electronics

Article

Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit

Hong Fan [1,*,†], Wu Du [1], Abdelghani Dahou [2], Ahmed A. Ewees [3], Dalia Yousri [4], Mohamed Abd Elaziz [5,6], Ammar H. Elsheikh [7], Laith Abualigah [8,9] and Mohammed A. A. Al-qaness [1,†]

[1] State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; d.y.wu.du@gmail.com (W.D.); alqaness@whu.edu.cn (M.A.A.A.-q.)
[2] LDDI Laboratory, Faculty of Science and Technology, University of Ahmed DRAIA, Adrar 01000, Algeria; Dahou.abdghani@univ-adrar.dz
[3] Department of Computer, Damietta University, Damietta 34511, Egypt; ewees@du.edu.eg
[4] Electrical Engineering Department, Faculty of Engineering, Fayoum University, Fayoum 63514, Egypt; day01@fayoum.edu.eg
[5] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt
[6] School of Computer Science and Robotics, Tomsk Polytechnic University, 634050 Tomsk, Russia; abd_el_aziz_m@yahoo.com
[7] Department of Production Engineering and Mechanical Design, Tanta University, Tanta 31527, Egypt; ammar_elsheikh@f-eng.tanta.edu.eg
[8] Faculty of Computer Sciences and Informatics, Amman Arab University, Amman 11953, Jordan; Aligah.2020@gmail.com
[9] School of Computer Science, Universiti Sains Malaysia, Penang 11800, Malaysia
* Correspondence: hfan3@whu.edu.cn
† These authors contributed equally to this work.

Citation: Fan, H.; Du, W.; Dahou, A.; Ewees, A.A.; Yousri, D.; Elaziz, M.A.; Elsheikh, A.H.; Abualigah, L.; Al-qaness, M.A.A. Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit. *Electronics* **2021**, *10*, 1332. https://doi.org/10.3390/electronics10111332

Academic Editors: Rashid Mehmood, Alessandro Lameiras Koerich and Savvas A. Chatzichristofis

Received: 22 April 2021
Accepted: 28 May 2021
Published: 1 June 2021

Abstract: Social media has become an essential facet of modern society, wherein people share their opinions on a wide variety of topics. Social media is quickly becoming indispensable for a majority of people, and many cases of social media addiction have been documented. Social media platforms such as Twitter have demonstrated over the years the value they provide, such as connecting people from all over the world with different backgrounds. However, they have also shown harmful side effects that can have serious consequences. One such harmful side effect of social media is the immense toxicity that can be found in various discussions. The word toxic has become synonymous with online hate speech, internet trolling, and sometimes outrage culture. In this study, we build an efficient model to detect and classify toxicity in social media from user-generated content using the Bidirectional Encoder Representations from Transformers (BERT). The BERT pre-trained model and three of its variants has been fine-tuned on a well-known labeled toxic comment dataset, Kaggle public dataset (Toxic Comment Classification Challenge). Moreover, we test the proposed models with two datasets collected from Twitter from two different periods to detect toxicity in user-generated content (tweets) using hashtages belonging to the UK Brexit. The results showed that the proposed model can efficiently classify and analyze toxic tweets.

Keywords: toxic; social media; brexit; twitter; BERT; sentiment analysis

1. Introduction

Social media has many positive aspects, one of which is the sense of community it provides to people [1]. Social media allows people to do in a day what would usually take a lifetime to achieve. It also gives people the opportunity to form a support network. People from all walks of life around the world can connect with the right individuals and build a mutually beneficial network. Social media also provides the ability to see what people in other parts of the world are thinking or doing. This can be very beneficial, for example, in instances where someone wants to see the immediate reactions people have during an election, sports game, award show, etc. Providing an outlet for self-expression is another area where social media excels. Most people are looking to frequently express themselves in some way, shape, or form. Some use social media as a sort of therapy, allowing them to talk about what is bothering them. In [2], the authors examine the benefits of social media in adolescence. One of the benefits they point out is identity exploration, which they say can help adolescents discover aspects of themselves.

Many applications have been addressed from user-generated contents in social media such as personal classification based on profiles in social media [3], group decision making [4], aggregation identification [5], domestic violence detection [6,7], clustering analysis [8,9], extremist affiliations [10], prediction and analysis of elections [11–13], analysis of happiness patterns [14], education evaluation [15], football events [16], multimedia summarization [17,18], and many other applications [19–22].

Despite the fact that social media offers a lot of good to the world, it also has a host of negative aspects. Social media has become a place of discord. People rarely agree on matters of discussion, and some of the individuals on social media take these disagreements to a higher intensity. They attack anyone who is at odds with what they believe in. This can lead to insensitive language being used from one person or group to another. Topics of contention range from politics, gender, movies, and more. People who want to discuss things they care about must be willing to receive opinions different from theirs. Said opinions can sometimes be toxic. This sometimes leads to users no longer expressing themselves, and eventually may halt the search for differing opinions because of the threat of abuse and harassment on social media. Platforms such as Twitter have had difficulty in trying to effectively facilitate conversations. This has led to large numbers of communities limiting or outright shutting down user comments.

Toxic behavior on social media has come to be expected, but it is increasingly not tolerated. Toxicity within the social sphere can be described as spreading unnecessary negativity or hate that ends up negatively affecting people who encounter it. Toxic individuals online look to spread malice and abuse other people in discussions. For instance, Kwak et al. [23] studied toxic behavior in team competition online games. They found that the result of a match is tied to the appearance of toxic behavior. Toxic comments on social sites, such as Twitter can be found on topics that are very difficult to discuss, such as Brexit, climate change, abortion, vaccines, and US elections. Toxic behavior is more prevalent in such topics because of their divisive nature. People tend to have different opinions when discussing such topics, which can lead to divisions. Groups of people that believe in a particular view are formed, with each group believing their views are right. There is rarely a middle ground in these clashes of words. Some people are civil when discussing such topics, but more often than not, people become frustrated by the other group and start using toxic language. A comment about climate change such as "They're stupid, it's getting warmer, we should enjoy it while it lasts", can be considered toxic. The previously mentioned comment, made by someone that believes in climate change, is a blatant attack on individuals that deny that climate change is happening.

Toxic behavior on social platforms can also indicate the presence of cyber-bullying. Cyberbullying is a type of bullying or harassment that is carried out online and is widespread on social media sites such as Twitter. It has become a common occurrence among people of all ages, but it mostly occurs amongst teenagers. According to a particular poll, 22% of teens use their preferred social media site more than 10 times a day [24]. Whittaker and Kowalski [25] carried out a study revealing that texting and social media are the most commonly used venues for cyberbullying. According to (Bullying statistics. http://www.bullyingstatistics.org/content/cyber-bullying-statistics.html (accessed on 1 January 2021)), over 25% of teens and adolescents have experienced cyberbullying, with most choosing not to tell their guardians when it occurs. Therefore, cyberbullying can be described as when individuals, usually teens, bully or harass others on social media. Cyberbullying is inherently toxic and harmful. Examples of harmful bullying behavior are threats, hate speech, posting vile rumors, making unsolicited sexual comments, as well as giving out someone's personal information without their permission. Fox et al. [26], for instance, show the effects of sexism and sexual harassment on social media. They deduce that a major factor is the anonymity of users. If left unchecked, cyberbullying can lead to low self-esteem in victims. In more extreme cases, victims end up committing suicide. Other emotional responses to such abuse include anger, frustration, depression, and fear for one's life.

Sometimes the source of these toxic comments are individuals who are known as internet trolls, or simply trolls. A troll is an individual who deliberately initiates quarrels or angers people online in an attempt to distract and stir up discord by posting provocative, off-topic comments to incite an emotional response from readers, either for their own amusement or to serve a specific goal. This act is known as trolling. The acts of trolls can be explained by the online disinhibition effect [27,28], which is the lowering of one's behavioral inhibitions in the online sphere. Trolls benefit significantly from the anonymity they have on social sites. Trolls can be detrimental to online communities in several ways by disrupting discussions (e.g., on Twitter), spreading lousy advice, and damaging the trust developed over time in an online community. Moreover, when the rate of deception is high, a group can become sensitized to trolling. This can lead to the rejection of honest, naive questions because they are considered trolling. In some situations, when a new user on a site such as Twitter decides to make their first post, they are immediately flooded with angry accusations. Despite the possibility that the accusations might be uncalled for, being labeled a troll can be quite harmful to one's online reputation.

All things considered, toxic behavior is an issue that needs to be dealt with head-on to foster civil, healthy, and open discussions on social media. That said, it is up to social platforms to decide the method of resolving this issue. Some have resorted to permanently banning individuals that were reported by other users, while others have chosen to let this behavior run rampant. The removal of toxic comments from social media can have a substantial positive impact, not just on conversations, but for people that may have been on the receiving end of said toxic comments. One solution to dealing with toxicity is through the use of sentiment analysis. A sentiment analysis system can be used to detect toxic comments by classifying the likelihood of such text as being toxic. Sentiment analysis has proven to be a successful approach to solving problems in numerous domains such as in [29–35]. In addition, optimization techniques can be used to optimize the classification parameters [36,37].

The goal of this study is to analyze and to detect toxic behavior in Twitter using user-generated content in social media, such as Twitter. Twitter sentiment analysis has received wide attention, and has been utilized in various domains, for example, political influences [38–40], consumer insight mining [41], transportation services [42], movements of stock markets [43], traffic congestion detection [44,45], happiness evaluation [46], and others [47].

Therefore, we propose a method to detect toxicity in social media that involves building a deep learning model that is trained on a toxic dataset and then tested with real-time tweets that were collected using the Twitter API.

In this study, a toxicity detection and classification model was built based on Bidirectional Encoder Representations from Transformers (BERT). The BERT is a language representation scheme proposed by Devlin et al. [48], which is considered an efficient method that is widely employed in various applications, and has showed excellent performance on different tasks.

To sum up, the primary contributions of this paper are as follows:

1. Propose an efficient deep learning model to classify toxic comments from user generated contents in social media;
2. Build our model based on the Bidirectional Encoder Representations from Transformers (BERT) model;
3. The BERT pre-trained model is fine-tuned on a Kaggle public dataset, "Toxic Comment Classification Challenge", and has been evaluated on two different datasets,

collected from Twitter in two different periods, using Twitter API by applying several search terms and hashtags such such as #Brexit, #BrexitBetrayal, and #StopBrexit;

4. We compare the BERT base model to three models, namely, Multilingual BERT, RoBERTa, and DistilBERT to verify its performance.

This paper is organized as follows. Related works are presented in Section 2. Section 3 describes the methodology. In Section 4, the experimental evaluation is described, where the conclusion is presented in Section 5.

2. Related Work

2.1. Applications of User Generated Contents in Social Media

Souri et al. [3] presented a personality classification model by analyzing user activities in Facebook using machine learning. The proposed method uses Facebook API to collect the data of 100 users. The proposed method also can recommend friends in Facebook groups. Morente-Molinera et al. [4] leveraged a sentiment analysis technique to address group decision making using social media. The proposed method helps experts to generate preference relations which may be applied to make a group decision. Risch and Krestel [5] proposed a deep learning based model to identify the aggregation of user-generated content in social media. The proposed method uses a recurrent neural network based on a bi-directional gated recurrent unit. They used a machine translated to augment a dataset used for model training.

Subramani et al. [6] proposed a domestic violence identification system using deep learning. The proposed system uses a dataset collected from Facebook. The proposed system uses a binary text classification approach that can detect if a content created by a user is recognized as critical or uncritical. Subramani et al. [7] also presented a deep learning-based domestic violence identification system. Unlike their previous work [6], this system can be applied for a multi-class posts categorization including five categories, namely general, empathy, awareness, personal story, and fund raising. The proposed system was evaluated using a dataset collected from Facebook users' generated contents and achieved a high accuracy rate.

Ahmad et al. [10] used a deep learning model with sentiment analysis technique to classify user-generated content on Twitter (tweets) as a binary classification (extremist tweet or non-extremist tweet). In their work, Budiharto and Meiliana [11] used sentiment analysis techniques to predict the result of the Indonesian presidential election. The prediction of the proposed method was correct according to the result of the Indonesian presidential election. Al Shehhi et al. [14] used sentiment analysis techniques to analyze tweets collected from Twitter users in the United Arab Emirates (UAE) to measure happiness. They used both English and Arabic tweets. They found that 7:00 am is the happiest hour, and Friday is the happiest day.

In [15], the authors proposed a sentiment analysis system that can be applied to analyze posts in teaching evaluation systems. The proposed system collects student feedback, and analyzes sentiment analysis phrases to obtain the classification of teaching attitude.

Aloufi and El Saddik [16] proposed a sentiment analysis approach to analyze football fans sentiments through tweets posted by them and their reaction to game events (i.e., penalty kick, scoring goals, etc.). The proposed approach can conclude fan interaction through football game events.

Ibrahim et al. [49] proposed a toxic detection model based on convolutional neural network (CNN), bidirectional gated recurrent units (GRU), and bidirectional long short-term memory (LSTM). They used a Wikipedia dataset to evaluate the proposed method and achieved an F-1-score of 87.2% for predicting toxicity types and 82.2% for the classification of toxic/non-toxic. In [50], the authors presented personal attacks analysis method based on a combination of machine learning and crowdsourcing. In [51], the authors proposed a deep learning-based approach to classify toxic comments. They used Kaggel data to test their approach.

Google and Jigsaw presented an API (Perspective API, https://www.perspectiveapi.com/ (accessed on 10 February 2021)) for classifying toxic comments. However, this API handles a simple binary text. Moreover, another study addressed the classification of toxic comments presented by [52]. This study also used a simple binary text classification to test the proposed model.

2.2. *Applications of Bidirectional Encoder Representations from Transformers (BERT)*

The BERT has received wide attention, and has been applied in numerous applications. Fang et al. [53] proposed a near-miss reports classification method based on the BERT model. They evaluated BERT using near-miss reports datasets that collected from real-world construction projects. They found that BERT has the ability to classify near misses from the datasets. Additionally, BERT outperforms other compared models. Fan et al. [54] applied BERT to detect adverse drug events. They used reviews from Drug.com and WebMD to detect unreported drug events. The evaluation outcomes showed that the BERT achieved 94% of AUC. Moradi et al. [55] presented a summarization method using contextualized embeddings generated by the BERT. Their model was applied to capture the context of the text of medical texts. The evaluation of the method confirmed that BERT improved the summarization for biomedical text.

Wang et al. [56] used different methods to normalize Chinese written procedure and diagnosis to the standard concepts in ICD (International Classification of Diseases). Among five well-known methods, the BERT showed the best performances for normalization of procedure and diagnosis. In addition, in [57], the authors presented a BERT-based model to measure semantic similarity of clinical trial outcomes. Moreover, another text analysis approach for medical applications was proposed by Zhang et al. [58] using BERT. In this study, they used Chinese clinical information of various types of notes for breast cancer. The BERT was evaluated with extensive comparisons to other models and it showed better performances.

Chen et al. [59] proposed a multi-source data fusion approach for aspect-level sentiment classification. Their proposed approach integrates data from word-level sentiment lexicons, sentence-level corpora, and aspect-level corpora using the power of the BERT model.

Furthermore, BERT is applied in many other applications, such as the classification of target-dependent sentiment [60], entity linking [61], image classification [62], medical text inferencing [63], occupational title embedding [64], and others [65]. Moreover, multiple transformer-based models have been developed such GPT3 [66,67], Megatron-lm [68], and Electra [69] which demonstrated remarkable performance.

3. Methodology

In this section, we build a classifier able to classify toxic comments from a selection of collected Twitter posts. In this work, the BERT pre-trained model and three of its variants has been fine-tuned on a well-known labeled toxic comment dataset (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge (accessed on 1 January 2021)) and used to classify the collected tweets. The methodology workflow consists of three main steps, (i) implementation and fine-tuning of the BERT model for toxic comment classification, (ii) collection, pre-processing, and characterization of tweets relevant to a pre-defined hashtag, and (iii) classification analysis in the toxicity trend of the collected tweets.

3.1. *Pre-Processing*

Posts originated from a particular hashtag in a specific period. Before building the model, we implemented light pre-processing steps. We removed punctuation, links, and non-English words. For tokenization, we also used the pre-trained tokenizer for “bert_base_uncased”.

3.2. BERT for Toxic Comments Classification

Transformer-based models are dominating a wide variety of NLP tasks leading to many state-of-the-art results [70]. Based on transformer architecture, BERT (Bidirectional Encoder Representations from Transformers) has been recently proposed by a Google AI language research team as a language model that has been trained on two tasks, which are masked tokens prediction (Masked LM (MLM)) and next sentence prediction (NSP). Transformers are based on the popular attention mechanism to perform language modeling. Previous techniques only generate their embeddings from a text sequence by either looking to it from left to right, combined left to right, or right to left during training. Contrary, BERT applies the bidirectional training of the transformer to devise and gain a deeper understanding of the language context and flow.

Since transformers integrate the attention mechanism, which incorporates two separate mechanisms (encoder and decoder), BERT makes use of the encoder only to generate a language model. The encoder is used to learn contextual relations where the input is a sequence of tokens (words or sub-words) embedded into vectors. Before converting words into vectors and feeding them to BERT for training the masked LM, word masking is applied. Training the Masked LM is done by predicting 15% of the randomly chosen tokens in each sequence. Where 80% of the chosen tokens are replaced with a [MASK] token, 10% with a random token, and the remaining 10% remain unchanged. At this stage, the model suffers from slow convergence compared to directional models as BERT focuses only on predicting the masked values when calculating the loss function and ignores the prediction of the non-masked tokens in each batch. In the NSP training process, pairs of sentences (A, B) are inputted to the model where the goal is to predict if the sentence B is the next sentence in the original document. At this stage, 50% of the inputs are paired in which sentence B is the subsequent sentence of sentence A in the original document. Whereas, the other 50% are pairs where sentence B is a random sentence from the corpus.

In this paper, BERT-base trained on a single language (English) is used during the fine-turning phase with its default configuration that has 12 encoder blocks, 768 hidden dimensions, 12 attention heads, 512 maximum sequence length, and a total of 110 M parameters [48]. In addition, the BERT multilingual base is a model trained on 104 languages using Wikipedia text and the MLM technique [48]. RoBERTa [71] is an optimized model based on Google's BERT where the authors modified the hyperparameters (mini-batche size and learning rate) and omitted the NSP objective. DistilBERT [72] is a model trained by distilling Google's BERT which reduced the number of parameters by 40% (66 M) compared to the BERT-based method and ran faster, almost matching BERT's performance.

3.3. Model Fine-Tuning for Toxic Comments Classification

The BERT-base model is fine-tuned on documents retrieved from Wikipedia by the Kaggle competition. The dataset is from the Toxic Comment Classification Challenge, which was provided by the Conversation AI team. Conversation AI is a research initiative that was formed by Google and Jigsaw, and both are a part of Alphabet Inc.

The provided class labels in the dataset were originally defined across six different types of toxicity, including toxic, severe toxic, obscene, threats, insults, and identity-based hate. In this study, we consider using all six classes and train/test samples provided in the original competition dataset to train, validate, and evaluate the model. Overall, the dataset contains 159,571 training samples and 153,165 testing samples.

Figure 1 shows the number of comments that belong to each category. A comment can exist in one, two, three, or more categories, as shown in Figure 2. In total, there are 16,225 comments with labels. There are also comments that do not belong to any of the categories, which are 143,346 comments. Such comments were deemed to be non-toxic. The number of non-toxic comments is shown under the none column.

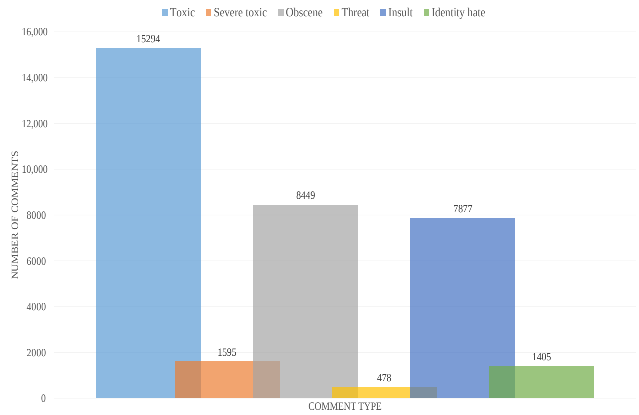


Figure 1. Comments distribution in all categories.

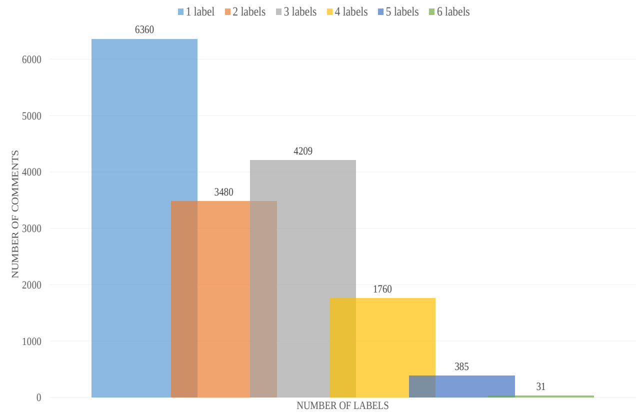


Figure 2. Comments having multiple labels.

Fine-tuning BERT-base on a multi-label text classification problem is done by adding a dropout and classification layer (simple feed-forward layer with standard Sigmoid). The layers are placed on top of the transformer output for the [CLS] token. In the fine-tuning, most hyper-parameters stay the same as in BERT's original training. Hence, a BERT-based compatible tokenization model was used, which contains around 30,000 tokens. During this stage, the classifier and the pre-trained model weights are trained jointly. For fine-tuning, we use a maximum sequence size of 169 tokens, a batch size of 16, and we set the learning rate to 2×10^{-5}. We consider the usage of AUC-ROC (Area Under the Receiver Operating Characteristics Curve) score as an evaluation metric to assess the classification model performance and choose the best fine-tuned model. For instance, the score for each class is calculated based on averaging the corresponding predicted AUCs for each class. The metric is the standard evaluation method used in the Kaggle competition which can be define as in Equation (1).

$$\delta(X,Y) = \begin{cases} 0 & \text{if } x < y \\ 0.5 & \text{if } x = y \\ 1 & \text{if } x > y \end{cases} \qquad AUC = \frac{1}{|X| \cdot |Y|} \sum_{1}^{|X|} \sum_{1}^{|Y|} \delta(x,y) \tag{1}$$

where δ indicates all the possible comparisons between subsets X and Y. Meanwhile, related metric to AUC such as specificity and sensitivity are calculate in Equations (2) and (3).

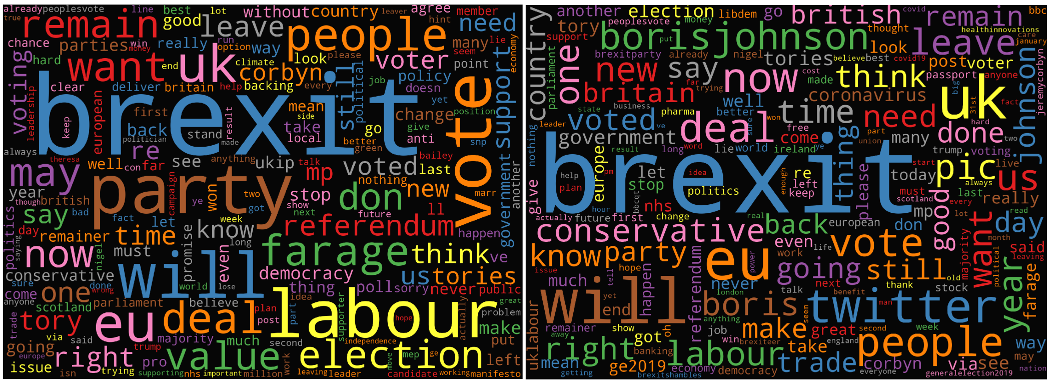
$$Sensitivity = \frac{True\ positive}{True\ positive + False\ negative} \tag{2}$$

$$Specificity = \frac{True\ negative}{True\ negative + False\ positive} \tag{3}$$

where true positive (TP) stands for positive samples correctly predicted, true negative (TN) are negative samples correctly predicted, false positive (TP) are positive samples misclassified as negative samples, and false negative are negative samples misclassified as positive samples.

4. Twitter Data Retrieval and Cleaning

The described methodological framework was applied on a Twitter data stream collected and divided into two time-periods: (i) The first four months of 2019 in which we collected around 14,000 tweets (Dataset1) (ii) and from 1 October 2019 to 31 March 2020 in which we collected around 10,000 tweets (Dataset2). Tweets were collected from Twitter via the Twitter API using several search terms and hashtags such as Brexit, #Brexit, #BrexitBetrayal, and #StopBrexit. Figure 3 shows the most frequent words in the collected datasets.



(**a**) (**b**)

Figure 3. Word clouds for (**a**) Dataset1 and (**b**) Dataset2.

The topic of Brexit has sparked a lot of tension among British citizens, and it is most prevalent on Twitter. Brexit is a term used to convey the process of the United Kingdom (UK), leaving the European Union (EU) [73]. A quick look at the several Brexit-related hashtags on Twitter shows the genuine disagreements people have on the matter. Some user tweets convey their thoughts in a more cordial manner, while others do not. People from all political spectrums have taken to social media sites like Twitter to vent their frustrations and give their opinions, among other things. The conversation around Brexit on Twitter ranges from sensible to chaotic. The current prime minister, Theresa May, proposed several deals that were all rejected by parliament. At such occurrences, there is a considerable uptick in the Brexit conversation on Twitter. The model proposed in the methodology will, therefore, be used to check the toxicity of conversations around Brexit on Twitter.

A look at the raw data will show a lot of unnecessary information within the tweets where only English tweets were included. Unnecessary in this case means bits and pieces of information that add no value to the model classification. Since these tweets were collected from hashtags, they contain hashtags. However, instead of removing the whole hashtag, only the "#" character was removed, leaving the text of the hashtag. Emojis, which are small digital icons used in online text like tweets, are also present in the collected data. As the model was not trained on emoji data, they were removed. Numbers, punctuations, and URL links in the tweets were also removed from the dataset.

5. Experiments and Results

We adopt BERT-base as the basis for all experiments based on the publicly available implementation of BERT (Tensorflow, https://github.com/huggingface/transformers (accessed on 1 February 2021), and we follow the fine-tuning regime specified in [48]). While many submissions to the Kaggle leaderboard tackling Toxic Comment Classification Challenge achieved high results, our fine-tuned model performed very well on the challenge test set where it scored 0.98561 (AUC-ROC) as a public score and 0.98603 (AUC-ROC) as a private score. Note that the maximum sequence length is set to 169 in our experiments where a batch size of 16 is used for fine-tuning, and 32 is used during the testing phase. The following sections present results and analysis for our experiments starting with the fine-tuning stage of the BERT model. Then, the classification of the collected tweets dataset (Dataset1 and Dataset2) and analysis of the model's predictions. Additionally, we compare the BERT-base model to three other models, called Multilingual BERT, RoBERTa, and DistilBERT. Table 1 lists the comparison results. It is clear that the BERT-base model recorded the best private score and public score of 0.9890, and 0.9856, respectively. The Multilingual BERT obtained the second rank, followed by DistilBERT and RoBERTa.

Table 1. Evaluated models for the fine-tuning task.

Model	Private Score	Public Score
BERT base	0.9860	0.9856
Multilingual BERT	0.9845	0.9852
RoBERTa	0.9772	0.9772
DistilBERT	0.9851	0.9848

5.1. Fine-Tuning BERT

Figure 4 shows the change in AUC-ROC value for each category during the fine-tuning (training) of BERT-base on the Toxic Comment Classification Challenge dataset described in Section 3.3. The reported changes over each train steps set shows that the fine-tuned model performed well in learning to classify different comments. In toxic, server toxic, obscene, and insult categories, the model scored less than 0.81 during the first 500 training steps, and then it starts fitting the data from the 1000 training steps till it reaches the highest AUC-ROC values on all categories. Whereas, in threat and identity hate categories, the model starts having AUC-ROC values greater than 0.90 after 2000 and 1000 training steps, respectively. This is due to the small size of training samples presented in threat and identity hate categories.

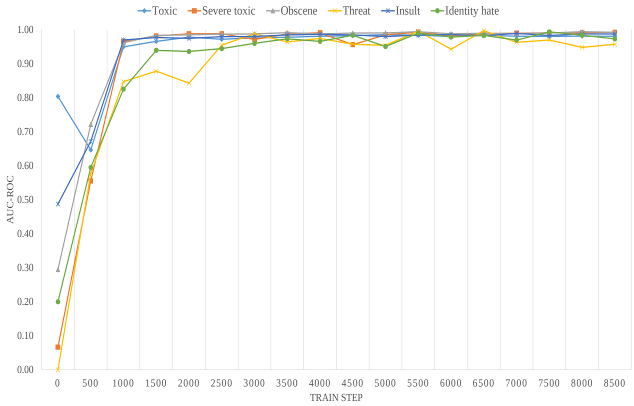


Figure 4. Training AUC-ROC in each training step for each comment type.

Figure 5 shows the training loss during the fine-tuning process, where the model reaches its lowest value (0.0583) after 8500 training steps, taking around 48 min.

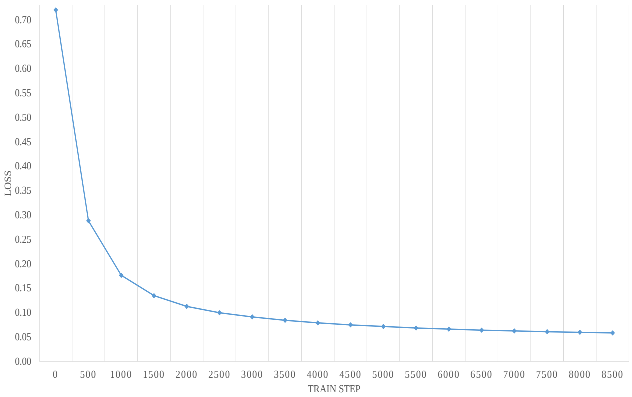


Figure 5. Loss in each training step during fine-tuning.

Figure 6 shows the obtained AUC-ROC values for each category during the fine-tuning (validation) of BERT-base where 10% of the training samples are used as a validation set. As shown in the figure, most of the categories have been classified with an AUC-ROC value greater than 0.98. Where the average AUC-ROC value for all categories is equal to 0.9878 with a loss value of 0.0374. These findings support the results obtained during the evaluation of the fine-tuned model on the testing set (the evaluation of the testing set predictions has been conducted on the Kaggle platform, https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge (accessed on 1 January 2021)).

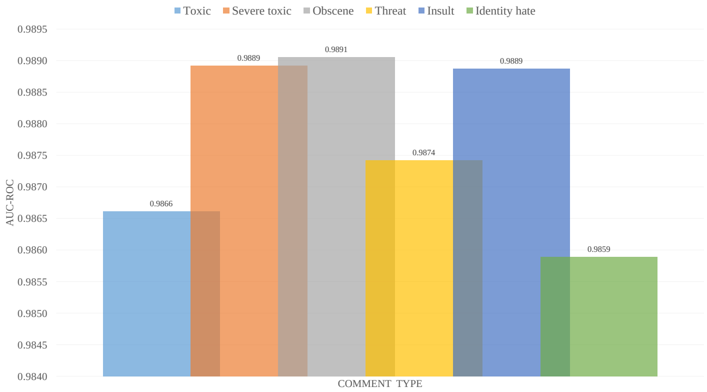


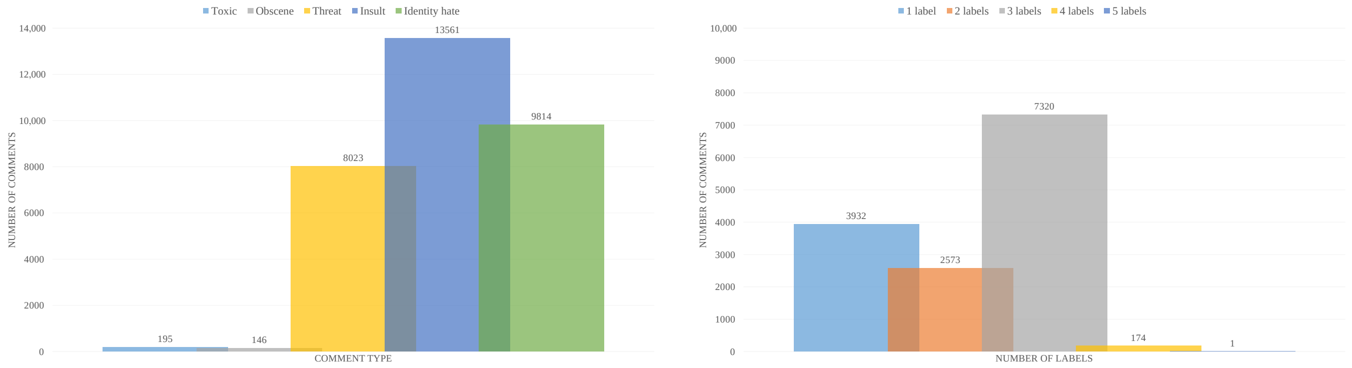
Figure 6. Validation AUC-ROC for each comment type.

5.2. Prediction Results

For Dataset1, 10,000 Tweets from the search term Brexit and hashtag # BrexitBetrayal were collected. After the tweets were cleaned and tokenized, they were fed to the model. Figure 7a shows the results of these predictions. The results show that most of the tweets were classified as an insult. Then, threat class in the second rank, and obscene class in the third rank. Identity hate class is the fourth, followed by toxic class. Figure 7b shows the number of tweets in Dataset1 having more than one label predicted by the model. The figure shows that 7320 tweets have three labels, which support the results in Figure 7, where threat, insult, and identity hate categories dominate the predicted classes.

Figure 8a shows the prediction results of Dataset2 of the BERT-base model. Insult class attracted the most tweets, followed by threat and obscene categories, respectively.

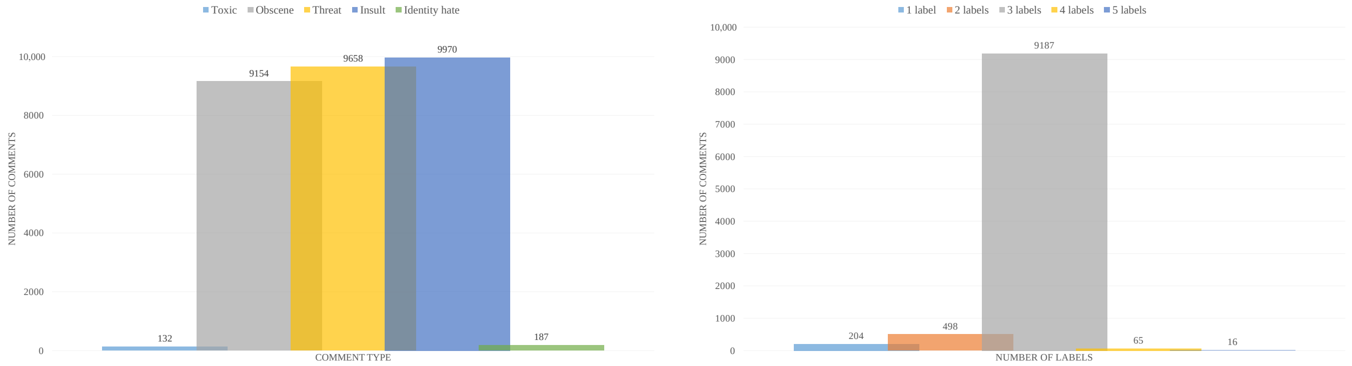
Figure 8b shows the number of tweets in Dataset2 having more than one label predicted by the BERT-base model. The figure shows that 9187 tweets have three labels, which is greater than the results reported in Dataset1 with few tweets having one or more labels.



(**a**) Dataset1 tweets in each class.

(**b**) Dataset1 tweets having multi-labels.

Figure 7. Dataset1 predictions using BERT base.

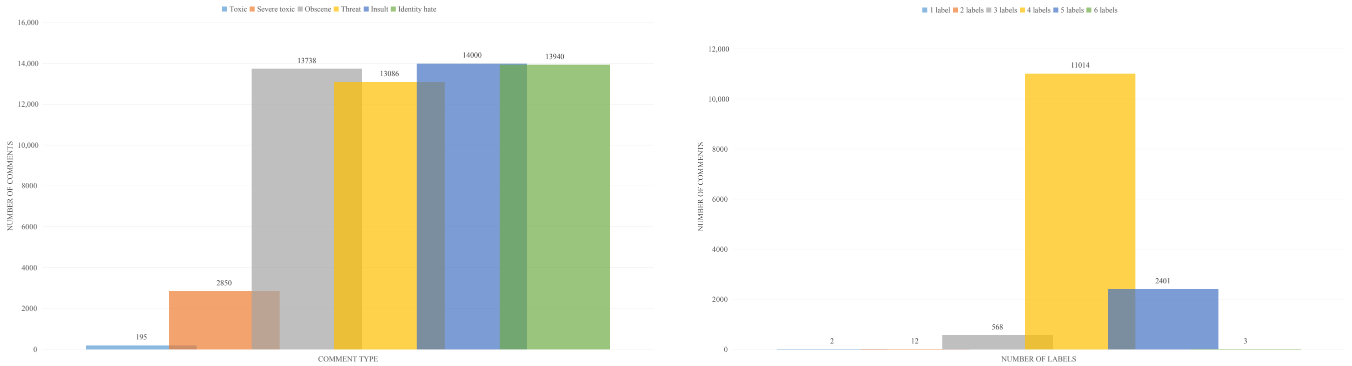


(**a**) Dataset2 tweets in each class.

(**b**) Dataset2 tweets having multi-labels.

Figure 8. Dataset2 predictions using BERT base.

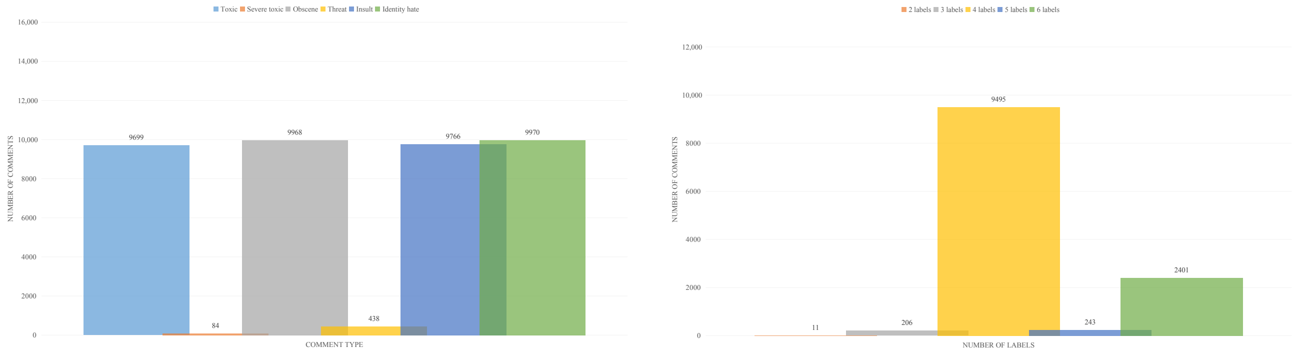
Furthermore, for the Multilingual BERT model using Dataset1, Figure 9a shows that most of the tweets were also classified as an insult. Figure 9b indicates that 11,014 tweets have three labels. For Dataset2, Figure 10a shows that identity hate class came in first. Figure 10a shows that 9495 tweets have three labels.



(**a**) Dataset1 tweets in each class.

(**b**) Dataset1 tweets having multi-labels.

Figure 9. Dataset1 predictions using Multilingual BERT.

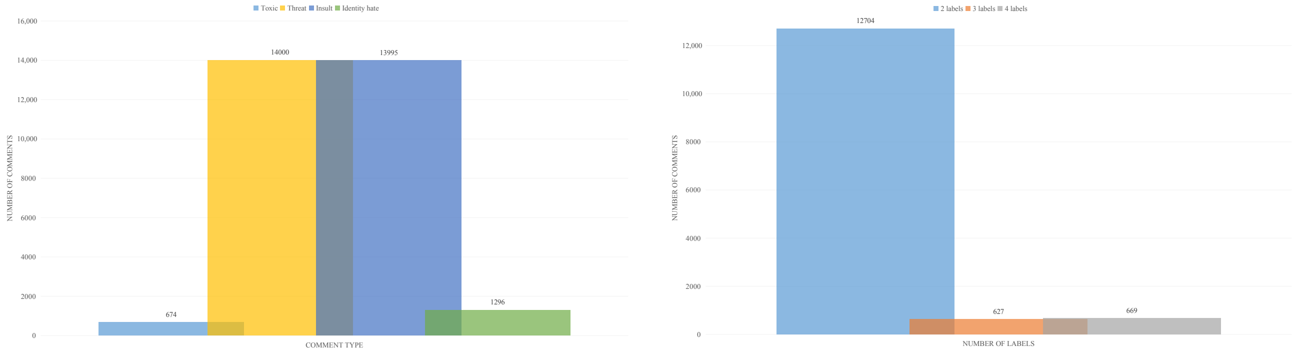


(**a**) Dataset2 tweets in each class.

(**b**) Dataset2 tweets having multi-labels.

Figure 10. Dataset2 predictions using Multilingual BERT.

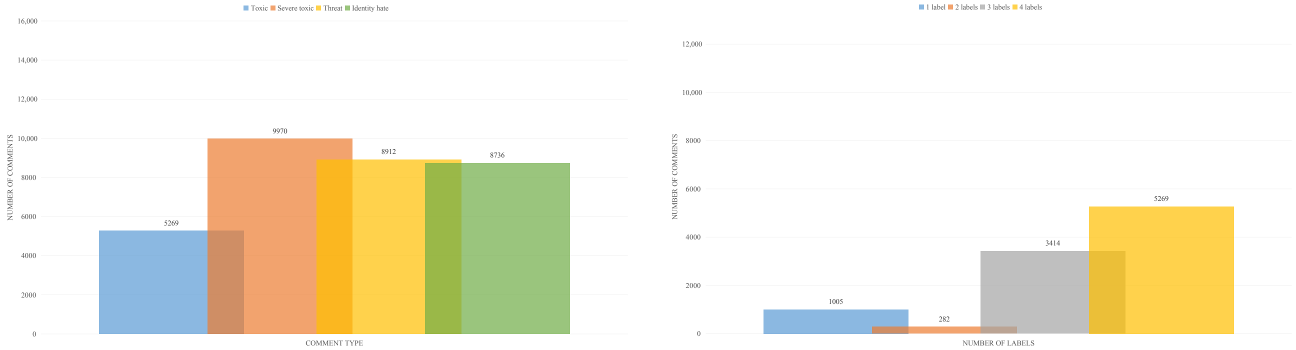
Additionally, for the RoBERTa model, in Dataset1, most of the tweets classified as insult and threat classes had a number of tweets at 13,995 and 14,000, respectively, as shown in Figure 11a. Moreover, Figure 11b indicates that 12,704 tweets have two labels. For Dataset2, as shown in Figure 12a severe toxic class came in first, with 9970 tweets. Figure 12b shows that about 5269 tweets have four labels.



(**a**) Dataset1 tweets in each class.

(**b**) Dataset1 tweets having multi-labels.

Figure 11. Dataset1 predictions using Roberta.

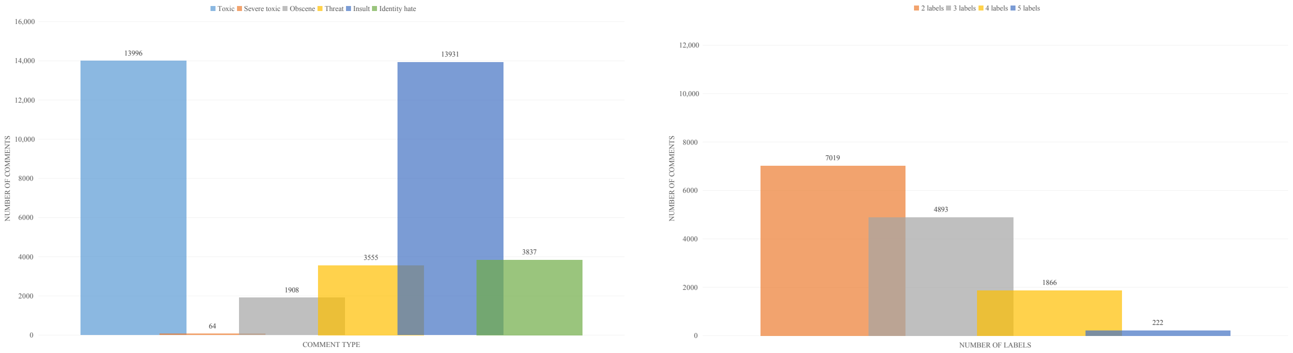


(**a**) Dataset2 tweets in each class.

(**b**) Dataset2 tweets having multi-labels.

Figure 12. Dataset2 predictions using Roberta.

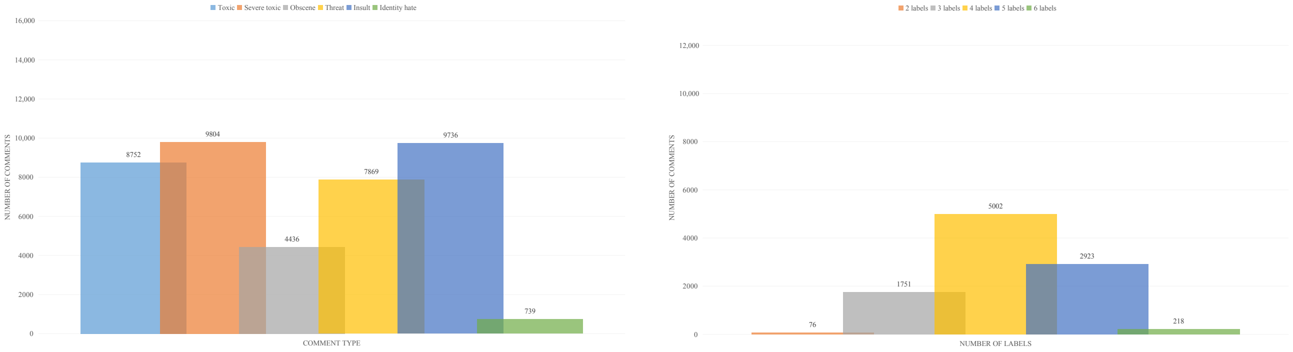
Finally, for the DistilBERT model, Figure 13a shows that most of the tweets in Dataset1 were classified as toxic and insult. Figure 13b indicates that 7019 tweets have two labels. Figure 14a shows that severe toxic, insult, and toxic are the most classified tweets, respectively. Moreover, about 5002 tweets have four labels, as shown in Figure 14b.



(**a**) Dataset1 tweets in each class.

(**b**) Dataset1 tweets having multi-labels.

Figure 13. Dataset1 predictions using DistilBERT.



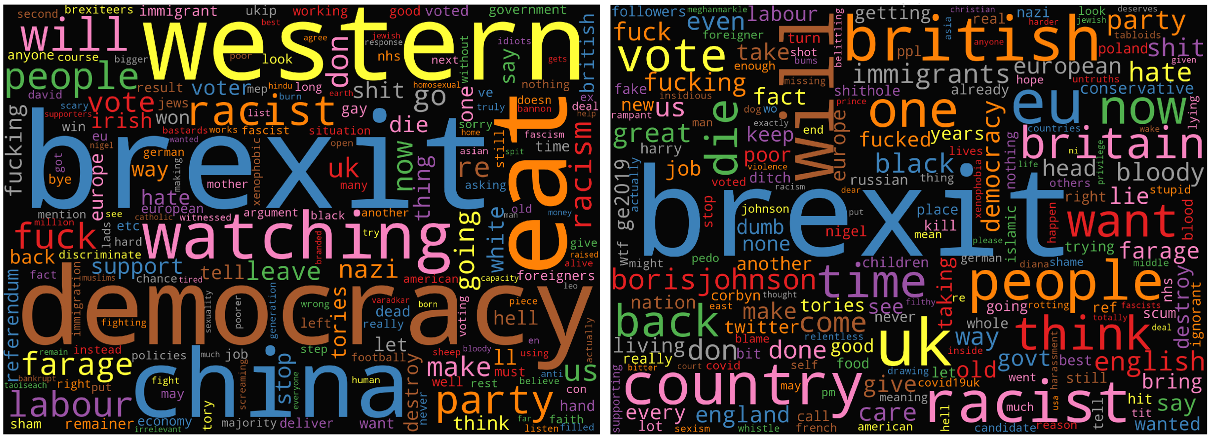
(**a**) Dataset2 tweets in each class.

(**b**) Dataset2 tweets having multi-labels.

Figure 14. Dataset2 predictions using DistilBERT.

5.3. Tweets Analysis

The following figures show the most common words in each class presented as word clouds. Figure 15 shows the most common words from tweets in the toxic class. The word Brexit comes up as the most used word, which is understandable considering the context. Even other words like a party, vote, and the UK appear. Harsh words like idiot, shit, and fuck are also among the most frequently used words in this class. However, the model did not classify any tweet as severe toxic.

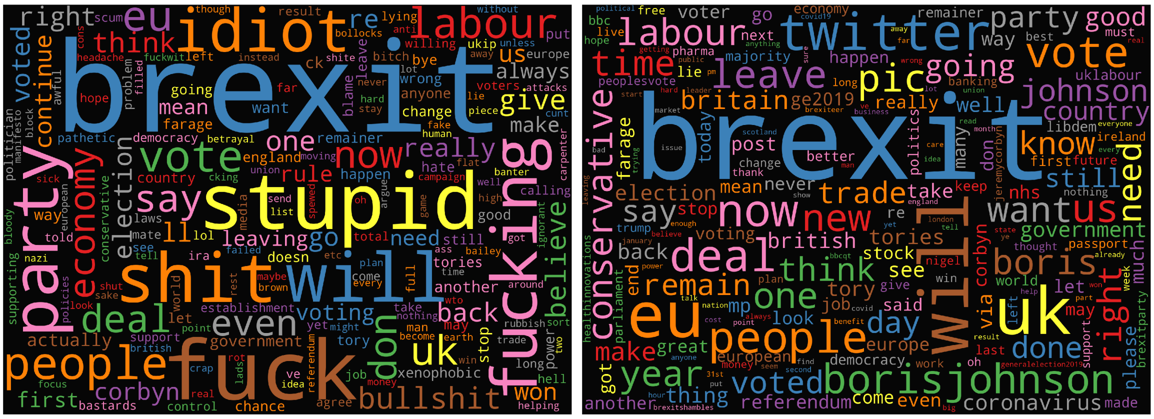


(**a**)

(**b**)

Figure 15. Word clouds for (**a**) toxic category in Dataset1 and (**b**) Dataset2.

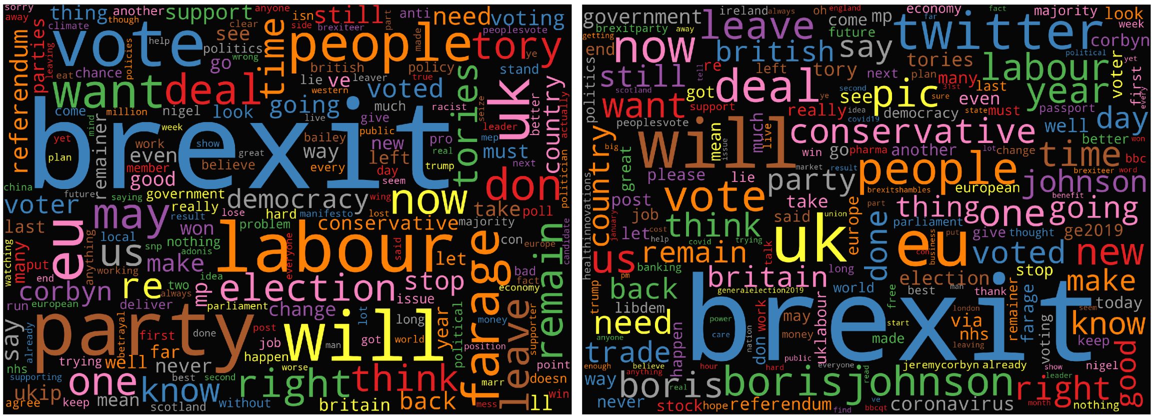
Figure 16 shows the obscene class common words that have Brexit as the most common among tweets in it. It also has harsh words showing up more frequently compared to the other classes.



(**a**) (**b**)

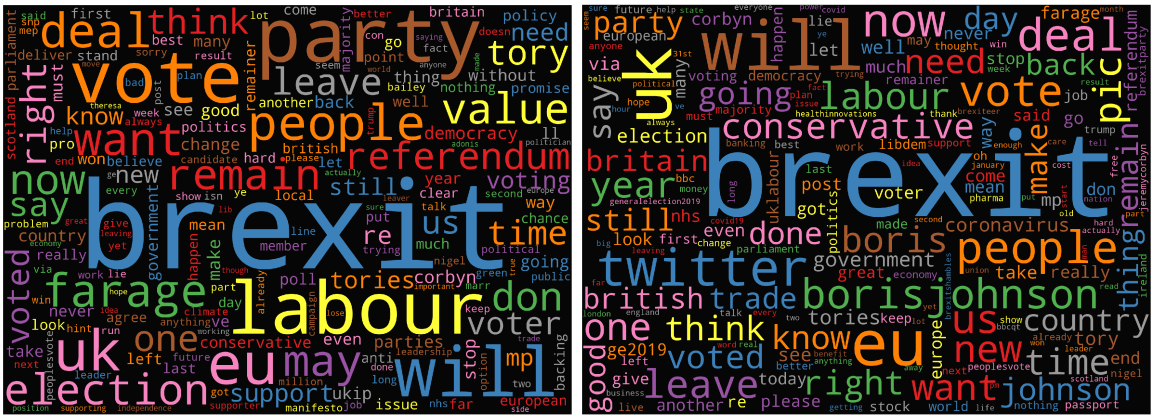
Figure 16. Word clouds for (**a**) obscene category in Dataset1 and (**b**) Dataset2.

In the same manner, Figure 17–19, show the common words in the threat class, insult class, and identity hate class, respectively.



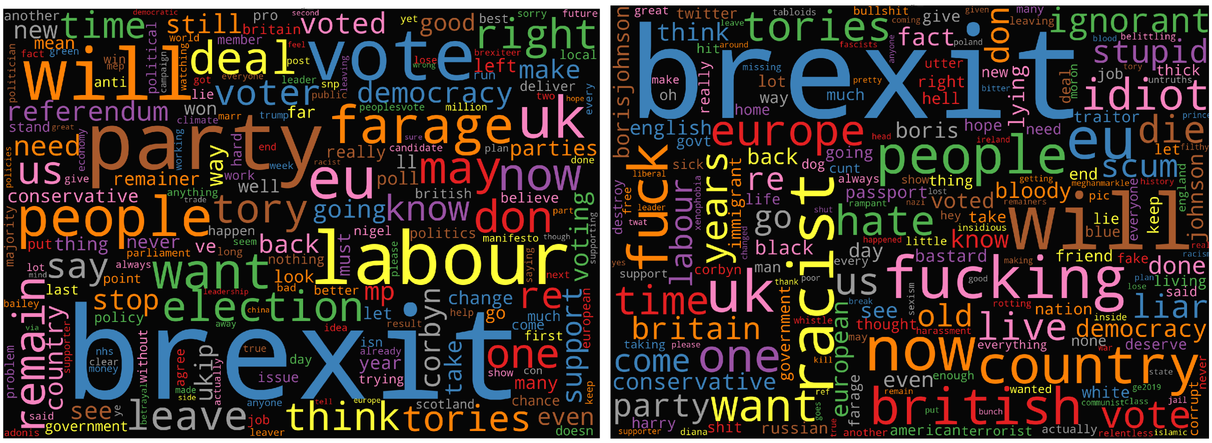
(**a**) (**b**)

Figure 17. Word clouds for (**a**) threat category in Dataset1 and (**b**) Dataset2.



(**a**) (**b**)

Figure 18. Word clouds for (**a**) insult category in Dataset1 and (**b**) Dataset2.



(**a**)Dataset1 (**b**)Dataset2

Figure 19. Word clouds for (**a**) identity hate category in Dataset1 and (**b**) Dataset2.

6. Conclusions

In this study, we addressed toxicity detection in social media using deep learning techniques. We adopt the Bidirectional Encoder Representations from Transformers (BERT) to classify toxic comments from user-generated data in social media, such as tweets. The BERT-base pre-trained model was fine-tuned on a well-known labeled toxic comment dataset, Kaggle public datasets. Moreover, the proposed model was tested on real-world data, two different tweets datasets, collected in two different periods based on a case study of the UK Brexit. The evaluation outcomes showed that BERT has the ability to classify and to predict toxic comments with a high accuracy rate. Moreover, we compared the BERT-base model to three models, called Multilingual BERT, RoBERTa, and DistilBERT. The BERT-base model outperformed all compared models and achieved the best results.

In future work, further work could be done to make the model better suited to dealing with specific social media data. The size of the dataset could be increased to include tweets to train the model with more Twitter-related data. These tweets would have to be hand-labeled, which would take a fair amount of time to get enough data to increase the accuracy of the model. One of the benefits of adding tweets is that tweets that have emojis can be kept in the dataset. This would allow the model to be trained to account for the presence of emojis. Data labeled toxic based on tweets is currently not available. However, the ability to label a massive dataset of this type of data could be hugely beneficial in the long run. Aside from just collecting Twitter data, text from other social media sites like Facebook, YouTube, and Reddit could be added to improve the dataset.

Author Contributions: Conceptualization, H.F. and M.A.A.A.-q.; data curation, A.A.E., A.D., D.Y., and M.A.E.; formal analysis, A.D. and M.A.A.A.-q.; funding acquisition, H.F.; investigation, H.F. and W.D.; Methodology, H.F. and A.D.; project administration, M.A.A.A.-q.; resources, A.A.E., A.D., L.A., and M.A.A.A.-q.; Software, A.D. and M.A.A.A.-q.; supervision, H.F.; writing—original draft preparation, H.F., A.D. and M.A.A.A.-q.; writing—review and editing, A.A.E., M.A.E., D.Y., A.H.E., and L.A.; Validation, M.A.E., D.Y., A.H.E., and L.A.; All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Key Research and Development Program of China (Grant No. 2019YFB1405600).

Data Availability Statement: Available by corresponding author upon request.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Abualigah, L.; Gandomi, A.H.; Elaziz, M.A.; Hussien, A.G.; Khasawneh, A.M.; Alshinwan, M.; Houssein, E.H. Nature-Inspired Optimization Algorithms for Text Document Clustering—A Comprehensive Analysis. *Algorithms* **2020**, *13*, 345. [CrossRef]
2. Uhls, Y.T.; Ellison, N.B.; Subrahmanyam, K. Benefits and costs of social media in adolescence. *Pediatrics* **2017**, *140*, S67–S70. [CrossRef] [PubMed]
3. Souri, A.; Hosseinpour, S.; Rahmani, A.M. Personality classification based on profiles of social networks' users and the five-factor model of personality. *Hum. Centric Comput. Inf. Sci.* **2018**, *8*, 24. [CrossRef]
4. Morente-Molinera, J.A.; Kou, G.; Samuylov, K.; Ureña, R.; Herrera-Viedma, E. Carrying out consensual Group Decision Making processes under social networks using sentiment analysis over comparative expressions. *Knowl. Based Syst.* **2019**, *165*, 335–345. [CrossRef]
5. Risch, J.; Krestel, R. Aggression identification using deep learning and data augmentation. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Santa Fe, NM, USA, 25 August 2018; pp. 150–158.
6. Subramani, S.; Wang, H.; Vu, H.Q.; Li, G. Domestic violence crisis identification from facebook posts based on deep learning. *IEEE Access* **2018**, *6*, 54075–54085. [CrossRef]
7. Subramani, S.; Michalska, S.; Wang, H.; Du, J.; Zhang, Y.; Shakeel, H. Deep Learning for Multi-Class Identification From Domestic Violence Online Posts. *IEEE Access* **2019**, *7*, 46210–46224. [CrossRef]
8. Abualigah, L.; Gandomi, A.H.; Elaziz, M.A.; Hamad, H.A.; Omari, M.; Alshinwan, M.; Khasawneh, A.M. Advances in Meta-Heuristic Optimization Algorithms in Big Data Text Clustering. *Electronics* **2021**, *10*, 101. [CrossRef]
9. Abualigah, L.M.Q. *Feature Selection And Enhanced Krill Herd Algorithm For Text Document Clustering*; Springer: Berlin/Heidelberg, Germany, 2019.
10. Ahmad, S.; Asghar, M.Z.; Alotaibi, F.M.; Awan, I. Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Hum. Centric Comput. Inf. Sci.* **2019**, *9*, 24. [CrossRef]
11. Budiharto, W.; Meiliana, M. Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. *J. Big Data* **2018**, *5*, 51. [CrossRef]
12. Prabhu, B.A.; Ashwini, B.; Khan, T.A.; Das, A. Predicting Election Result with Sentimental Analysis Using Twitter Data for Candidate Selection. In *Innovations in Computer Science and Engineering*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 49–55.
13. Cury, R.M. Oscillation of tweet sentiments in the election of João Doria Jr. for Mayor. *J. Big Data* **2019**, *6*, 42. [CrossRef]
14. Al Shehhi, A.; Thomas, J.; Welsch, R.; Grey, I.; Aung, Z. Arabia Felix 2.0: a cross-linguistic Twitter analysis of happiness patterns in the United Arab Emirates. *J. Big Data* **2019**, *6*, 33. [CrossRef]
15. Pong-inwong, C.; Songpan, W. Sentiment analysis in teaching evaluations using sentiment phrase pattern matching (SPPM) based on association mining. *Int. J. Mach. Learn. Cybern.* **2018**, *10*, 2177–2186. [CrossRef]
16. Aloufi, S.; El Saddik, A. Sentiment identification in football-specific tweets. *IEEE Access* **2018**, *6*, 78609–78621. [CrossRef]
17. Amato, F.; Castiglione, A.; Moscato, V.; Picariello, A.; Sperlì, G. Multimedia summarization using social media content. *Multimed. Tools Appl.* **2018**, *77*, 17803–17827. [CrossRef]
18. Amato, F.; Moscato, V.; Picariello, A.; Sperlí, G. Multimedia social network modeling: A proposal. In Proceedings of the 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 4–6 February 2016; pp. 448–453.
19. Li, Z.; Fan, Y.; Jiang, B.; Lei, T.; Liu, W. A survey on sentiment analysis and opinion mining for social multimedia. *Multimed. Tools Appl.* **2019**, *78*, 6939–6967. [CrossRef]
20. Angadi, S.; Reddy, R.V.S. Survey on Sentiment Analysis from Affective Multimodal Content. In *Smart Intelligent Computing and Applications*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 599–607.
21. Chiranjeevi, P.; Santosh, D.T.; Vishnuvardhan, B. Survey on Sentiment Analysis Methods for Reputation Evaluation. In *Cognitive Informatics and Soft Computing*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 53–66.
22. Alaei, A.R.; Becken, S.; Stantic, B. Sentiment analysis in tourism: capitalizing on big data. *J. Travel Res.* **2019**, *58*, 175–191. [CrossRef]
23. Kwak, H.; Blackburn, J.; Han, S. Exploring cyberbullying and other toxic behavior in team competition online games. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Korea, 18–23 April 2015; pp. 3739–3748.
24. O'Keeffe, G.S.; Clarke-Pearson, K. The impact of social media on children, adolescents, and families. *Pediatrics* **2011**, *127*, 800–804. [CrossRef] [PubMed]
25. Whittaker, E.; Kowalski, R.M. Cyberbullying via social media. *J. Sch. Violence* **2015**, *14*, 11–29. [CrossRef]
26. Fox, J.; Cruz, C.; Lee, J.Y. Perpetuating online sexism offline: Anonymity, interactivity, and the effects of sexist hashtags on social media. *Comput. Hum. Behav.* **2015**, *52*, 436–442. [CrossRef]
27. Lapidot-Lefler, N.; Barak, A. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Comput. Hum. Behav.* **2012**, *28*, 434–443. [CrossRef]
28. Kim, H.; Chang, Y. Managing Online Toxic Disinhibition: The Impact of Identity and Social Presence. SIGHCI 2017 Proceedings. Available online: https://aisel.aisnet.org/sighci2017/1 (accessed on 1 February 2021).
29. Joyce, B.; Deng, J. Sentiment analysis of tweets for the 2016 US presidential election. In Proceedings of the 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 3–5 November 2017; pp. 1–4.

30. You, Q.; Luo, J.; Jin, H.; Yang, J. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
31. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 12–15 December 2016; pp. 439–448.
32. Li, X.; Xie, H.; Chen, L.; Wang, J.; Deng, X. News impact on stock price return via sentiment analysis. *Knowl. Based Syst.* **2014**, *69*, 14–23. [CrossRef]
33. Wöllmer, M.; Weninger, F.; Knaup, T.; Schuller, B.; Sun, C.; Sagae, K.; Morency, L.P. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intell. Syst.* **2013**, *28*, 46–53. [CrossRef]
34. Arias, M.; Arratia, A.; Xuriguera, R. Forecasting with twitter data. *ACM Trans. Intell. Syst. Technol. (TIST)* **2013**, *5*, 8. [CrossRef]
35. Jansen, B.J.; Zhang, M.; Sobel, K.; Chowdury, A. Twitter power: Tweets as electronic word of mouth. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 2169–2188. [CrossRef]
36. Abualigah, L.; Diabat, A.; Mirjalili, S.; Abd Elaziz, M.; Gandomi, A.H. The arithmetic optimization algorithm. *Comput. Methods Appl. Mech. Eng.* **2021**, *376*, 113609. [CrossRef]
37. Abualigah, L.; Yousri, D.; Abd Elaziz, M.; Ewees, A.A.; Al-qaness, M.; Gandomi, A.H. Aquila Optimizer: A novel meta-heuristic optimization Algorithm. *Comput. Ind. Eng.* **2021**, *107250*.
38. Ringsquandl, M.; Petkovic, D. Analyzing political sentiment on Twitter. In Proceedings of the 2013 AAAI Spring Symposium Series, Stanford, CA, USA, 25–27 March 2013.
39. Kušen, E.; Strembeck, M. Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. *Online Soc. Netw. Media* **2018**, *5*, 37–50. [CrossRef]
40. Haselmayer, M.; Jenny, M. Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Qual. Quant.* **2017**, *51*, 2623–2646. [CrossRef]
41. Rathan, M.; Hulipalled, V.R.; Venugopal, K.; Patnaik, L. Consumer insight mining: Aspect based Twitter opinion mining of mobile phone reviews. *Appl. Soft Comput.* **2018**, *68*, 765–773.
42. Anastasia, S.; Budi, I. Twitter sentiment analysis of online transportation service providers. In Proceedings of the 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Malang, Indonesia, 15–16 October 2016; pp. 359–365.
43. Pagolu, V.S.; Reddy, K.N.; Panda, G.; Majhi, B. Sentiment analysis of Twitter data for predicting stock market movements. In Proceedings of the 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, India, 3–5 October 2016; pp. 1345–1350.
44. Alomari, E.; Mehmood, R. Analysis of tweets in Arabic language for detection of road traffic conditions. In *International Conference on Smart Cities, Infrastructure, Technologies and Applications*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 98–110.
45. Al-qaness, M.A.; Abd Elaziz, M.; Hawbani, A.; Abbasi, A.A.; Zhao, L.; Kim, S. Real-Time Traffic Congestion Analysis Based on Collected Tweets. In Proceedings of the 2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS), Shenyang, China, 21–23 October 2019; pp. 1–8.
46. Frank, M.R.; Mitchell, L.; Dodds, P.S.; Danforth, C.M. Happiness and the patterns of life: A study of geolocated tweets. *Sci. Rep.* **2013**, *3*, 2625. [CrossRef]
47. Giachanou, A.; Crestani, F. Like it or not: A survey of twitter sentiment analysis methods. *ACM Comput. Surv. (CSUR)* **2016**, *49*, 1–41. [CrossRef]
48. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
49. Mai, I.; Marwan, T.; Nagwa, E.M. Imbalanced Toxic Comments Classification Using Data Augmentation and Deep Learning. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 875–878. [CrossRef]
50. Wulczyn, E.; Thain, N.; Dixon, L. Ex machina: Personal attacks seen at scale. In Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, Perth, Australia, 3–7 May 2017; pp. 1391–1399.
51. Saeed, H.H.; Shahzad, K.; Kamiran, F. Overlapping Toxic Sentiment Classification Using Deep Neural Architectures. In Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW), Singapore, 17–20 November 2018; pp. 1361–1366.
52. Georgakopoulos, S.V.; Tasoulis, S.K.; Vrahatis, A.G.; Plagianakos, V.P. Convolutional neural networks for toxic comment classification. In Proceedings of the 10th Hellenic Conference on Artificial Intelligence, Patras, Greece, 9–12 July 2018; p. 35.
53. Fang, W.; Luo, H.; Xu, S.; Love, P.E.; Lu, Z.; Ye, C. Automated text classification of near-misses from safety reports: An improved deep learning approach. *Adv. Eng. Inform.* **2020**, *44*, 101060. [CrossRef]
54. Fan, B.; Fan, W.; Smith, C.; Garner, H. Adverse drug event detection and extraction from open data: A deep learning approach. *Inf. Process. Manag.* **2020**, *57*, 102131. [CrossRef]
55. Moradi, M.; Dorffner, G.; Samwald, M. Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. *Comput. Methods Programs Biomed.* **2020**, *184*, 105117. [CrossRef]

56. Wang, Q.; Ji, Z.; Wang, J.; Wu, S.; Lin, W.; Li, W.; Ke, L.; Xiao, G.; Jiang, Q.; Xu, H.; et al. A study of entity-linking methods for normalizing Chinese diagnosis and procedure terms to ICD codes. *J. Biomed. Inform.* **2020**, *105*, 103418. [CrossRef]
57. Koroleva, A.; Kamath, S.; Paroubek, P. Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. *J. Biomed. Inform.* **2019**, *4*, 100058. [CrossRef]
58. Zhang, X.; Zhang, Y.; Zhang, Q.; Ren, Y.; Qiu, T.; Ma, J.; Sun, Q. Extracting comprehensive clinical information for breast cancer using deep learning methods. *Int. J. Med Inform.* **2019**, *132*, 103985. [CrossRef]
59. Chen, F.; Yuan, Z.; Huang, Y. Multi-source data fusion for aspect-level sentiment classification. *Knowl. Based Syst.* **2020**, *187*, 104831. [CrossRef]
60. Gao, Z.; Feng, A.; Song, X.; Wu, X. Target-Dependent Sentiment Classification With BERT. *IEEE Access* **2019**, *7*, 154290–154299. [CrossRef]
61. Yin, X.; Huang, Y.; Zhou, B.; Li, A.; Lan, L.; Jia, Y. Deep Entity Linking via Eliminating Semantic Ambiguity With BERT. *IEEE Access* **2019**, *7*, 169434–169445. [CrossRef]
62. He, J.; Zhao, L.; Yang, H.; Zhang, M.; Li, W. HSI-BERT: Hyperspectral Image Classification Using the Bidirectional Encoder Representation From Transformers. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 165–178. [CrossRef]
63. Lee, L.H.; Lu, Y.; Chen, P.H.; Lee, P.L.; Shyu, K.K. NCUEE at MEDIQA 2019: Medical text inference using ensemble BERT-BiLSTM-Attention model. In Proceedings of the 18th BioNLP Workshop and Shared Task, Wurzburg, Germany, 16–20 September 2019; pp. 528–532.
64. Liu, J.; Ng, Y.C.; Wood, K.L.; Lim, K.H. Ipod: An industrial and professional occupations dataset and its applications to occupational data mining and analysis. *arXiv* **2019**, arXiv:1910.10495.
65. Zhang, Z.; Zhang, Z.; Chen, H.; Zhang, Z. A Joint Learning Framework With BERT for Spoken Language Understanding. *IEEE Access* **2019**, *7*, 168849–168858. [CrossRef]
66. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165.
67. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv* **2019**, arXiv:1906.08237.
68. Shoeybi, M.; Patwary, M.; Puri, R.; LeGresley, P.; Casper, J.; Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv* **2019**, arXiv:1909.08053.
69. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv* **2020**, arXiv:2003.10555.
70. Rogers, A.; Kovaleva, O.; Rumshisky, A. A Primer in BERTology: What we know about how BERT works. *arXiv* **2020**, arXiv:2002.12327.
71. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
72. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
73. Chung, S.W.; Kim, Y. The Truth behind the Brexit Vote: Clearing away Illusion after Two Years of Confusion. *Sustainability* **2019**, *11*, 5201. [CrossRef]