feature's standard deviation, 366.62 and mean value, 62.5. From these statistics, we apply a three standard deviations cut off technique, commonly used for removing outliers [37]. This technique only considers news articles with a number of comments bellow three times the standard deviation, in other terms, news with 1,160 comments or less. Applying the mentioned data cleaning method results in the removal of 241 news articles from our data.

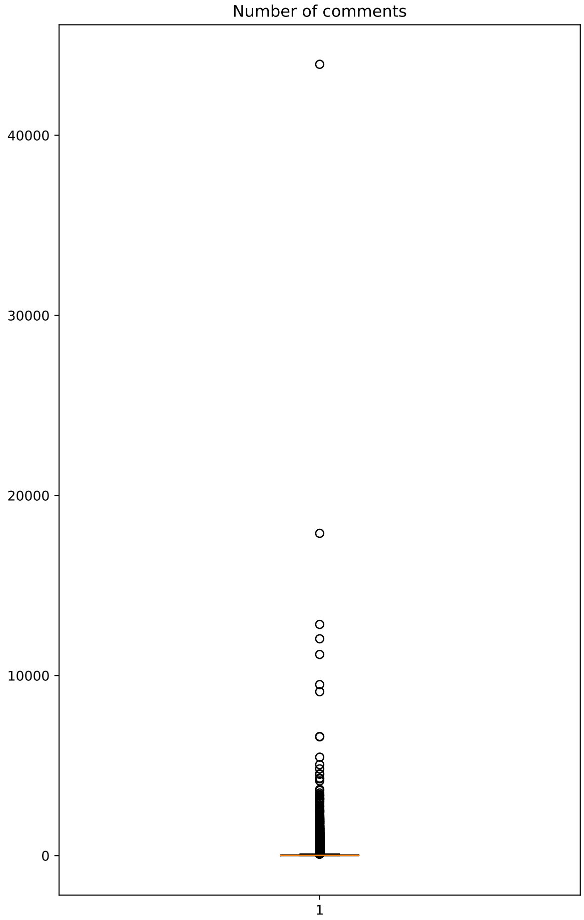

Figure 5.6: Number of comments feature boxplot.

From the distribution of the number of comments feature without outliers (Fig. 5.7), we can see that the data is very positively skewed. With the intent of smoothing and resolving the skewed nature of the feature, in turn, attempting to improve the model's performance using this feature, we performed a logarithmic transformation. We maintain the number of comments as a numerical value, not encoding it in any form.

5.2.1.5 News topic category

With the purpose of capturing the news topic and integrate it as a feature, a news topic classifier was developed using spaCy's *TextCategorizer* module [35], which uses an ensemble of a BoW model with a CNN to classify text. As a means to achieve a model capable of predicting a news article topic based on its headline, we make use of a public dataset called *News classification dataset* [53]. This dataset is comprised of 202,372 news articles from events between 2012 and 2018 from the American Huffpost newspaper, containing information about the headline, category,

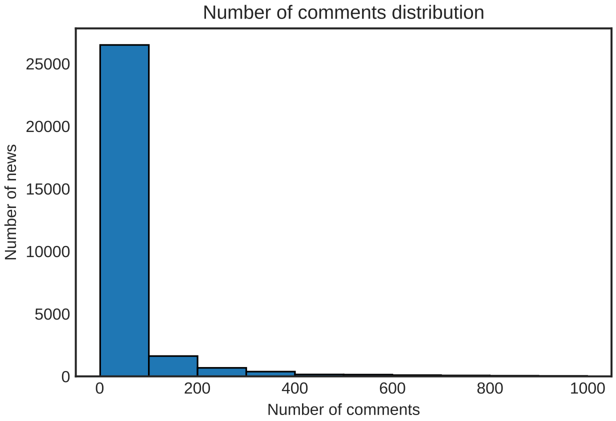

Figure 5.7: Number of comments feature distribution without outliers.

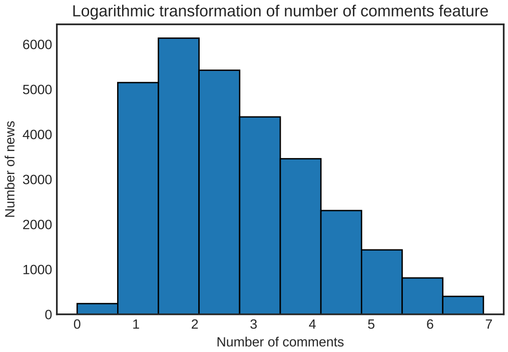

Figure 5.8: Number of comments feature with logarithmic transformation.

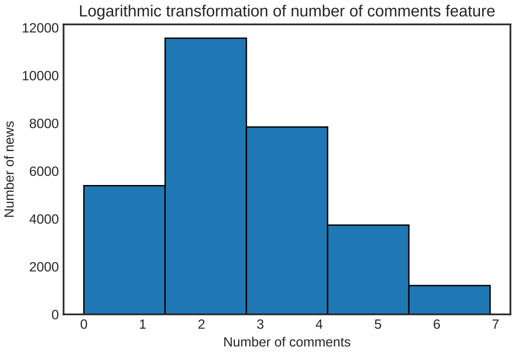

Figure 5.9: Number of comments within five bins.

author, link, date and a short description. We decided on using this dataset since it had a close context and information as our dataset. We also make use of news articles from the USA Huffpost newspaper and use similar information as the one provided to this topic classifier. Since there are a total of 30 categories in the *News classification dataset* and based on the fact that the categories are very unbalanced, we started by skewing the number of classes. To achieve this goal, we started by running our model and analyse its capacity to classify news using all categories. From this

Table 5.2: Final news topic categories considered and their instances from *News classification dataset* [53].

Category	**Instances**
Politics	32,739
Entertainment	16,058
Healthy Living	6,694
Queer Voices	6,314
Business	5,937
Sports	4,884
Parents	3,955
The Worldpost	3,664

first experiment, we could see in Table 5.3 that there were many classes that the model could not make good predictions, namely due to the large volume of categories and due to the few instances of news several categories had. Thereby we only considered the ten categories that had most instances. Running the model again and only predicting the selected ten categories, we could observe an overall increase in performance in the classification of each class. Analysing the results, we decided to drop any class that had a weak performance, as was the case of the comedy class, which had an 0.49 F1 score. Following the same principle for the third experiment, we dropped the black voices class, since it had an F1 score of 0.48. As we further experimented leaving out the Business class, as it had the lowest F1 score from all classes, we saw that the model would not benefit the model's performance. This resulted in our final set of eight categories enumerated in Table 5.2. With this set of labels, the model reaches an accuracy of 0.81.

The trained CNN was then applied to our dataset, extracting the news topic, resulting in one of the eight classes enumerated in Table 5.2. In the case of the predicted class having a probability score lower than 0.3, it would be assigned a distinct category, representing all news that did not fit any of the categories provided. Similar to what was done in other features, we experimented models with various types of encodings, namely label, frequency and OHE.

Table 5.3: News topic classifier experiments results with different categories classes as target (bold values indicate categories chosen for the next experiment).

News	**Experiment 1**			**Experiment 2**			**Experiment 3**			**Experiment 4**		
Category	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Black Voices	0.00	0.00	**0.00**	0.64	0.44	**0.52**	0.54	0.43	0.48	-	-	-
Business	0.32	0.51	**0.40**	0.59	0.53	**0.56**	0.67	0.49	**0.56**	0.65	0.52	0.58
Comedy	0.60	0.43	**0.50**	0.65	0.39	0.49	-	-	-	-	-	-
Entertainment	0.59	0.79	**0.67**	0.72	0.81	**0.76**	0.77	0.85	**0.76**	0.82	0.86	0.84
Healthy Living	0.45	0.63	**0.53**	0.68	0.69	**0.68**	0.66	0.70	**0.68**	0.70	0.69	0.69
Parents	0.50	0.67	**0.57**	0.67	0.67	**0.67**	0.77	0.63	**0.69**	0.73	0.66	0.69
Politics	0.69	0.85	**0.77**	0.82	0.90	**0.86**	0.86	0.89	**0.88**	0.86	0.91	0.88
Queer Voices	0.69	0.65	**0.67**	0.81	0.64	**0.72**	0.79	0.65	**0.71**	0.82	0.67	0.73
Sports	0.59	0.72	**0.65**	0.78	0.71	**0.75**	0.73	0.72	**0.72**	0.76	0.75	0.75
The Worldpost	0.44	0.64	**0.53**	0.74	0.69	**0.71**	0.68	0.76	**0.72**	0.76	0.69	0.72
Accuracy	-	-	0.79	-	-	0.76	-	-	0.78	-	-	0.81

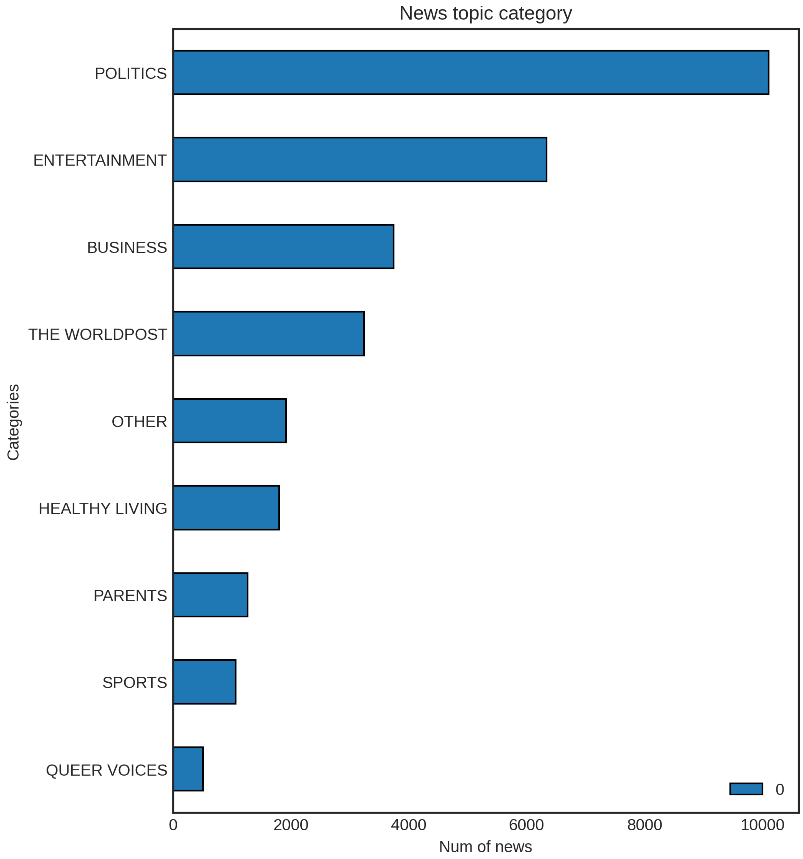

Figure 5.10: Topic category feature distribution obtained from the *Stop PropagHate* dataset.

5.2.1.6 Title entities

Aiming to extract entities present in news titles, we make use of the named entities recognition module provided by spaCy [33; 34]. Table 5.4 enumerates and describes the entities that the model is able to identify. As news articles may have more than one type of entity simultaneously in its titles, we encode the feature using OHE.

Upon the extraction of the 18 distinct named entities, we could observe in Fig. 5.11 that four main entities stand out, person, organization (ORG), geopolitical entities (GPE), and nationalities, religious or political groups (NORP). With the interest using this features in our model further on, we analyze in more detail each of the mentioned entities, we examine the top ten most frequent terms in each of the entities, represented in Fig. 5.12, 5.13, 5.14 and 5.15. Inspecting the most frequent person entities identified in news titles, we could recognize that the “Trump” and “Donald Trump” are the most mentioned person entities, followed by “Mum” and “Meghan Markle”. We could also notice that “Brexit” stands as a miss identification. Regarding the most mentioned ORG entities (Fig. 5.13), despite not being an organization itself, “Brexit” is the most present entity in news titles by a large margin. Its presence is not unexpected since most news articles belong to British newspapers. Despite this fact, “Brexit” is identified as most as double of the times in comparison to the second most frequent ORG entity “House”, which refers to the official residence

Table 5.4: Named entities and their description [34].

Entity	Description
Person	People, including fictional.
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. (Not services.)
Event	Named hurricanes, battles, wars, sports events, etc.
Work of art	Titles of books, songs, etc.
Law	Named documents made into laws.
Language	Any named language.
Date	Absolute or relative dates or periods.
Time	Times smaller than a day.
Percentage	Percentage, including ”%“.
Money	Monetary values, including unit.
Quantity	Measurements, as of weight or distance.
Ordinal	“first”, “second”, etc.
Cardinal	Numerals that do not fall under another type.

of the President of the United States of America. Besides the two mentioned organizations, it is possible to detect not only the identification of newspapers such as the “ABC”, “BBC” or “ITV” but also congresses, “European Union” (EU) and “Emmerdale”, an ITV soap opera. Regarding the most frequent geopolitical entities (GPE) (Fig. 5.14), we can see the “UK” and “US” as most prevalent, with “China”, “Russia” and “Syria” as the remaining most frequently identified countries. “Brexit” makes again the shelf of the most frequent entities of this type, which indicates that spaCy’s model has a broad definition for this term. Finally, investigating the nationalities or religious or political groups(NORP) top entities (Fig. 5.15), we see that “Democrats” is the most frequent, being the only political term present in the most frequent NORP entities. Additionally, we can find “Instagram” that does not belong to any entities of this type.

5.2.1.7 Title entities frequency

Derived from the conclusions taken from the analysis of the most frequent entities present in news titles, we chose to extract all the entity terms so that we could calculate their frequencies and use it to create an additional feature: the title entity frequency. This new feature represents the term frequency of each entity present in the news headline. Since news titles may have more than one entity type present in them, we chose to encode similarly as to OHE, with the frequency value in the array element representing the entity category, Table 5.5 exemplifies the extraction of the title entities frequency of a news title.

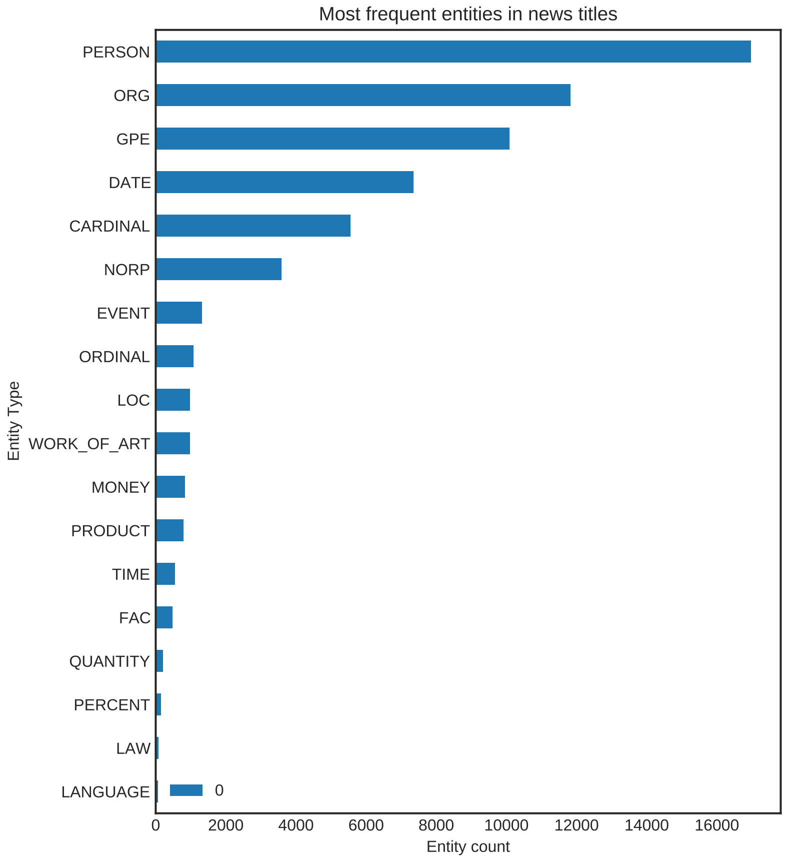

Figure 5.11: Most frequent entities present in news titles.

5.2.1.8 Title keywords

Besides extracting entities present in titles and their frequencies, we try to use the keywords existing in news titles. In order to use textual data from news titles, sentences must be first processed and consequently encoded as numerical values. Therefore, tokenization converts each of the words in the sentence into a token. Such a process is not a difficult task, Table 5.6 exemplifies the tokenization of a news title. Tweets, in contrast, have proven to have a more informal nature posing a challenge, due to nuances. The majority of news tweets gravitate towards a semi-structured and informal environment, often containing hyper links to the news article. Nevertheless, tweets may contain hashtags, mentions, abbreviations and non-alphanumeric characters. As a means to extract

Table 5.5: Example of entities term frequency feature.

News title	*China factory shrinks for the first time in two years.*
Entities terms	*'China' - (Gpe), 'two years' - (Date), 'first' - (Ordinal)*
Entities term frequencies feature	[0, 0, 0, 0, 0.087, 0, 0, 0, 0, 0, 0, 0.001, 0, 0, 0, 0, 0.002, 0]

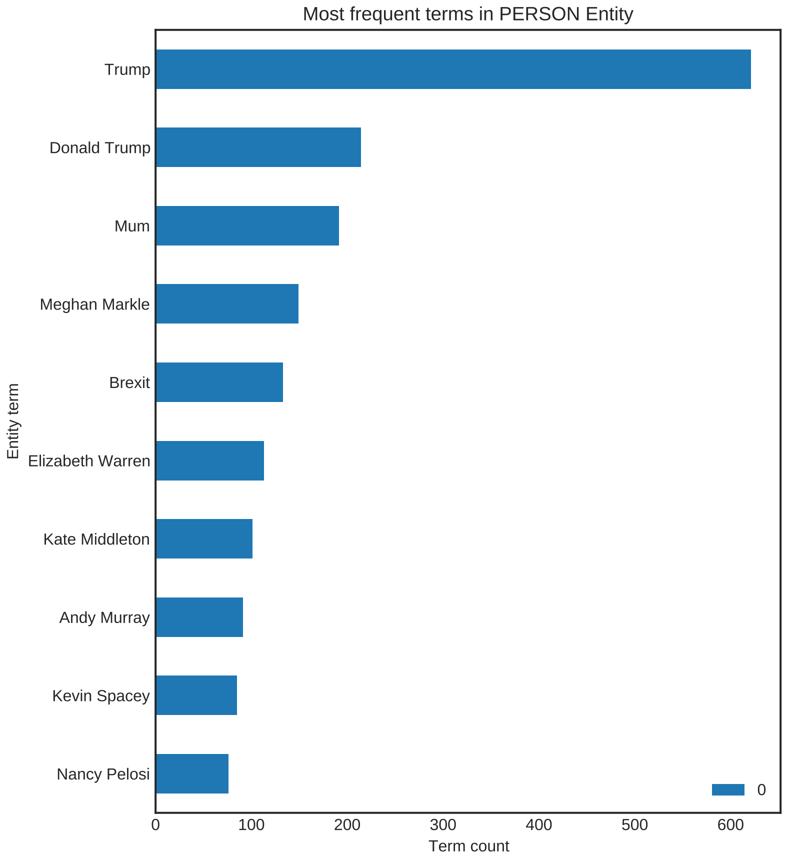

Figure 5.12: Ten most frequent PERSON entities identified in news titles.

essential features from such texts, it is crucial to do text preprocessing, so that we can reduce noise in data and improve the feature's potential for enhancing the model's performance.

With the intent of tokenizing and preprocess news titles to extract keywords, we started by experimenting with common libraries used in *Python* programming language. We experimented using a *Python* implementation of Rapid Automatic Keyword Extraction (RAKE) algorithm using the **NLTK** library [69; 72] as well as the **Gensim**, [67; 66] keyword module. As either libraries have different ways to extract keywords from news titles, we examine the process and outcome of each in the attempt of choosing the option that best suits our needs.

RAKE-NLTK In order to extract keywords from documents or text, the algorithm performs a set of tasks and applies several NLP techniques. First, RAKE converts the given document to lowercase. Following the preprocessing method, the document tokenization is performed, dividing

Table 5.6: Exemplification of a news title tokenization.

Sentence	*China factory shrinks for the first time in two years.*
Tokens	"China", "factory", "shrinks", "for", "the", "first", "time", "in", "two", "years", "."

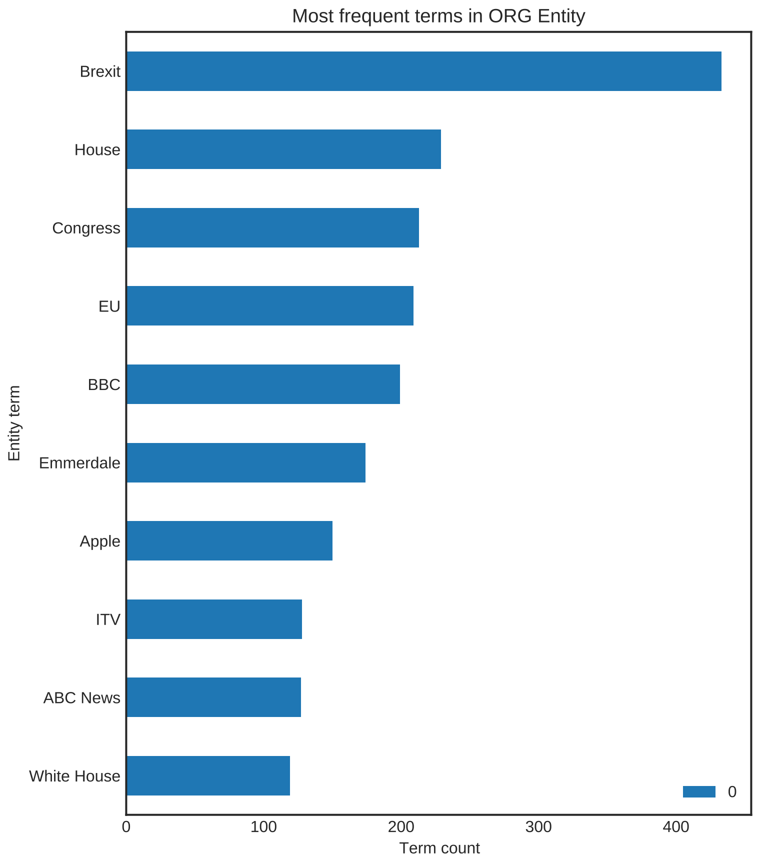

Figure 5.13: Ten most frequent ORG entities identified in news titles.

the text into tokens, maintaining all non-alphanumeric characters and punctuation, as Table 5.6 demonstrates. With the tokenization process finished, the algorithm follows with the removal of punctuation (e.g., "(", "&", "!") and stopwords (e.g., "the", "this", "from"). Having a list of relevant words, RAKE proceeds with the relevance calculation in order to select the tokens that will be considered as keywords. As means to achieve this measure, the algorithm calculates the word frequency $freq(w)$, how many times a particular word appeared among all candidate keywords, and the word degree $deg(w)$, the frequency that a particular word occurs with other candidate keywords. This way, the resulting keyword score is provided by calculating using the Equation 5.1. Table 5.7 illustrates an example of a news title keyword extraction with the RAKE algorithm using NLTK.

$$Ks(w) = (deg(w)/freq(w)) \tag{5.1}$$

Table 5.7: Keyword extraction example using RAKE with NLTK.

News tweet	*China factory activity shrinks for the first time in 2 years.*
Preprocessed tokens	*'china', 'factory', 'activity', 'shrinks', 'first', 'time', '2', 'years'*
RAKE Keyword scores	*('china factory activity shrinks', 16.0), ('first time', 4.0), ('2 years', 4.0)*

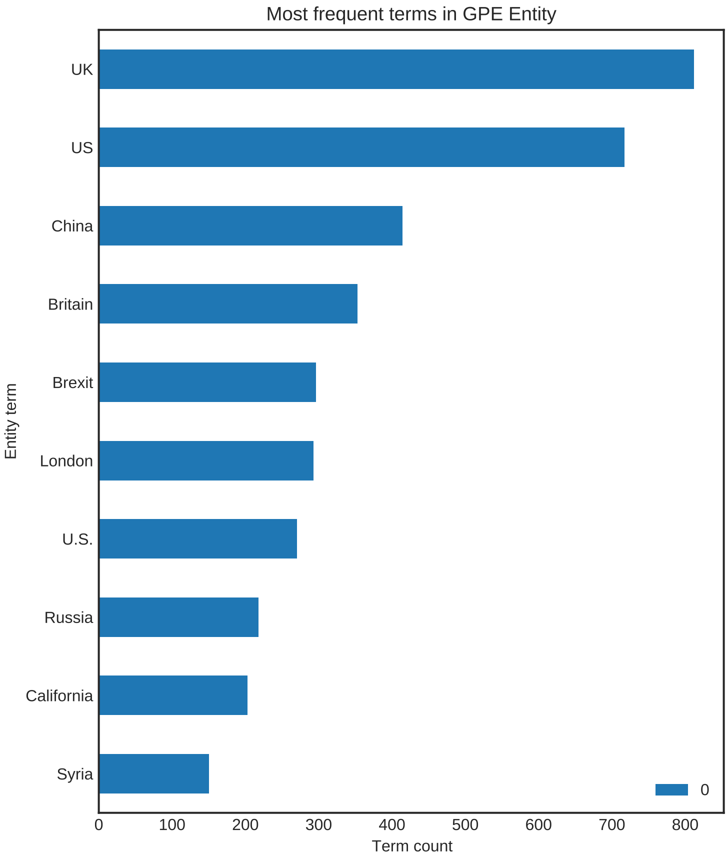

Figure 5.14: Ten most frequent geopolitical entities (GPE) identified in news titles.

Gensim keywords Intending to extract keywords from text, Gensim makes use of the TextRank algorithm. The algorithm first tokenizes text, following with text stemming and removal of stop-words. Concluded the preprocessing step, the algorithm calculates word embeddings and the similarity scores based on the given metric on the embeddings obtained. With this, TextRank constructs a graph, using words as nodes and similarities scores as vertices. From the graph, words and respective scores can be extracted, resulting in a list of keywords sorted by score. Table 5.8 illustrates an example of keywords extraction using Gensim library.

Table 5.8: Keyword extraction example using TextRank with Gensim library.

News tweet	*China factory activity shrinks for the first time in 2 years.*
Preprocessed tokens	*'china', 'factory', 'activity', 'shrinks', 'first', 'time', '2', 'years'*
TextRank Keyword scores	('factory activity shrinks', 0.436)

From examining each library resulting keywords, we could conclude that for short documents, such as news titles, Gensim demonstrated difficulties on extracting relevant words, which in cases the algorithm could not identify any keywords. We further analysed the distribution of the number of keywords in "ABC" news titles for each library, demonstrated in Fig. 5.16. From the histogram of the new titles keywords using Gensim, we could recognise that it could not extract keywords of

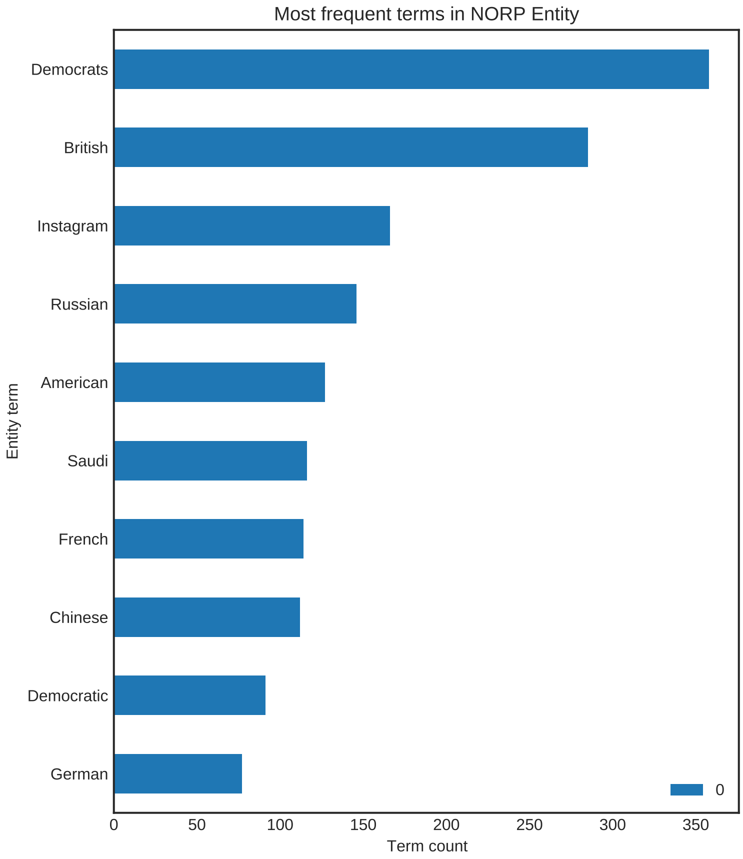

Figure 5.15: Ten most frequent nationalities or religious or political groups (NORP) identified in news titles.

more than 400 titles from a sample of 1,454 news titles, comparatively to 100 titles using NLTK library. For this reason, we chose to extract keywords using the NLTK library, which proved more consistent, as provided by the number of keywords distribution (right side) in Fig. 5.16.

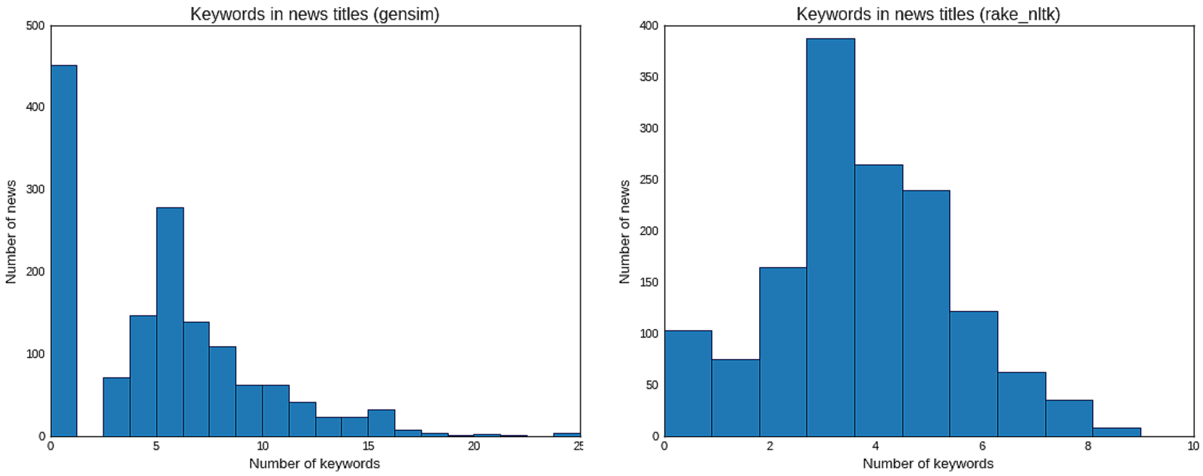

Figure 5.16: ABC news titles keyword distribution using Gensim and NLTK libraries respectively.

Since many machine learning algorithms do not support direct text as features, textual data must be converted into numerical data if we intend to use it. With this purpose in mind, we include news title keywords as features using BoW. Despite a simple method of encoding textual data, this method is frequently followed by "the curse of dimensionality", [12]. As having several distinct keywords leads to a high dimension of features, the time and computational complexity can increase exponentially, which in cases may cause other issues, such as lack of data, [12], overfitting and numerical instabilities [76]. In an attempt to lower the dimension of features, we calculate all keywords frequencies and only consider the top 1,000 most frequent keywords. This way, keywords are represented using a BoW of the 1,000 most common keywords.

5.2.1.9 Body entities

To extract features related to the news body and explore alternative ways to reduce noise, we extract relevant entities from news bodies using spaCy named entity recognition module [33]. We chose to change the extraction technique in comparison to title keywords because, despite named entities restricting the identified terms in news bodies, these have the advantage of producing a lower noise. This resulted in a smaller number of cases where words with no meaning were being represented as features.

Having no interest in numerical and ordinal terms, we excluded several types of entities. With this said, the chosen entities used for the extraction of news body features were persons, nationalities or religious or political groups (NORP), facilities (FAC), organization (ORG), geopolitical entities (GPE), locations (LOC), product, event, work of art, law and language. Therefore, we follow a similar encoding approach as for the title keywords feature. As having all news body entities represented would result in a high dimensionality, we only reproduce the 1,000 most common news body entity terms using BoW. Analyzing the resulting 50 most frequent body keywords, we can see from Fig. 5.17 that "Trump" is the most common term in news bodies by a large margin, with "London" as the second most frequent term in news bodies. We can see that the majority of terms are either from person ("Trump"), NORP ("American"), GPE ("Washington").

5.2.2 Feature correlation

As we extracted a set of features that derive from a parent feature such is the case as the number of comments features and title entities frequency features, we decided to analyse the correlation between these features and the remaining others. The decision to examine the correlation is due to the fact that we only intend to use one feature, the parent or child feature when testing our models. This way, we could make an informed decision on which had a stronger relationship with the remaining features, and thus lead to better performance.

Analysing the correlation of the features derived from the number of comments features (number of comments bins), we can examine in Fig. 5.18 that the number of comments feature with the logarithmic transformation increases the correlation with all remaining features, with an increase from 0.26 to 0.43 relative to the target variable we called "label". We can additionally observe

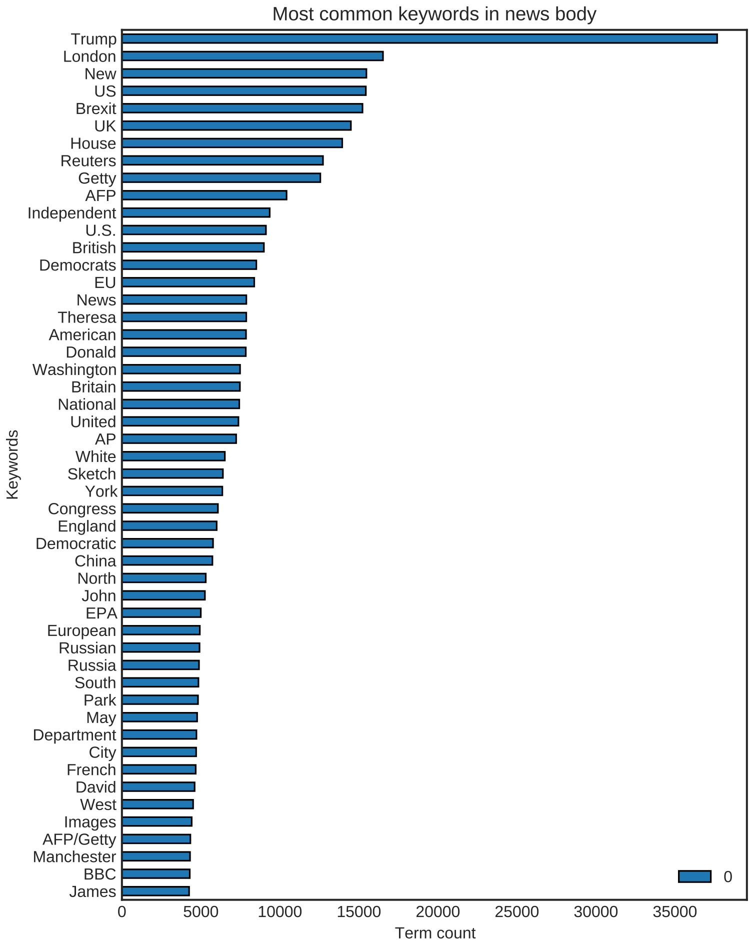

Figure 5.17: Most frequent entity terms in news bodies.

that the derived feature, number comments bins has a weaker correlation relative to all other features. Comparing the correlation of the title entities and title entities frequencies features to the remaining features (Fig 5.19 and 5.20), we can observe that although the correlation of every entity category is very small, a detailed comparison revealed that the title entity frequency feature demonstrates a weaker relationship to the other features.

5.2.3 Feature selection

Having explored the correlation between features and how the derivations of primary features influence the correlation between them, we selected the features that would be used on the classification task, and those that would be dropped. We decided that all meta-data features of news articles, such as the country, newspaper, time of publication and number of comments would be selected and used in the classification task. Furthermore, since the logarithmic transformation of the number of comments feature had a stronger correlation with all reaming features, we decided that this variation would also be used in the classification. The same could not be said to the number of comments bins feature, that we decided to drop due to the lower correlation when

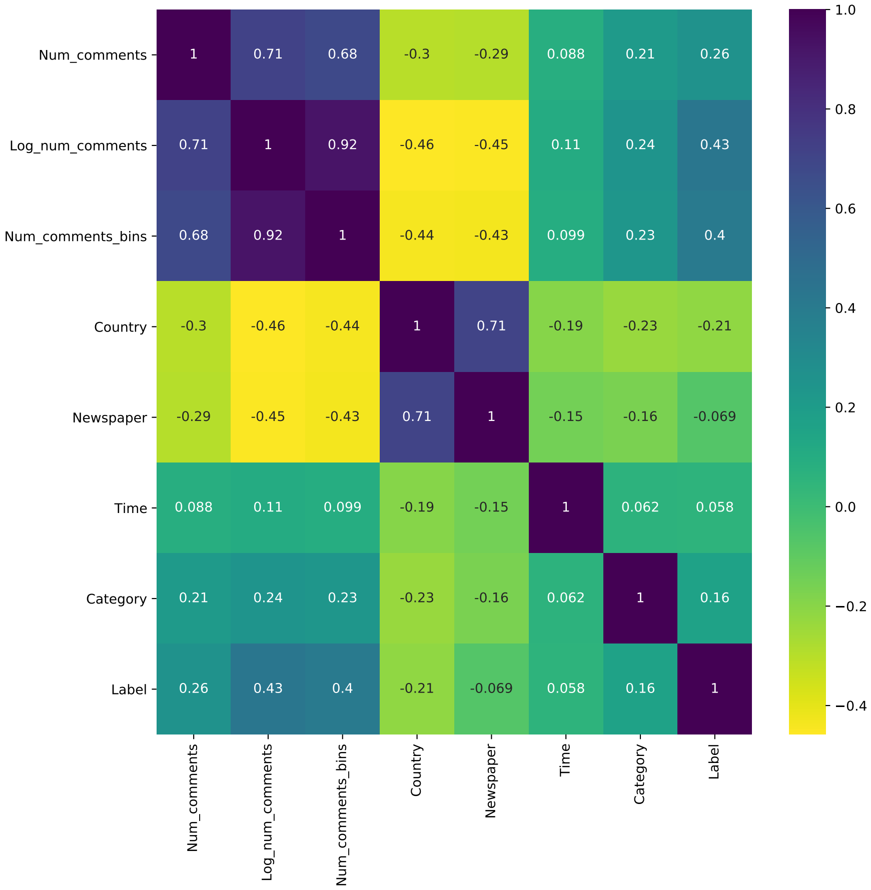

Figure 5.18: Correlation matrix with features derived from the number of comments feature.

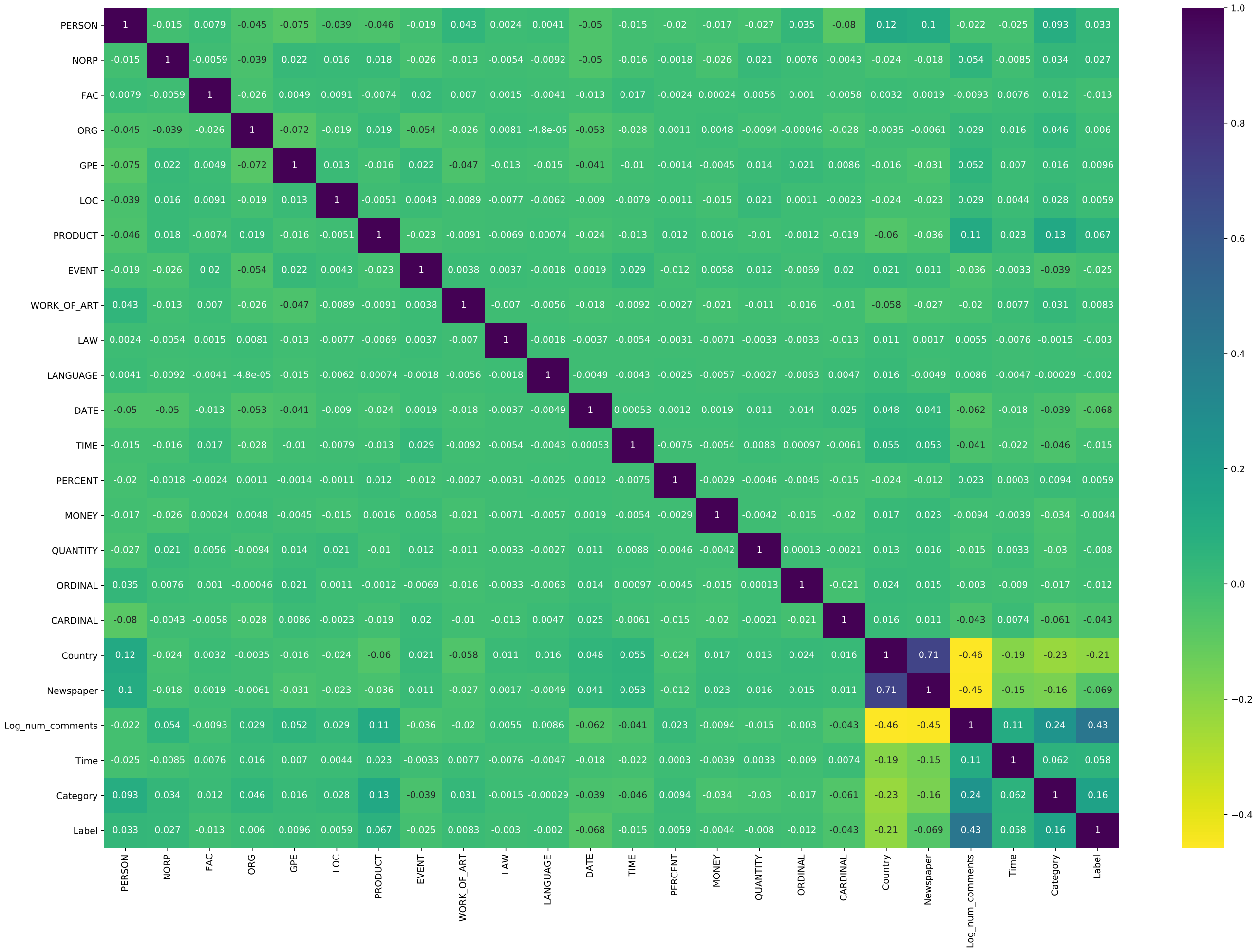

Figure 5.19: Correlation matrix with the title entities feature.

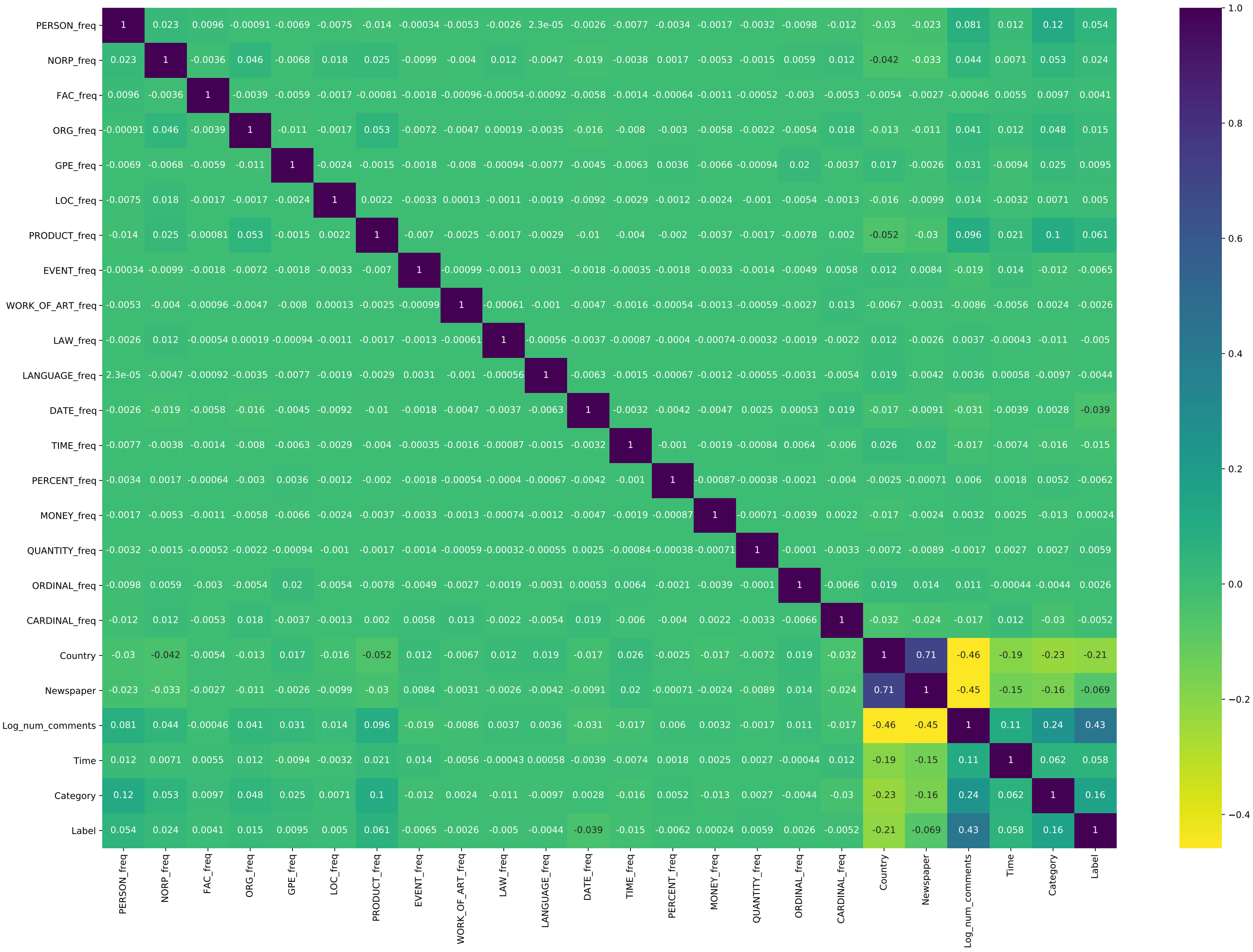

Figure 5.20: Correlation matrix with the title entities frequency feature.

Table 5.9: Final selected features used in the classification task.

Features
Body entities
Country
News title entities
Newspaper
Number of comments
Time of publication
Title keywords
Topic category

compared to the logarithmic transformation of the number of comments feature. We decided to add all news-content features, namely the topic category feature, the title keywords, body entities and the title entities features. Since the derived news title entities frequency feature demonstrated having a lower correlation when compared with its parent title entities feature, we decided that this variation would not be included in experiments during the classification task. Table 5.9 lists all the features that would we used in the classification task.

5.3 Results comparison and discussion

In order to understand how different models perform and analyse how feature combination affects the results of a model, we start by defining the baseline. Following this step, we continue with investigating the contribution in performance of each feature alone comparing improvements to the baseline. Further along, we select various models experimented throughout the development of this thesis, comparing the results obtained across all models. For the comparison, we not only analyse the prediction from the cross-validation using the training set data but also compare the results obtained using the data from the test set. Furthermore, we provide a description of the Receiver Operating Characteristic (ROC) curve and Precision-Recall curve of the model with the best performance. Lastly, we present an analysis of the feature correlation and feature importance obtained from the best performing model obtained in our work, in the attempt of analysing which features contributed to the classification task.

5.3.1 Choosing classifiers

With the intent of comparing various classifiers and analyse, which could label toxicity-generating news more accurately, we chose four different classifiers, SVM with linear kernel, logistic regression, random forest, and Gradient Boosting Decision Tree (GBDT). This decision is based on an analysis of the standard use of machine learning classifiers in the area of news classification presented in Chapter 3. Additionally, we chose to incorporate GBDT since, although being used with variations, this algorithm has gained popularity in machine learning competitions, providing solutions with promising scores, and greater efficiency [71; 77].

5.3.2 Baseline

In order to compare and understand the improvements provided by the addition of each feature, we need to establish a baseline first. Since our work makes use of a dataset constructed for the purpose of predicting hate speech generation in news, we consider our baseline to be the best performing model obtained in that work [10]. Despite the differences in research, such as the phenomenon, the metrics used to classify hate speech, the algorithms and the features extracted, we chose to consider its best experiment as a baseline since we make use of the same data and use features derived from the same components. As described in more detail in Chapter 3, the model we consider our baseline is composed of a CNN with an LSTM sequentially arranged, using the news titles as features. The proposed model achieved an F1 score of 0.607.

As an attempt to understand the contribution of each individual feature to the classification task, we compared the performance of each algorithm when having only one feature alone. Analysing the results (Table 5.10), we could conclude that the logarithmic transformation of the number of comments feature is the most significant contributor to the classification of toxicity-generating news, with all classifiers achieving an F1 score above 0.620, higher than our baseline. Comparing the results between the logarithmic transformation and the number of comments, we could conclude that the transformation increased the performance of the classifiers significantly for the SVM and logistic regression. We can also notice that the title keywords features is the second feature that aids the classification task the most, with the country feature following with an 0.522 F1 score for all algorithms.

Additionally, we chose to examine the contribution of the news-content features grouped (body entities, title entities, title keywords and topic category) as well as the dataset meta-data features grouped together (country, newspaper, number of comments, time of publication). From Table 5.11, we can examine that the algorithms reach higher performances when using only meta-data features compared to news-content features. We can observe that the GBDT reaches an F1 score of 0.723, far superior to the results obtained using the baseline. From the comparison of these two feature groups, we can conclude that news-content features do not contribute as much to the prediction of toxicity-generating news. This fact indicates that alternative methods of extracting features relative to news titles and body may be explored in order to maximize the model's ability to learn from such information.

5.3.3 Model comparison

Having explored the individual and grouped contribution of features to the classification task, we selected a set of feature combination experiments that we had an interest in exploring. Table 5.12 details each of the five experiments of feature combination with respective encoding. Although our dataset has a toxicity class imbalance that can be considered as irrelevant, throughout the review of model results, we chose to additionally provide the F1-micro and F1-macro scores. We decided to additionally provide these variations of the F1 score since we wanted to analyse scenarios where a class imbalance is not of interest (F1-macro) and scenarios otherwise (F1-micro).

Table 5.10: Performance of each algorithm when having one of the features.

Algorithms	Features	Precision	Recall	F1
Meta-data				
SVM	Number of comments (log. transf.)	0.679	0.570	0.620
Log. Reg.		0.677	0.608	0.641
Random Forest		0.673	0.664	0.668
GBDT		0.664	0.699	**0.680**
SVM	Number of comments	0.771	0.311	0.443
Log. Reg.		0.747	0.364	0.489
Random Forest		0.678	0.649	0.663
GBDT		0.669	0.677	0.673
SVM	Country	0.603	0.509	0.522
Log. Reg.		0.603	0.509	0.522
Random Forest		0.603	0.509	0.522
GBDT		0.603	0.509	0.522
SVM	Newspaper	0.560	0.309	0.398
Log. Reg.		0.551	0.343	0.423
Random Forest		0.648	0.525	0.580
GBDT		0.650	0.504	0.568
SVM	Time of publication	0.0	0.0	0.0
Log. Reg.		0.510	0.423	0.462
Random Forest		0.510	0.423	0.462
GBDT		0.510	0.423	0.462
News-content				
SVM	Title keywords	0.638	0.462	0.536
Log. Reg.		0.618	0.518	0.564
Random Forest		0.612	0.571	0.591
GBDT		0.651	0.409	0.502
SVM	Topic category	0.533	0.611	0.570
Log. Reg.		0.600	0.420	0.494
Random Forest		0.597	0.439	0.506
GBDT		0.597	0.439	0.506
SVM	Body entities	0.563	0.412	0.476
Log. Reg.		0.549	0.472	0.508
Random Forest		0.537	0.506	0.521
GBDT		0.570	0.427	0.488
SVM	Title entities	0.546	0.121	0.197
Log. Reg.		0.359	0.532	0.429
Random Forest		0.529	0.334	0.407
GBDT		0.532	0.359	0.429

Table 5.11: Performance of each classifier for meta-data features and news-content features.

Algorithms	**Feature group**	**Precision**	**Recall**	**F1**
SVM	Meta-data	0.680	0.717	0.698
Log. Reg.		0.697	0.660	0.678
Random Forest		0.701	0.664	0.682
GBDT		0.695	0.753	0.723
SVM	News-content	0.643	0.585	0.613
Log. Reg.		0.640	0.621	0.630
Random Forest		0.640	0.574	0.606
GBDT		0.650	0.620	0.635

Analysing the results from running our first model, which we named ***m_base***, using 10-fold cross-validation, we could examine that results were far from satisfactory such was the case as SVM and logistic regression, which had an F1 score of 0.357 and 0.452 respectively (Table 5.13). A cause for this score was that having 10-folds and being our data ordered by newspapers, a fold (2,996 news article features) could in cases represent the entire data from a single newspaper, hindering the model's ability to learn from that data. To improve the model's ability to learn from data relative to all newspapers, we tested the same model with the stratified shuffle split cross-validation method. Analysing the results (Table 5.13) we could examine significant improvements particularly to SVM and logistic regression, 0.576 and 0.577 F1 score respectively. From the improvement, we decided on using stratified shuffle split as the cross-validation method in all future experiments.

Having experimented with a simpler model, we continued with the addition of the topic category feature, which we called ***m_base+cat***. As a first attempt, we represented the additional feature using label encoding. Analysing the impact of the feature (Table 5.16), we can observe that, compared to the previous experiment (***m_base***), there are minimal improvements only regarding the logistic regression and random forest classifiers. This fact indicates that the topic category, at least with the current encoding, did not provide the classifiers with any additional information that can correlate to the generation of toxicity in news.

As an iteration from the experiment with ***m_base+cat***, we not only change the encoding method of various features but also use the logarithmic variation of the number of comments feature, this experiment we called ***m_base+log_transf***. We change the topic category feature encoding from label to frequency encoding and change the country and newspaper features encoding

Table 5.12: Description of each experiment with feature combinations and respective encoding method.

Experiments	**Features**								
	Title Entities	**Publication Time**	**Country**	**Newspaper**	**Num. comments**	**Num. comments (log. transf.)**	**Topic Category**	**Title Keywords**	**Body Keywords**
m_base	OHE	Label enc.	Label enc.	Label enc.	Numerical value	-	-	-	-
m_base+cat	OHE	Label enc.	Label enc.	Label enc.	Numerical value	-	Label enc.	-	-
m_base+log_ransf	OHE	Label enc.	OHE	OHE	-	Numerical value	Freq. enc.	-	-
m_base+title	OHE	Label enc.	OHE	OHE	-	Numerical value	OHE	BoW	-
m_base+body	OHE	Label enc.	OHE	OHE	-	Numerical value	OHE	BoW	BoW

Table 5.13: ***m_base*** performance metrics for both cross-validation methods for different classifiers (bold values represent significant improvements).

m_base		**F1-micro**	**F1-macro**	**F1-score**	**Precision**	**Recall**
10-fold cross-validation	SVM	0.604	0.451	**0.357**	0.771	0.412
	Log. Regression	0.636	0.526	**0.452**	0.728	0.450
	Random Forest	0.655	0.626	0.634	0.619	0.658
	GBDT	0.688	0.648	0.690	0.657	0.755
10-fold stratified shuffle split	SVM	0.619	0.615	**0.576**	0.621	0.538
	Log. Regression	0.640	0.632	**0.577**	0.666	0.508
	Random Forest	0.688	0.687	0.672	0.682	0.662
	GBDT	0.729	0.729	0.729	0.704	0.756

from label to OHE. With this changes, we could examine in Table 5.16 that the classification task regarding the SVM and logistic regression had significant improvements with an increase of 11.8% and 13.5% respectively relative to their F1 scores. This increase in performance shows that data processing and treatment are crucial components for the increase in performance seen in this experiment. As previously exposed in Table 5.10, the replacement of features from number of comments to its derived logarithmic transformation is mostly responsible for the increase in performance of the SVM and logistic regression algorithms. Additionally, the results obtained in this experiment suggests that the changes in encoding methods also contributed to an increase in performance, indicating that some algorithms can extract more information from features with a different encoding.

As we advanced into the addition of new features, we consider the addition of news title keywords. With this in mind, we not only add the representation of the title keywords but also change the encoding of the topic category to OHE. With these changes made, we named the new experiment as ***m_base+title*** and analysed how the selected classifiers performed. Analysing the performance reports from Table 5.16 we could observe that the SVM and random forest performances had benefited from the new feature resulting in an improvement of its F1 score by 1.9% and 1.3% respectively. These results indicate that, although the conclusion from Table 5.11 that news-content features had a smaller contribution to the classification task, the features extracted from the news titles have aided the prediction of toxicity-generating news. This fact indicates that the title keywords feature is relevant to the prediction of toxicity in news, but can possibly increase its significance.

As a final experiment named ***m_base+body***, we considered taking advantage of the information present in news bodies. For this experiment, we chose not to change any of the remaining features encodings, since to that time, our best performing model was ***m_base+title***. Analysing the results from Table 5.16 we could notice that the addition of the news body features did not improve the overall performance of the algorithms. We could observe that, although the random forest classifier registered a minimal increase in performance (0.3% F1 score), the remaining algorithms have a lower F1 score. This observation indicates that our approach to extract features from this element was not fruitful, and other ways to perceive information from this element had

Table 5.14: Confusion matrix of different algorithms using test data.

	TP	**FP**	**TN**	**FN**
SVM	2,770	1,107	2,587	968
Logistic Regression	2,889	988	2,456	1,099
Random Forest	2,866	1,011	2,452	1,103
GBDT	2,776	1,101	2,641	914

to be explored.

From an overall comparison of the performance of the training phase results of the different models, it could be concluded that ***m_base+title*** resulted as the overall best model. Despite different classifiers achieving better results in other models, such as GBDT in ***m_base*** and random forest in ***m_base+body*** and logistic regression in ***m_base+log_transf***, the most consistent model considering all classifiers is ***m_base+title***, having all algorithms reaching performance over 0.70. Furthermore, we can see that when analysing the influence of the imbalanced nature of ***m_base+title***, this model continues to be our best performing model with F1-micro scores above 0.72.

Concluding that our best model was ***m_base+title***, we continued with evaluating the prediction results using the test data, never accessible to the model during the training phase. First, we analysed the confusion matrix, examining where the model's difficulties lie. Secondly, we analysed the performance metrics so that we could observe how the model performs with the test data. Analysing the confusion matrix of all algorithms using the test data (Table 5.14), we could recognise that logistic regression and random forest have a higher number of true positives, achieving 2,889 and 2,866 respectively. Regarding true negatives, we could detect that GBDT had the highest value with 2,641 negative instances correctly identified.

We could also witness from Fig 5.21 that SVM and GBDT had a higher number of false positives, and a lower identification of true positive when compared to logistic regression and random forest. With this information, we could conclude that SVM and GBDT had more difficulty of correctly identifying positive cases of toxicity compared to the remaining classifiers. In counterpart, we could see that SVM and GBDT were skilled in identifying negative cases. Regarding the logistic regression and random forest, we could conclude the contrary.

From the performance metrics calculated from the test set data using ***m_base+title*** (Fig. 5.22), we could observe that, despite the proximity of the values for each classifier, the logistic regression had the highest Precision of all algorithms. The high Precision indicated that the logistic regression had a higher proportion of true positives and the total positive predicted news. When comparing the Recall of different algorithms, we could observe that GBDT performed best. Having a higher Recall value indicates that GBDT had a higher proportion of true positives of the total of toxicity-generating news. Lastly, we could see that the algorithm with the highest F1 score is GBDT, with a margin of 0.008 to the second-best performing algorithm, logistic regression.

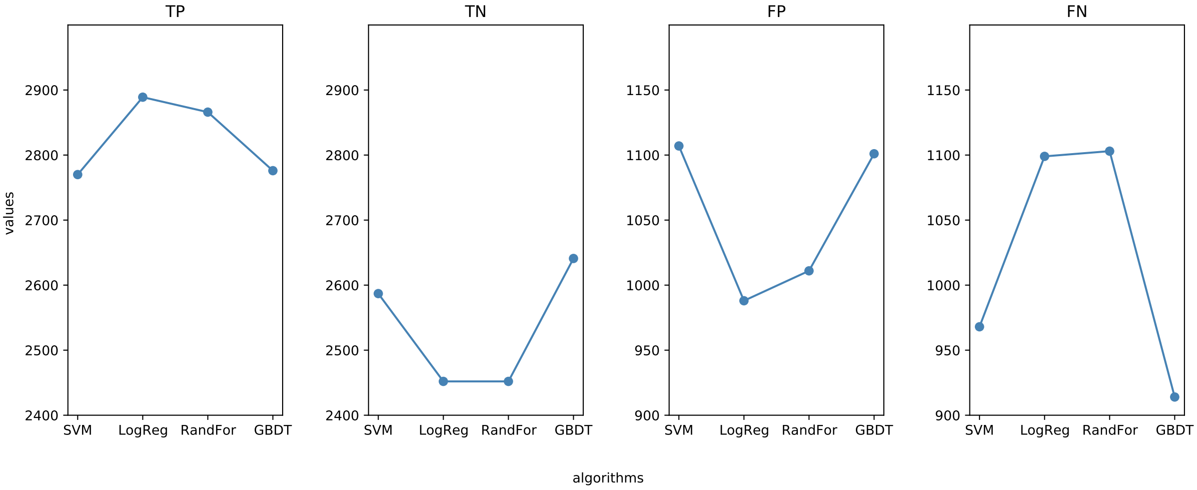

Figure 5.21: Comparison of the ***m_base+title*** test set confusion matrix for each algorithm.

Table 5.15: ***m_base+title*** performance metrics using test set data.

	Precision	**Recall**	**F1**
SVM	0.714	0.741	0.727
Logistic Regression	0.745	0.724	0.734
Random Forest	0.739	0.722	0.730
GBDT	0.733	0.752	0.742

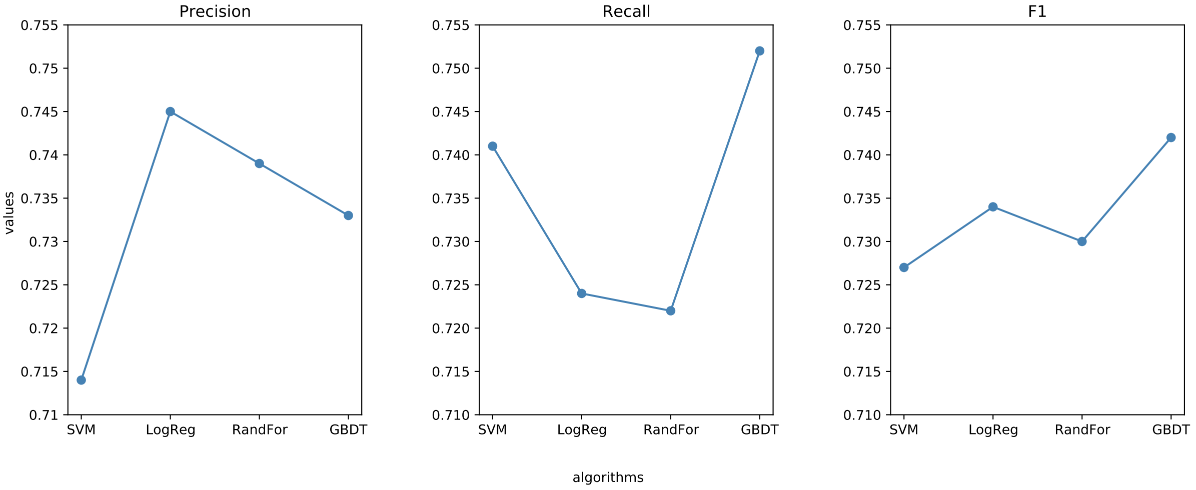

Figure 5.22: Comparison of ***m_base+title*** test set performance metrics for each algorithm.

Table 5.16: Comparison of the performance metrics from all experiments (bold values represent best performance of each algorithm).

	F1-micro	**F1-macro**	**F1-score**	**Precision**	**Recall**
m_base					
SVM	0.619	0.615	0.576	0.621	0.538
Log. Regression	0.640	0.632	0.577	0.666	0.508
Random Forest	0.688	0.687	0.672	0.682	0.662
GBDT	0.729	0.729	**0.729**	0.704	0.756
m_base+cat					
SVM	0.619	0.615	0.576	0.622	0.537
Log. Regression	0.643	0.635	0.580	0.671	0.510
Random Forest	0.693	0.692	0.680	0.684	0.676
GBDT	0.730	0.730	**0.729**	0.706	0.753
m_base+log_transf					
SVM	0.720	0.718	0.697	0.721	0.675
Log. Regression	0.726	0.725	**0.715**	0.711	0.720
Random Forest	0.702	0.702	0.690	0.688	0.692
GBDT	0.729	0.729	0.725	0.704	0.747
m_base+title					
SVM	0.722	0.722	**0.716**	0.701	0.733
Log. Regression	0.720	0.719	0.705	0.711	0.699
Random Forest	0.720	0.719	0.703	0.713	0.694
GBDT	0.728	0.728	0.723	0.705	0.742
m_base+body					
SVM	0.712	0.712	0.698	0.700	0.696
Log. Regression	0.713	0.712	0.698	0.703	0.692
Random Forest	0.721	0.721	**0.706**	0.713	0.700
GBDT	0.727	0.727	0.722	0.704	0.741

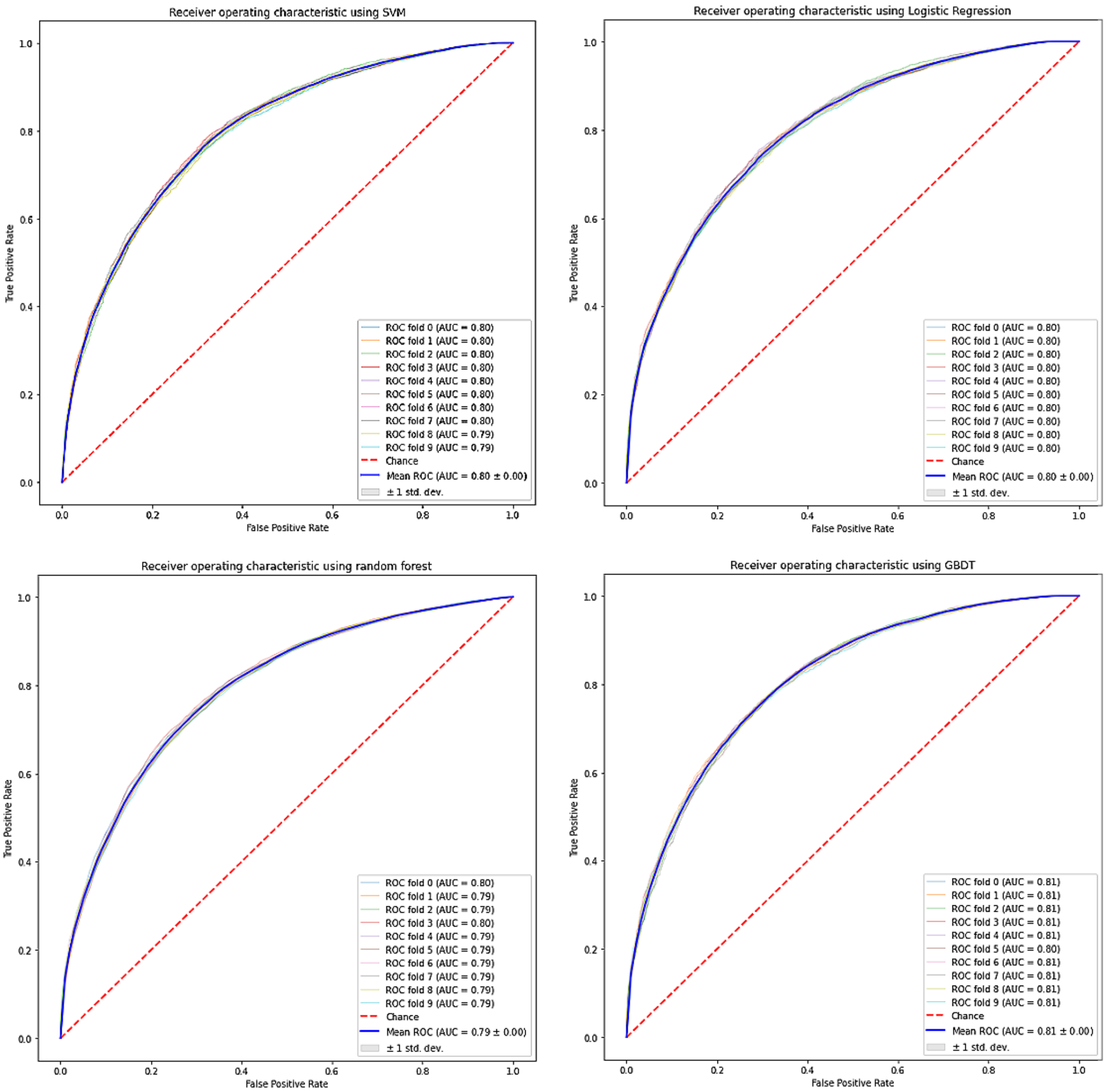

Figure 5.23: ROC curve of all classifiers using ***m_base+title***.

5.3.4 ROC and Precision-Recall curve

Analysing the results from the feature combination experiments performed, we could conclude that when considering the F1 score of each classifier our best performing model was the ***m_base+title***. With the intent of further validating the capability of our model, we calculated the ROC and Precision-Recall curves for each classifier. As seen in Fig. 5.23 our model is skilful in identifying true non-toxicity-generating news, with an average Area Under Curve (AUC) ranging from 0.79 (random forest) to 0.81 (GBDT).

Despite this promising fact, ROC plots have an optimistic view of the performance of a model when using an imbalanced dataset [20]. Furthermore, researchers indicate that ROC plots may be deceiving and lead to incorrect interpretations of the performance of its models when dealing with

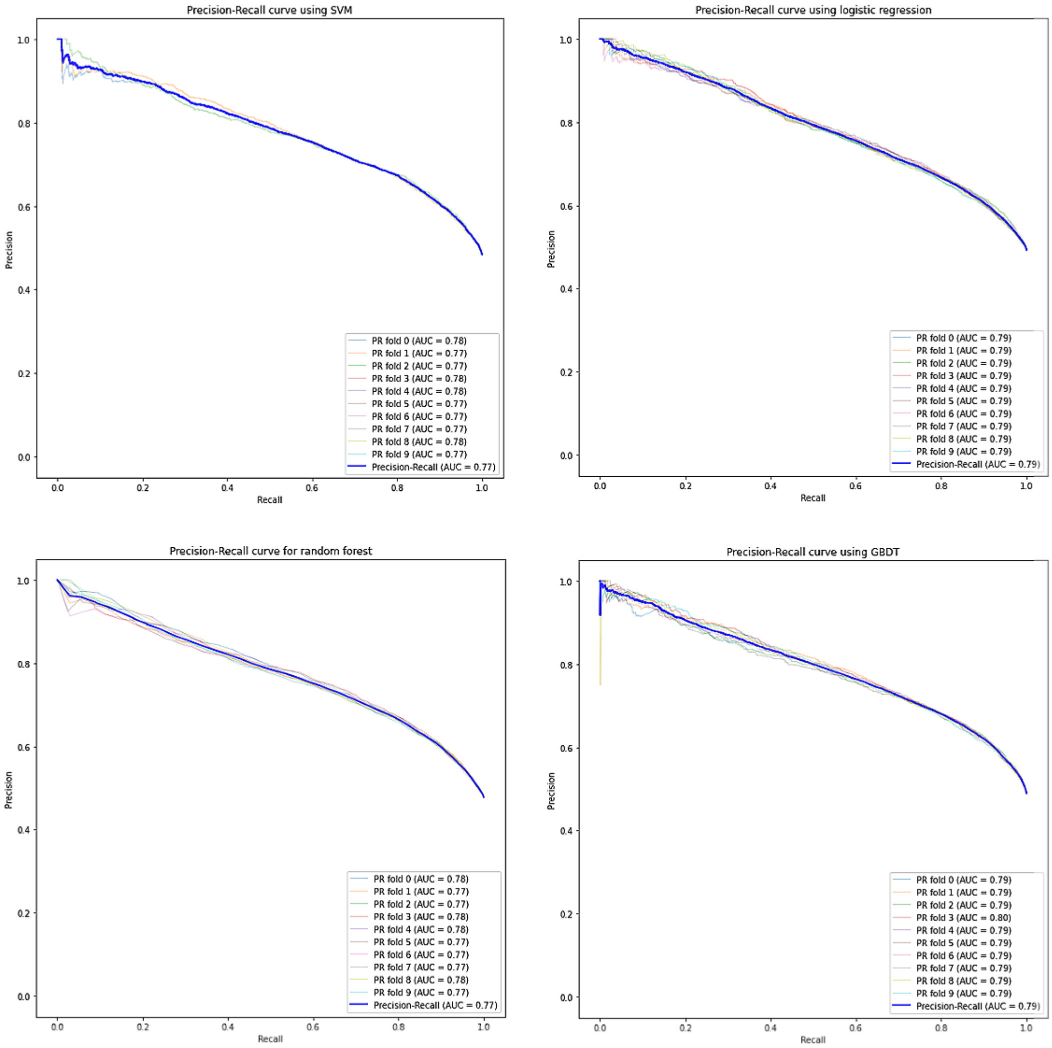

Figure 5.24: Precision-Recall curve of all classifiers using ***m_base+title***.

imbalanced datasets [70]. Although our data may be considered balanced since the delta between the positive and negative class is so small (additional 1,284 negative classes), we chose to give no room for deceiving results. For this reason, we further explore the Precision-Recall curve. The Precision-Recall curve only concerns the correct prediction of the class with a minority, in our case, the positive (toxicity-generating news) class. Therefore, this type of comparison is better suited for classification tasks that deal with class imbalance.

For the reasons stated above, we examine the Precision-Recall curves of all the selected classifiers for our best performing model. We can notice from Fig. 5.24, that the AUC of Precision-Recall ranges from 0.77 (random forest) to 0.79 (GBDT, logistic regression). Although these values are lower relative to the ones regarding the ROC, it solidified the reasonable ability of our model to identify toxicity-generating news.

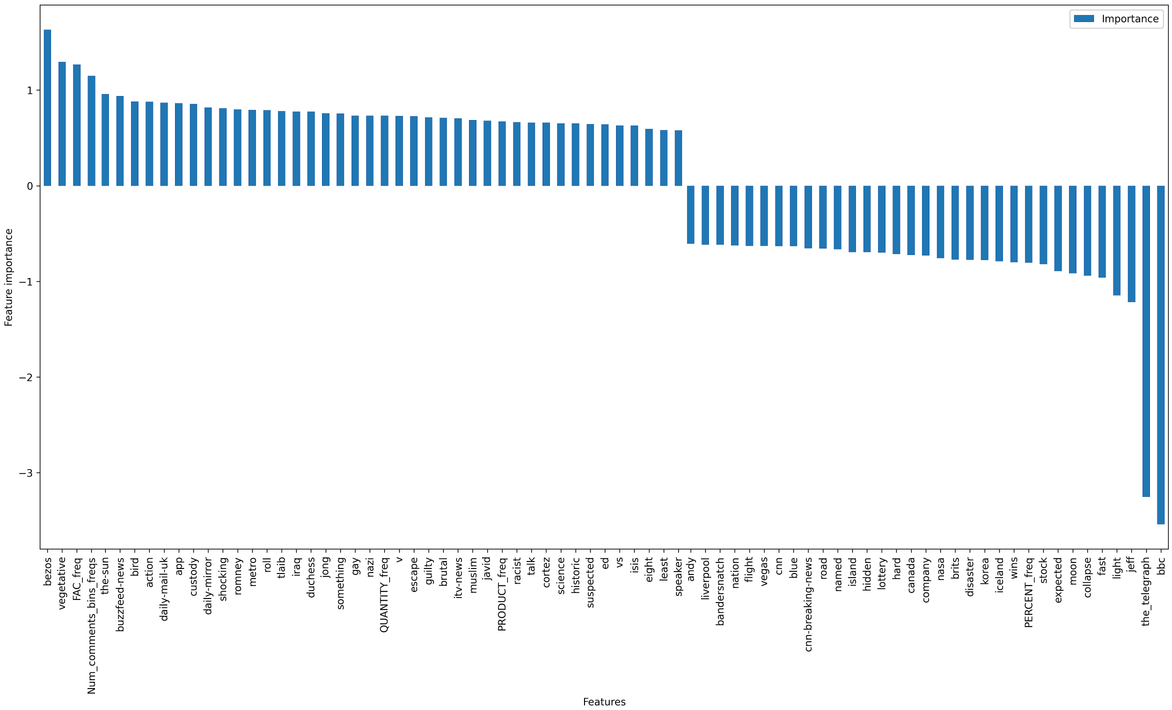

Figure 5.25: Feature importance using the SVM classifier.

5.3.5 Feature importance and correlation

In order to better understand our most competent model, and analyse which features influenced the generation of toxicity, we examined feature importance for each of the classifiers used. Furthermore, we explored the correlation between features and our target variable in the attempt of reaching a conclusion of which features influence the most when predicting toxicity-generating news.

Regarding feature importance, we started by analysing, which features contributed the most for the identification of toxicity-generating news when using SVM. From Fig. 5.25 we could conclude that news title terms such as “bezos” (Amazon founder and CEO) and “vegetative” are features that contributed the most for the identification of toxicity-generating news. Furthermore, we could see that the number of comments contributed to the positive identification as well. Regarding news outlets, we could detect that “The Sun” and “Buzzfeed News” were the two newspapers that contributed to the positive classification of news as toxicity-generating. “The Guardian” and “BBC” were the two predominant features that influence (negatively) the classification of news as toxicity-generating. The presence of newspapers in both extremes of most valued features of the classification of toxicity-generating news, suggests that the culture and ideology influenced the type of responses their users have on news.

When analyzing the feature importance for the logistic regression classifier, we could recognize from Fig. 5.26 that scores ranged from 1.5 to negative 6. We could conclude that for this classifier most of the significant contributors to the positive or negative labelling of news as

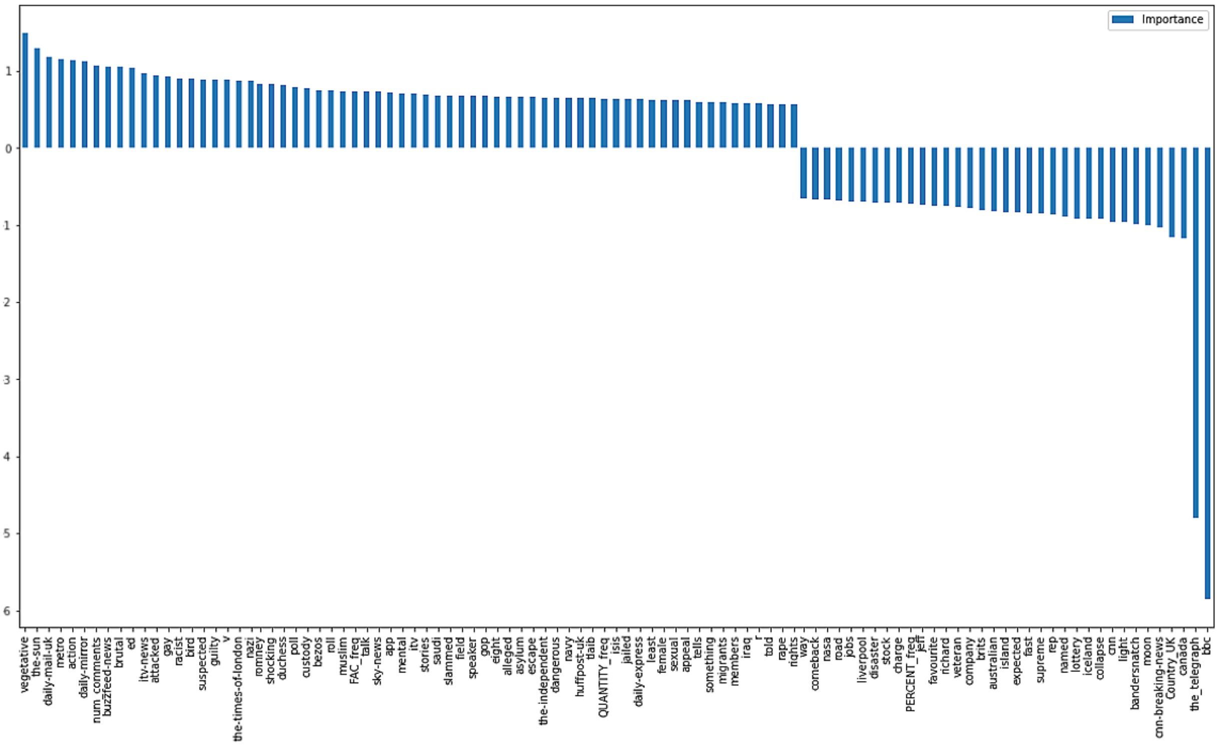

Figure 5.26: Feature importance using the logistic regression classifier.

toxicity-generating represented either newspapers (e.g., “The Sun”, “The Daily Mail”, “BBC” and “The Telegraph”) or specific social terms present in news titles (e.g., “gay”, “racist”). The significant negative importance of the “BBC” and “The Telegraph” newspapers supports our claims when analyzing the feature importance of the SVM classifier. The significance of these newspapers makes us believe that the newspaper culture and ideology posed a strong indicator of the production of toxicity-generating news. Furthermore, we could conclude that terms present in news titles such as “brutal”, “gay”, “racist” and “nazi” are feature terms that contributed the most for the identification of toxicity-generating news.

Concerning which features best aided the random forest algorithm on the labelling task of toxicity-generating news, we could analyse in Fig. 5.27 that the number of comments was the most relevant feature with an importance score above 0.08. The second most contributing feature for the classification task was the topic category, followed by the publication time. We could also conclude that the “BBC” newspaper, despite having a lower importance score of 0.02, is the only newspaper that stood out.

Finally, analysing the feature importance of the GBDT classifier, we could conclude in Fig. 5.28 that the significant contributing feature for the classification of toxicity-generating news was the number of comments, with an importance greater than 0.5. Additionally, we could conclude that the second most prominent features were the “BBC” and “The Telegraph” newspapers. The fact these two features were represented in all classifiers as significant contributors to the classification task proposes that the newspaper culture and journalistic ideology were factors that influenced the

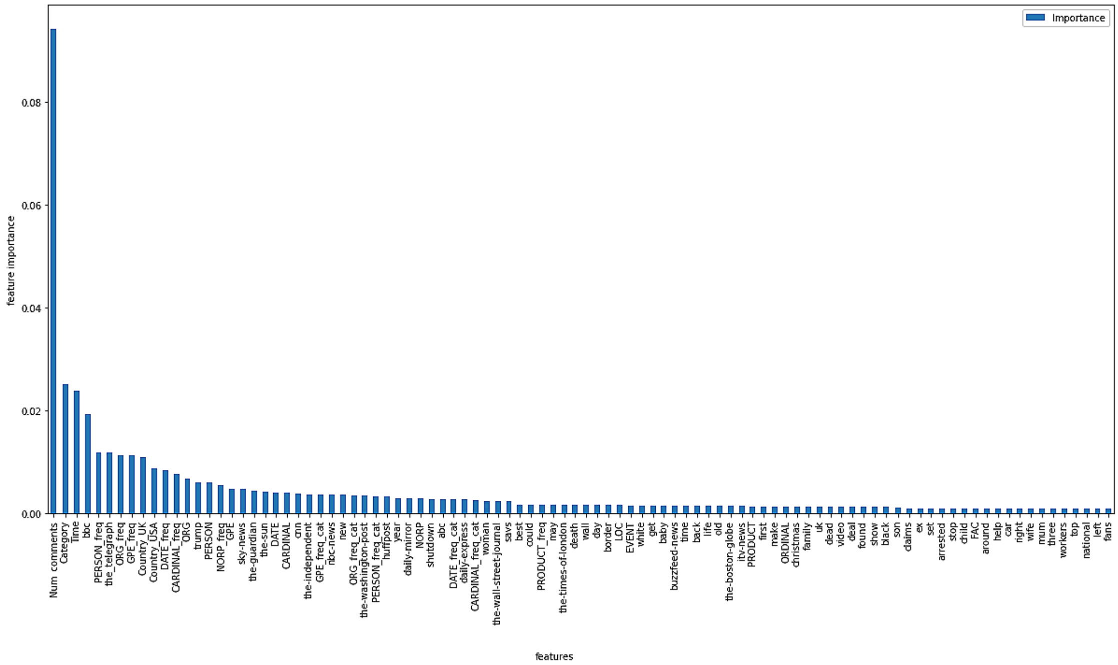

Figure 5.27: Feature importance using the random forest classifier.

tendency to generate toxicity. Moreover, the presence of the feature term “Trump” in the most contributing features for random forest and GBDT, demonstrates that political news and topics tend to lead to a better prediction of toxicity-generating news. We could also recognise feature terms relevant to logistic regression, such as “gay” and “racist” that solidifies our belief that controversial social themes pose as foundations to the identification of toxicity-generating news.

Analysing the correlation matrix (Fig. 5.29) of the best performing model obtained (***m_base+title***), we could see that none of the used features has a high correlation to our target label. Nevertheless, we could see that the logarithmic transformation of the number of comments posed as the feature with the strongest correlation to the target label. The country feature was the second most contributing, with the category and title keywords features following. The fact that no feature had a strong correlation with the target label indicates that the model's performance is not justified by a set of features but the conjunction of all.

5.4 Conclusions

In the current chapter, we not only analysed how our model performed when predicting toxicity-generating news but also decided to investigate what features benefited the prediction. In order to achieve such conclusions, we developed a pipeline used for the experiment with feature combination. Based on the results collected from the experiments performed, we could observe that the best performing model was what we named ***m_base+title***. We could conclude that for this model, the GBDT was the algorithm with the best performance when reaching an F1 score of

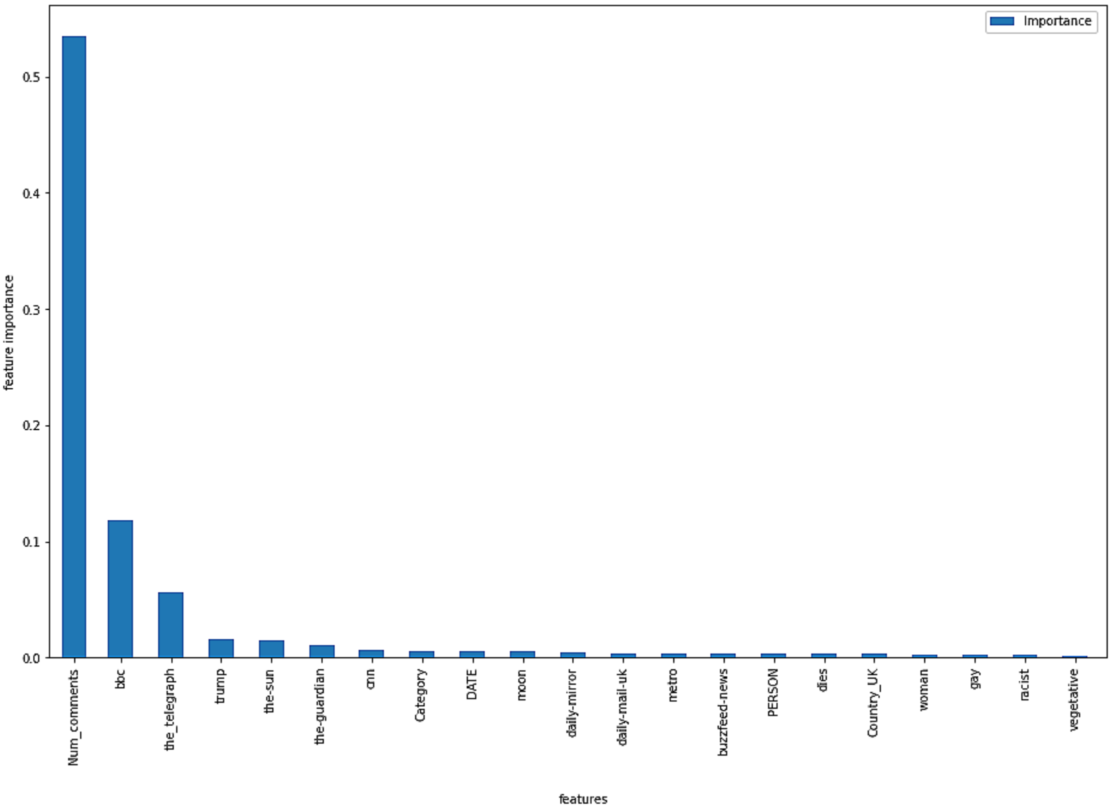

Figure 5.28: Feature importance using the GBDT classifier.

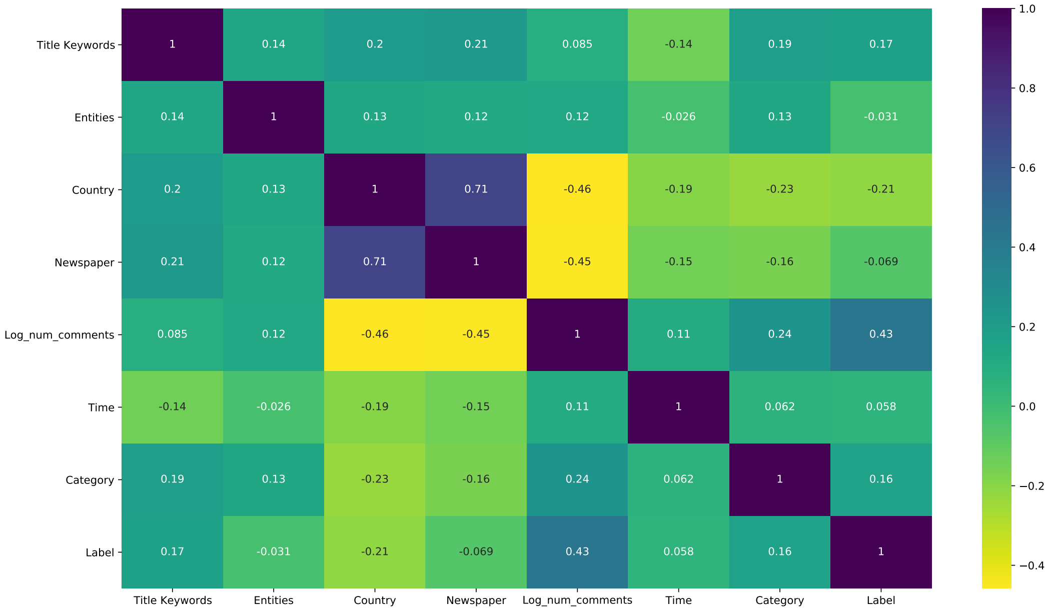

Figure 5.29: Correlation matrix of ***m_base+title*** model.

0.742. Furthermore, it could be observed that this algorithm has a better capacity in identifying a considerable amount of toxicity-generating news due to its high Recall value.

On the other hand, the random forest classifier has a higher Precision value, indicating that it has higher reliability in the identification of the positive class. We could examine the frequent trade-off between Precision and Recall. In situations such as the identification of toxicity-generating news, we share the opinion that it is of higher importance, having a higher Recall since we want to identify as much toxicity-generating news as possible. Although this decision would lead to a higher number of false positives, a simple solution could be the addition of a filter that would in a latter stage refine improperly identified news.

Comparing our results with the studies reviewed in the state of the art in Section 3.3, we observed that relative to the work on uncivil speech in political news [48], from the Precision and Recall scores reached by the researchers, we could calculate that their work achieved an F1 score of 0.639. We concluded that, although their research engages different datasets and phenomena, both our and their work involves the prediction of harmful behaviour in news. Furthermore, we use similar metrics for the classification of comments and news. Having our work a better performance, we can suppose that having meta-data from news articles, such as the news source and the number of comments are significant points for an increase in the ability to predict harmful behaviour from news. Regarding the results reported in the study on incivility generation in Facebook pages of German news [21; 81] the improved model reached an F1 score of 0.68. Despite the two mentioned incivility studies having different classification metrics, data and language, one possible factor for the enhanced results of the German research may be the presence of meta-information since the latter includes features such as the number of likes and number of words present in comments. This assumption can be further extended to the improvements in results our study reports since our work makes no use of word n-grams and PoS that were present in the study regarding German news.

Investigating our model's feature correlation, we could recognize that no feature had a strong correlation with the target variable, meaning that the performance of the model was not due to a set of features but rather the combination of all. Further analyzing the feature importance from each of the algorithms, we observed that the number of comments, newspaper and title keywords were the main features that contributed the most for the identification of toxicity-generating news. The presence of newspapers as features with significant positive and negative importance indicates that the cultural background and journalistic ideology are factors that impact the generation of toxicity in news. The presence of controversial terms such as "racist", "nazi" and "gay" suggests that news regarding highly debated social topics tends to cause toxicity. Lastly, the presence of politic entities such as "Trump", reveals that political news and topics lead to an increase of toxicity in news.

Finally, the work done in these experiments can be further improved. In future work, we prioritize the improvement of the news title and body features by changing the representation from BoW to TF-IDF and bag-of-bigrams, since *"Bag-of-words approaches tend to have high recall but lead to high rates of false positives"* [19] and *"a bag-of-bigrams representation is much* *more powerful than bag-of-words, and in many cases proves very hard to beat"* [30]. Furthermore, regarding the news title feature, other more complex text mining features can be explored, such as word embeddings, which are commonly used in NLP. Additionally, more features regarding textual data can be extracted from the news body, such as text length and keywords instead of the named entities used.

Chapter 6

Conclusions and future work

The present chapter has the intent of reporting the work conveyed in this thesis. We analyse the goals set and their achievement, reviewing the results returned. Additionally, we analyse future improvements to work done and discuss points susceptible to discussion.

6.1 Goals of our work

The main goal of this thesis was to contribute to the field of prediction of toxicity-generation news. To achieve our primary purpose, we branched out smaller objectives, the first being the exploration of the state of the art of the field of study. From the overview of the state of the art, we concluded that there had been vast research in various forms of harmful communication such as offensive, hate and uncivil speech. With the intent of correctly detecting toxicity-generating news, we started by understanding the differences between toxicity and these terms and how they overlap. With this, we summarized the differences and similarities of each definition and proposed our definition of toxicity, to which we would guide our work.

Additionally, we concluded that although offensive and hate speech have numerous published studies in regards to their automatic detection in text, we found that research regarding incivility and toxicity detection in news has been less studied. Furthermore, the review found in our area of research begins with the composition of datasets for the purpose of their study, indicating the lack of publicly available datasets for the classification of toxic or uncivil behaviour. This lack of public datasets not only slows the progress of research in these areas but difficulties the sharing of results, hindering the development in the field. Since there is little work in the area of toxicity-generating news, we chose to investigate research done in the area of news classification. This way, we could investigate standard news extraction techniques and algorithms used in the area, such as TF-IDF and BoW with SVM and naive bayes.

Since there has been no work done in the area of prediction of toxicity-generating news, we make use of a dataset constructed for the prediction of hate generating news [10] in an earlier stage of the *Stop PropagHate* project [2] our work also is inserted in. The data gathered represented a set of 72,342 news articles with 3,026,270 comments from Twitter users relative to those news articles. With a second objective of investigating the toxicity phenomenon in our dataset, and with the intent of focusing the classification task to news and not comments, we took the decision of using *Perspective API* to identify comments as toxic. With the conclusion of the classification of comments, we determined that news having a toxicity percentage higher than the dataset median toxicity percentage (11.1%) would be flagged as toxicity-generating. The findings returned by the classification of comments and respective news as toxic, suggests that the cultural variances and distinct orientation to which journalists draw news articles contributes to the readers' toxic behaviour. Furthermore, the analysis from toxicity in news implied that having a larger number of news did not directly lead to a larger number of toxic comments. This fact suggests that the newspaper personality and target audience poses as a main contributing factor to the growth of toxicity in news.

A final goal of this thesis was to investigate which features contributed to the identification of toxicity-generating news. To achieve this goal, we started by selecting features that would be experimented in various models throughout our work. We extracted meta-data features such as the news country, newspaper, date of publication and number of comments. Additionally, news-content features were extracted, such as the news titles named entities, keywords, topic category and news body entities. Throughout the experiments done with feature combination, we concluded that our best models represented a combination of news-content features and meta-data features. We experimented with a variety of machine learning algorithms chosen based on the standard use in the area of text and news classification. The best performing model was the GBDT, reaching an F1 score of 0.743. This algorithm also identifies the highest number of positive cases of toxicity-generating news. This is due to the fact the GBDT classifier reaches the highest Recall value, despite having a lower Precision compared to other algorithms. In situations such as the identification of toxicity-generating news, we benefit a higher Recall, since this meant identifying as much toxicity-generating news as possible. Although privileging more Recall would lead to a higher number of false positives, a simple solution could be the addition of a filter that would in a latter stage refine improperly identified news.

Having our best performing model, we started with the investigation of which features had higher contributions to the classification of toxicity-generating news. We analysed the correlation matrix, which indicated that the resulting performance was not related to a subset of features, but rather the contribution of all used features. Such conclusion was due to the low correlation coefficient between all features and the target variable, with only the number of comment feature having a higher correlation (0.43). Additionally, we studied the feature importance for all four algorithms. From the analysis, we concluded that from the meta-data features, the newspaper and number of comments were the features that significantly contributed to the identification of toxicity-generating news. This fact suggested that the cultural background and journalistic ideology were factors that impacted the generation of toxicity in news. Regarding news-content features, news title keywords such as "racist", "nazi" and "gay" suggested that news regarding highly debated social topics tend to cause toxicity. Moreover, the presence of political entities such as “Trump” as contributing features to the classification task suggested that political news and topics lead to an increase of toxicity in comments relative to that news. With all this said we can assert that the objectives we established in the beginning of this thesis we all met.

From the work developed in this thesis, we encountered some challenges namely the high dimensionality of some news-content features such as the news title keywords and body entities, which due to this reason we decided only to consider the 1,000 most frequent keywords. The same decision applied to the news body entities. Additionally, the imbalanced nature of the distribution of comments and news between newspapers makes the distribution less homogeneous, leading to a difficulty towards classification.

6.2 Future work

Although the study and identification of offensive terms are vastly researched, such as the case of hate speech, research on toxicity in news has been less explored. This fact represents an opportunity for the investigation of new approaches, namely with the use of novel deep learning algorithms that have proven to be effective in similar areas such as hate speech detection. As exposed in this thesis, several terms such as hate and offensive speech share similarities with toxicity. Since such definitions have a subjective nature attached to their interpretation, we believe that an opening point in the research of toxicity in news is the uniformization of the meaning of toxicity. The provision of guidelines for the uniformization would provide a more transparent and homogeneous definition, which would consequently transpose to the models, leading to more effective identification of the phenomenon.

Furthermore, the complex subjective nature of the field in study proves that providing clear guidelines would improve the learning process of annotators with the task of labelling datasets gathered to the study of toxicity. A second opportunity would be the development of public datasets based on established definitions of toxicity. Public datasets would benefit research in this field of study since novelty approaches could compare results to other studies that use the same data.

Regarding the experiments described in this thesis, future improvements can be extended. Firstly the news-content features such as the news title keywords and body entities can be changed from BoW to TF-IDF or bag-of-bigrams. This may be a good improvement since *“Bag-of-words approaches tend to have high recall but lead to high rates of false positives”* [19] and *“a bag-of-bigrams representation is much more powerful than bag-of-words, and in many cases proves very hard to beat”* [30]. Furthermore, despite the dataset used in our work not having user meta information, the extraction of features regarding such data can be beneficial and a possible point for further development of the model.

Finally, the gain of popularity of deep learning may in the area of text classification might be of interest for further developments. This is due to the reports of such approaches having better result when compared to machine learning algorithms. The experimentation of novel algorithms using deep learning can contribute to a step towards a better classification of toxicity in news.

References

[1] Conversation ai. Available at https://conversationai.github.io/. Online (accessed March 20, 2020).

[2] Stop propaghate. Available at http://stop-propaghate.inesctec.pt/. Online (accessed March 20, 2020).

[3] T. G. Almeida, B. A. Souza, F. G. Nakamura, and E. F. Nakamura. Detecting hate, offensive, and regular speech in short comments. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, WebMedia '17, pages 225–228, New York, NY, USA, 2017. ACM.

[4] H. Almerekhi, B. J. Jansen, H. Kwak, and J. Salminen. Detecting toxicity triggers in online discussions. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, HT 2019, pages 291–292, Hof, Germany, 2019. ACM.

[5] A. A. Anderson, D. Brossard, D. A. Scheufele, M. A. Xenos, and P. Ladwig. The "nasty effect:" online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, 19(3):373–387, 4 2014.

[6] U. Ansaldo, J. Don, and R. Pfau, editors. *Parts of Speech: Empirical and theoretical advances*. John Benjamins, Amsterdam, Philadelphia, 1st edition, 2010.

[7] A. Antoci, A. Delfino, F. Paglieri, F. Panebianco, and F. Sabatini. Civility vs. incivility in online social interactions: An evolutionary approach. *PLOS ONE*, 11(11):1–17, 11 2016.

[8] J. B. Jacobs. Hate crime: Criminal law and identity politics: Author's summary. *Theoretical Criminology - THEOR CRIMINOL*, 6:481–484, 11 2002.

[9] X. Bai, X. Merenda, C. Zaghi, T. Caselli, and M. Nissim. Rug @ EVALITA 2018: Hate speech detection in italian social media. In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018) co-located with the Fifth Italian Conference on Computational Linguistics*, CLiC-it2018, Turin, Italy, 12 2018.

[10] R. Barros. Predicting the impact of news stories in reactions containing hate speech. Master's thesis, Faculdade de Ciências da Universidade do Porto, 11 2019.

[11] E. Berk and E. Filatova. Incendiary news detection. In *Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference*, FLAIRS-32, pages 161–166, Sarasota, USA, 5. AAAI Press.

[12] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[13] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 7 2017.

[14] P. Bourgonje, J. Moreno-Schneider, A. Srivastava, and G. Rehm. Automatic classification of abusive language and personal attacks in various forms of online communication. In G. Rehm and T. Declerck, editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10713 LNAI, pages 180–191, Berlin, Germany, 2018.

[15] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Neural Information Processing Systems 2014 Workshop on Deep Learning*, NIPS 2014, Montreal, Canada, 12 2014.

[16] K. Coe, K. Kenski, and S. A. Rains. Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64(4):658–679, 8 2014.

[17] C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273—-297, 9 1995.

[18] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani. 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, pages 112–116, Washington, USA, 9 2016. IEEE Computer Society.

[19] T. Davidson, D. Warmsley, M. W. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th Internacional Conference on Web and Social Media*, ICWSM-17, pages 512–515, Montereal, Canada, 2017. AAAI Press.

[20] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233—-240, New York, USA, 2006. Association for Computing Machinery.

[21] J. Daxenberger, M. Ziegele, I. Gurevych, and O. Quiring. Automatically detecting incivility in online discussions of news media. In *Proceedings - IEEE 14th International Conference on eScience, e-Science 2018*, pages 318–319, Washington, USA, 11 2018. IEEE Computer Society.

[22] R. P. de Pelle, C. Alcântara, and V. P. Moreira. A classifier ensemble for offensive text detection. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, WebMedia '18, pages 237–243, New York, USA, 2018. ACM.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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, pages 4171–4186, Minneapolis, USA, 6 2019. Association for Computational Linguistics.

[24] I. Dilrukshi, K. De Zoysa, and A. Caldera. Twitter news classification using svm. In *Proceedings of the 8th International Conference on Computer Science and Education*, ICCSE 2013, pages 287–291, Washington, USA, 2013. IEEE Computer Society.

[25] Y. Dinkov, I. Koychev, and P. Nakov. Detecting toxicity in news articles: Application to bulgarian. In *International Conference Recent Advances in Natural Language Processing, RANLP*, pages 247–258, Varga, Bulgaria, 9 2019. Incoma Ltd., Shoumen, Bulgaria.

[26] M. Duggan. Online harassment. Technical report, Pew Research Center, Washington, USA, 10 2014. Available at https://radimrehurek.com/gensim/summarization/keywords.html.

[27] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding. In *Proceedings of the Twelfth International Conference on Web and Social Media*, ICWSM 2018, Standford, USA.

[28] K. Erjavec and M. Poler Kovačič. "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society*, 15:899–920, 11 2012.

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

[30] Y. Goldberg. *Neural Network Methods for Natural Language Processing*, volume 37 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool, San Rafael, USA, 2017.

[31] S. Helmstetter and H. Paulheim. Weakly supervised learning for fake news detection on twitter. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM 2018, pages 274–277, Washington, USA, 10 2018. IEEE Computer Society.

[32] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735––1780, 11 1997.

[33] M. Honnibal and I. Montani. spacy named entities. Available at https://spacy.io/usage/spacy-101#annotations-ner. Online (accessed March 20, 2020).

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

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

[36] ILGA-Europe. Hate crime hate speech. Available at https://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech, 2019. Online (accessed April 20, 2019).

[37] I. F. Ilyas and X. Chu. *Data Cleaning*. Association for Computing Machinery, New York, USA, 1st edition, 2019.

[38] T. Jay and K. Janschewitz. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2), 1 2008.

[39] Jigsaw and google's counter abuse technology team. Perspective api. Available at https://www.perspectiveapi.com/#/home. Online (accessed February 13, 2020).

[40] Jigsaw and google's counter abuse technology team. Perspective api attributes. Available at https://github.com/conversationai/perspectiveapi/blob/master/2-api/models.md. Online (accessed February 13, 2020).

[41] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, 4 2017. Association for Computational Linguistics.

[42] D. Jurafsky and J. H. Martin. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J., 1st edition, 2009.

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

[44] J. Kazil and K. Jarmul. *Data Wrangling with Python: Tips and Tools to Make Your Life Easier*. O'Reilly Media, Inc., 1st edition, 2016.

[45] M. Kuhn and K. Johnson. *Applied predictive modeling*. Springer-Verlag, New York, USA, 1st edition, 2013.

[46] P. Langley and S. Sage. Induction of selective bayesian classifiers. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, UAI'94, pages 399—-406, San Francisco, USA, 1994. Morgan Kaufmann Publishers Inc.

[47] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio. Object recognition with gradient-based learning. In *International Workshop on Shape, Contour and Grouping in Computer Vision*, pages 319–345, Berlin, Germany, 5 1999. Springer Verlag.

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

[49] S. Malmasi and M. Zampieri. Challenges in discriminating profanity from hate speech. *Journal of Experimental and Theoretical Artificial Intelligence*, 30(2):187–202, 3 2018.

[50] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 1st edition, 2008.

[51] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Conference Track Proceedings, 1st International Conference on Learning Representations*, ICLR 2013, pages 1–12, Scottsdale, USA, 5 2013.

[52] D. Milošević, Y. Tang, and Q. Zu. *Human Centered Computing: 5th International Conference, HCC 2019, Čačak, Serbia, August 5–7, 2019, Revised Selected Papers*. Lecture Notes in Computer Science. Springer International Publishing, Cacak, Serbia, 1st edition, 2020.

[53] R. Misra. News category dataset. https://www.researchgate.net/publication/332141218_News_Category_Dataset, 06 2018. Online (accessed April 6, 2020).

[54] H. Mubarak, K. Darwish, and W. Magdy. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, Canada, 8 2017. Association for Computational Linguistics.

[55] N. Naaz, Y. Malik, and K. P. Adhiya. Hate speech detection in twitter-a survey. *International Journal of Management, Technology And Engineering*, 9(1):1272–1277, 2019.

[56] A. Natekin and A. Knoll. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7:21, 2013.

[57] N. Newman, R. Fletcher, K. Antonis, D. Levy, and R. Nielsen. The 2017 digital news report. Technical report, Reuters Institute, Washington, USA, 6 2017. Available at https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf.

[58] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW ’16, pages 145–153, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[59] J. T. Nockleby. Hate speech. *Encyclopedia of the American Constitution, (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al.)*, pages 1277–1279, 6 2000.

[60] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1):1—135, 1 2008.

[61] H.-A. Park. An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, 43:154–164, 4 2013.

[62] J. Pavlopoulos, N. Thain, L. Dixon, and I. Androutsopoulos. 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*, SemEval-2019, pages 571–576, Minneapolis, USA, 6 2019. Association for Computational Linguistics.

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

[64] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2014, pages 1532–1543, Doha, Qatar, 10 2014. ACL.

[65] D. M. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, 2(1):37–63, 2011.

[66] R. Řehůřek. Gensim keyword textrank. https://radimrehurek.com/gensim/summarization/keywords.html. Online (accessed March 20, 2020).

[67] R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 5 2010. ELRA. http://www.lrec-conf.org/proceedings/lrec2010/workshops/W10.pdf.

[68] S. Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520, 2004.

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

[70] T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):1–21, 3 2015.

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

[72] V. B. Sharma. Rake using nltk. https://csurfer.github.io/rake-nltk/_build/html/_modules/rake_nltk/rake.html. Online (accessed March 20, 2020).

[73] C. B. Stone and Q. Wang. From conversations to digital communication: The mnemonic consequences of consuming and producing information via social media. *Topics in Cognitive Science*, 11(4):774–793, 2019.

[74] U. Suleymanov, S. Rustamov, M. Zulfugarov, O. Orujov, N. Musayev, and A. Alizade. Empirical Study of Online News Classification Using Machine Learning Approaches. In *IEEE 12th International Conference on Application of Information and Communication Technologies*, pages 1–6, Washington, USA, 10 2018. IEEE Computer Society.

[75] Twitter. Hateful conduct policy. Available at https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy, 2019. Online (accessed April 20, 2019).

[76] M. Verleysen and D. François. The curse of dimensionality in data mining and time series prediction. In *Proceedings of the 8th International Conference on Artificial Neural Networks: Computational Intelligence and Bioinspired Systems*, IWANN'05, pages 758—-770, Berlin, Heidelberg, 2005. Springer-Verlag.

[77] M. Volkovs, G. W. Yu, and T. Poutanen. Content-based neighbor models for cold start in recommender systems. In *Proceedings of the Recommender Systems Challenge 2017*, RecSys Challenge '17, pages 1–6, New York, USA, 2017. Association for Computing Machinery.

[78] Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the North American Association for Computational Linguistics Student Research Workshop*, pages 88–93, San Diego, California, 6 2016. Association for Computational Linguistics.

[79] E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1391–1399, New York, USA, 2017. ACM Press.

[80] YouTube. Hate speech policy. Available at https://support.google.com/youtube/answer/2801939?hl=en, 2019. Online (accessed April 20, 2019).

[81] M. Ziegele, J. Daxenberger, O. Quiring, and I. Gurevych. Developing automated measures to predict incivility in public online discussions on the facebook sites of established news

media. In *Proceedings of the 68th Annual Conference of the International Communication Association (ICA)*, ICA18, Prage, Czech Republic, 5 2018. Advanced copy. https://public.ukp.informatik.tu-darmstadt.de/UKP_Webpage/publications/2018/2018_ICA_Ziegele_DevelopingAutomatedMeasures.pdf.