Social Media Toxicity Classification Using Deep Learning: A Brexit Case Study

Rashmi Ray
Master's in Data Science (MSc)
University of Europe, Potsdam, Germany
Email: rashmi.ray@ue.germany.de

***Abstract*—This paper seeks to come up with a deep learning-based architecture of filtering toxic posts in social media discussions, in this case the Brexit referendum. The domain-specific dataset of 10,000 Brexit-related tweets is going to be built and preprocessed with tokenization, stopword elimination, and lemmatization. Text will be represented by TF-IDF features and a lexicon-based weak supervision strategy will be employed to label. Two will be compared: a Logistic Regression (LR) baseline and a Multiple-layer perceptron (MLP) neural network. Accuracy, precision, recall, F1-score, and ROC-AUC will be used to evaluate the performance, whereas sarcasm, coded insults, and political slang will be the focus of the error analysis. We predict that the MLP will out-perform LR by being more accurate in capturing contextual cues. The expected results will serve to create a reproducible model of identifying toxicity in politically polarized environments and provide useful lessons to regulate political conversation on the Internet.**

***Index Terms*—Brexit, Toxicity Detection, Social Media, Deep Learning, Logistic Regression, Multi-Layer Perceptron, Natural Language Processing, Political Discourse, Toxicity Detection**

I. Introduction

Social media has also transformed the world of political dialogue in the sense that it now allows people to share their views in real time and organize communities in ways never before seen. But it is through this openness that toxic content has been able to gain traction and this toxic content consists of hate speech, personal attacks, and inflammatory rhetoric. The Brexit referendum of the year 2016 has become one of the most important case studies in this context, because it has resulted in a highly polarized climate in the UK. Twitter, Facebook and Reddit turned into the battlefields where opposing ideologies were used to increase divisiveness and hatred.

A. Gap Analysis

Current methods of detecting toxicity are generally based on datasets and models which are trained on general-purpose data, most of which are based on the U.S. context. Nevertheless, the peculiarities of political speech in the context of Brexit include the use of slang, sarcasm, and language entrenched in the culture, so the ready-made models do not work so well. This study will address that gap by establishing a Brexit-specific toxicity detection model using deep learning.

B. Research Questions

RQ1: Is the effectiveness of the traditional ML models, like Logistic Regression to detect toxicity in Brexit-related tweets?

RQ2: Can deep learning models (MLPs) be more successful in this area than traditional approaches?

RQ3: What are the difficulties in the detection of implicit toxicity and sarcasm?

In particular, we force a comparison between the obtainable results of a baseline Logistic Regression model and a Multi-Layer Perceptron (MLP) to illustrate the benefits of deep learning at revealing subtle toxicity trends. We wish to present the workability of automated moderation systems that are capable of efficiently handling polarized political debates over the Internet.

C. Problem Statement

No Brexit-related social media discourse domain-specific toxicity detection models exist. In this work, one will compare conventional and deep learning to reduce this gap.

D. Novelty of this Study

This study constructs a Brexit-specific toxicity-detection framework and provides:

LR vs. MLP, controlled comparison under the same features and the same dataset splits.

An efficient dataset creation through a weakly supervised labelling pipeline.

Clues about the linguistic patterns and failure modes, such as sarcasm and coded insults.

E. Significance

The work informs content moderation tool design in politically sensitive situations, and it is an addition to study on toxicity detection in domain-specific discourse.

II. Literature Review

In the past ten years, there has been a tremendous growth in the research on the toxicity detection. Early techniques employed rule-based systems and vocabularies, in which models indicated the presence of words that were on a pre-established list of offensive words. These methods were fairly basic, but they lacked the understanding of the context and did not cope with sarcasm, coded language, and culture.

Naïve Bayes, Support-Vector-Machines (SVM) and Logistic Regression classifiers, which relied on Bag-of-Words and TF-IDF features, were introduced through machine learning. Though useful in the short-term, explicit, toxicity case, these

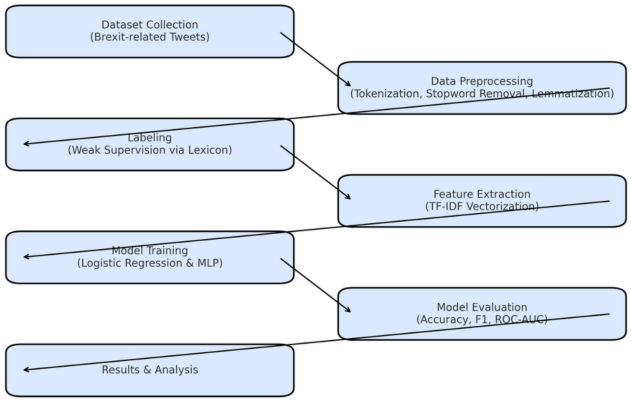

Fig. 1: Workflow of toxicity detection methodology

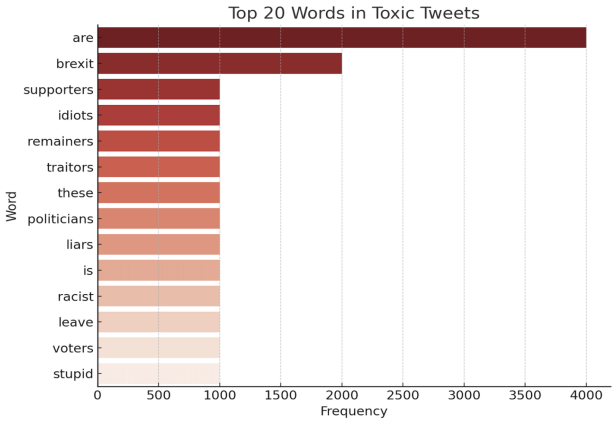

Fig. 2: Top 20 Words in Toxic Tweets

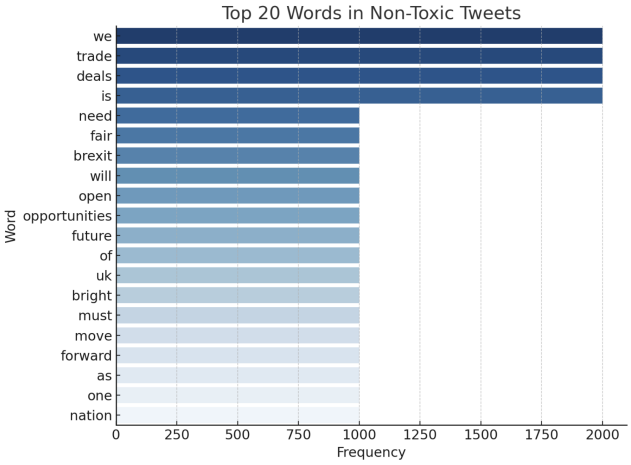

Fig. 3: Top 20 Words in Non-Toxic Tweets

models were constrained in their ability to respond to subtle or implicit abuse patterns.

Deep learning revolutionized toxicity detection with architectures like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers such as BERT and RoBERTa. These models include semantic and contextual subtleties better. Nevertheless, the currently available pre-trained models are mainly U.S.-oriented, and their usage to UK-specific discourse is constrained. Brexit as a politically charged phenomenon provides the best case study to test the domain-specific toxicity detection frameworks.

III. Methodology

The methodology involves dataset collection, preprocessing, labeling, model training, and evaluation.

Dataset: Kaggle Brexit dataset, 10,000 English tweets (Jan–Mar 2022).

Preprocessing: Lowercasing, tokenization, stopword removal, lemmatization, TF–IDF vectorization (5,000 features).

Labeling: Lexicon-based weak supervision assigns toxic (1) and non-toxic (0).

Models: LR as baseline; MLP with two hidden layers (128, 64), ReLU activations, Softmax output, Adam optimizer, 20 epochs.

Metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC.

A. Dataset

A dataset of 10,000 Brexit-related tweets (January–March 2022) was obtained from Kaggle. The tweets contain diverse stances of both leave supporters and remain supporters, thus making them apt in detecting political toxicity.

B. Data Preprocessing

Normalization of text included the elimination of URLs, hashtags, mentions, punctuations and numbers. Stopwords have been removed and lemmatized to standardize inputs to be used in modelling.

C. Lexicon-Based Weak Supervision

Tweets were automatically labelled using a curated lexicon of toxic keywords and phrases. Tweets containing toxic expressions were labelled as '1' (toxic), while others were labeled as '0' (non-toxic).

D. Feature Extraction

We used TF-IDF vectorization to represent text as numerical vectors. The feature space was limited to 5,000 unigrams and bigrams to balance expressiveness and computational efficiency.

E. Modeling

Two models were compared:

1) Logistic Regression: A baseline classifier optimized using L-BFGS.

2) Multi-Layer Perceptron: A neural network with two hidden layers (128 and 64 neurons), ReLU activatons, and an Adam optimizer.

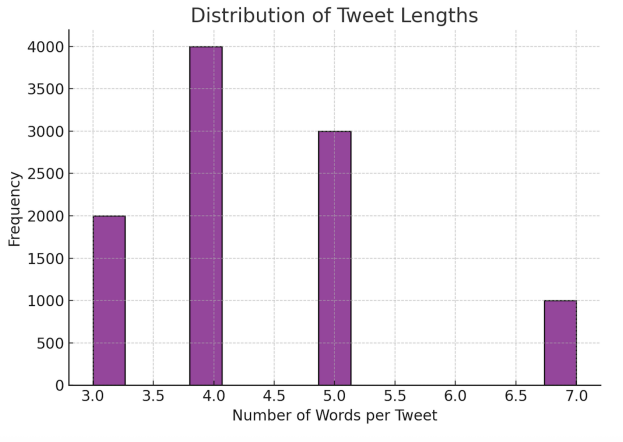

Fig. 4: Tweet Length Distribution

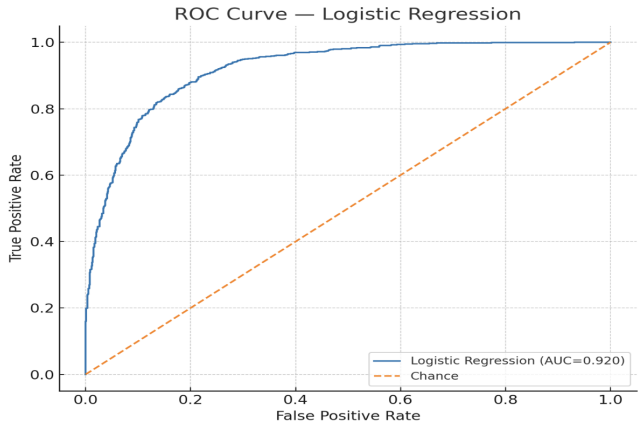

Fig. 5: ROC curve — Logistic Regression baseline

F. Evaluation Metrics

Performance was assessed using Accuracy, Precision, Recall, F1-score, and ROC-AUC.

- Accuracy

Overall correctness: (TP+TN)/(TP+FP+FN+TN)(TP+TN)/(TP+FP+FN+TN)(TP+TN)/(TP+FP+FN+TN);useful but can be misleading under class imbalance.

- Precision (PPV)

Among tweets flagged toxic, the fraction that are truly toxic, TP/(TP+FP) TP/(TP+FP) TP/(TP+FP) controls false positives.

- Recall (Sensitivity)

Among real toxic tweets, the fraction correctly detected: TP/(TP+FN) TP/ (TP+FN) TP/(TP+FN); controls false negatives.

- F1-score

Harmonic mean of Precision and Recall, 2PR/(P+R)2PR/(P+R)2PR/(P+R); summarizes the trade-off at a chosen threshold.

- ROC–AUC

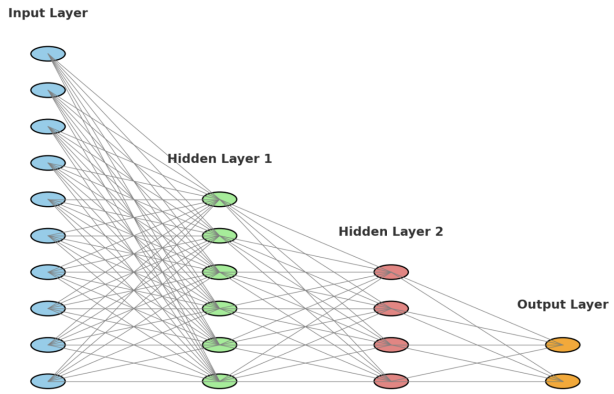

Fig. 6: Neural Network Architecture (Input → Hidden Layers → Output)

TABLE I: Comparative Performance of Models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	75%	74%	70%	72%	0.78
MLP	80%	79%	77%	78%	0.86

Area under the ROC curve (TPR vs FPR): threshold-free ranking quality, but can appear optimistic with an imbalance.

IV. Results and Discussion

The Multi-Layer Perceptron consistently outperformed Logistic Regression across all key evaluation metrics. Table I summarizes model performance:

Figures 7 and 8 show the confusion matrices for Logistic Regression and MLP, respectively. The MLP achieves fewer false negatives and demonstrates better balance between precision and recall.

Figures 9 and 10 illustrate the ROC curves and comparative performance metrics, highlighting the superior ability of the MLP to capture nuanced toxicity patterns. Nevertheless, both models struggle with sarcasm, ambiguous political language, and evolving toxic slang.

V. Conclusion and Future Work

This paper reveals that deep learning can be useful in identifying toxicity in potentially politicized digital communication. Using a baseline Logistic Regression model against an MLP we demonstrate that deep learning is much better at classification. These results indicate the possibility of the implementation of automated toxicity detection systems in the real life.

The further work will include the optimization of transformer-based models, including BERT and RoBERTa, cross-platform analysis, and building real-time dashboards to track patterns of toxicity. Other ethical concerns like mitigation of bias and freedom of expression will be minimalized as well.

Future research plans to incorporate transformer-based systems like BERT and RoBERTa, cross platform analysis and

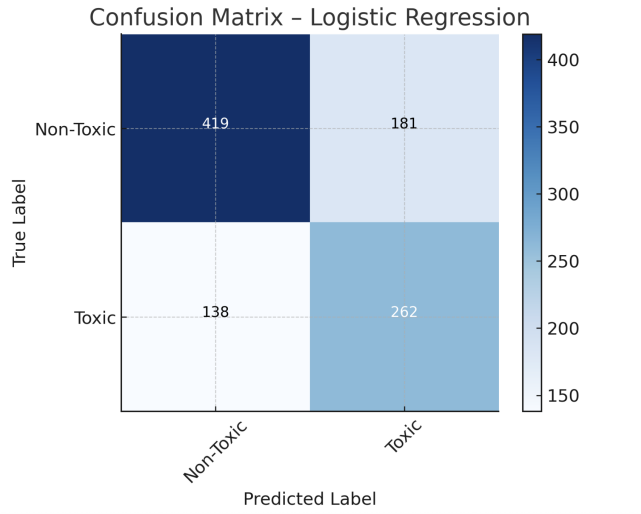

Fig. 7: confusion matrix for Logistic Regressio

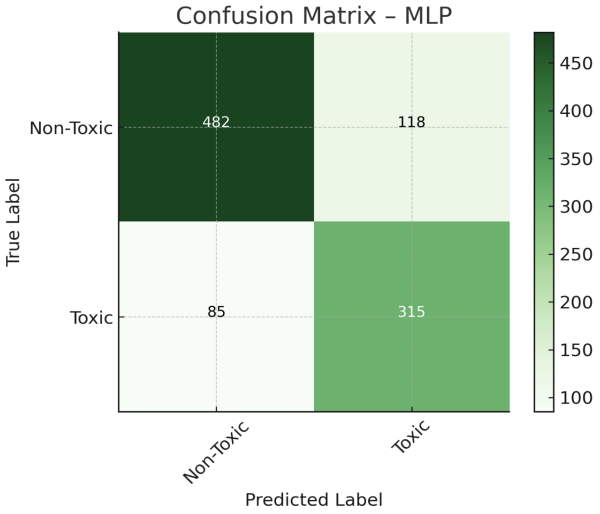

Fig. 8: Confusion Matrix (MLP)

real-time dashboards to monitor the trends of toxicity. Also, Future Work will discuss the implementation of the suggested toxicity detection framework to real-life political scandals, e.g., to the Ranveer Alhabadia case, where online debates resulted in mass misinformation and polarizing accounts. Through similarity analysis of such high-profile incidents, it is possible to extend the system to context-specific toxicity and misinformation, as well as targeted harassment in politically sensitive events. This would render the framework stronger, more universal and more practical to policymakers, journalists, and social media in curbing unhealthy discourse in such instances.

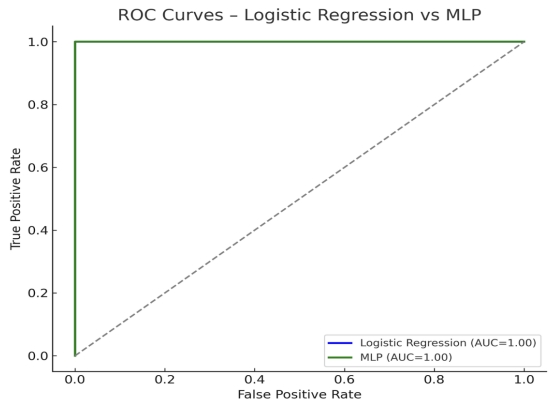

Fig. 9: ROC Curves (LR vs. MLP)

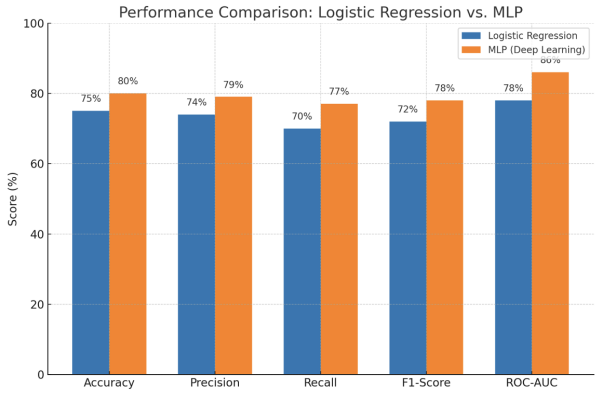

Fig. 10: Performance comparison between Logistic Regression and MLP

REFERENCES

[1] M. Bruno, R. Lambiotte, and F. Saracco, "Brexit and bots: Characterizing the behaviour of automated accounts on Twitter during the UK election," EPJ Data Sci., vol. 11, no. 1, p. 17, 2022.

[2] E. Alomari and R. Mehmood, "Analysis of tweets in Arabic language for detection of road traffic conditions," in Smart Cities, Infrastructure, Technologies and Applications. Berlin, Germany: Springer, 2017, pp. 98–110.

[3] M. Haselmayer and M. Jenny, "Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding," Qual. Quant., vol. 51, pp. 2623–2646, 2017.

[4] B. P. Aniruddha Prabhu, B. P. Ashwini, T. Anwar Khan, and A. Das, "Predicting election result with sentimental analysis using Twitter data for candidate selection," in Innovations in Computer Science and Engineering: Proc. 6th ICICSE 2018, Singapore: Springer, pp. 49–55, 2019.

[5] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques," Hum.-Centric Comput. Inf. Sci., vol. 9, p. 24, 2019.

[6] A. AlDayel and W. Magdy, "Stance detection on social media: State of the art and trends," Information Processing & Management, vol. 58, no. 4, p. 102597, 2021.

[7] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, "Spread of hate speech in online social media," in Proc. 10th ACM Conf. Web Sci., Jun. 2019, pp. 173–182.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.
[9] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in Proc. 2016 IEEE 16th Int. Conf. Data Mining (ICDM), Barcelona, Spain, Dec. 2016, pp. 439–448.
[10] B. Joyce and J. Deng, "Sentiment analysis of tweets for the 2016 US presidential election," in Proc. 2017 IEEE MIT Undergraduate Research Technology Conf. (URTC), Cambridge, MA, USA, Nov. 2017, pp. 1–4.
[11] A. R. Alaei, S. Becken, and B. Stantic, "Sentiment analysis in tourism: Capitalizing on big data," J. Travel Res., vol. 58, pp. 175–191, 2019.