CS3244 Project 2 Group 14

Jared Tan, Luo Dan, Shi Kexin, Shim Jaejun, Tan Jia Jun

18 March 2023

1 Introduction

Toxic or harmful comments on the internet have become increasingly prevalent in recent years, affecting users in various ways. These comments can include hate speech, harassment, cyberbullying, and trolling, among others. Such behavior can have severe negative impacts on the mental health and wellbeing of users, including increased stress, anxiety, and depression. Additionally, toxic comments can lead to a toxic culture online, creating an environment that discourages positive interactions and discourages people from participating in online discussions. The negative effects of toxic comments can be particularly damaging to vulnerable communities such as minorities, women, and children. Therefore, it is essential to address this issue by promoting healthy online behavior and enforcing regulations that protect users from harmful content.

2 An outline of the application

Our application is targeted toward social media platforms that wish to give its users the option to filter out malignant comments and hate speech that is abusive in nature, or to censor such comments from users who are underaged. Currently, many social media platforms rely only on a simple filtering of prohibited words (e.g. YouTube filters), or their users manually flagging or reporting malignant comments (e.g. Reddit) before a manual review by admins takes place, resulting in inefficiencies and delays. Depending on the platform, many users who frequently make malignant comments may also never be reported and are allowed to stay active on the platform, without a manual review of their account activities by site admins.

Some of the use cases include:

- Allowing users to censor all hateful comments towards them, instead displaying that the comment has been hidden, and allowing users to view the comment only if they wish to. This is especially useful for those who are concerned about their mental health.
- For public figures who are active on social media to automatically block out hate speech towards them instead of hiring PR personnel to filter out the comments for them.
- For staff members, to flag comments for manual review, so as to identify users who frequently violate the community guidelines.
- For microblogging sites like Reddit, to analyse which forums have the highest proportion of comments that are flagged as malignant, so as to direct more site administrators to these forums etc.

We assume that such platforms can gather information consistently to update and improve the model application.

The performance criteria used would mainly be precision and recall. We wish to first minimise the proportion of false negatives (indicated by recall) because we do not wish to overlook comments that are malignant, therefore effectively achieving the intended goal of the application. At the same time, we wish to minimise the proportion of false positives (indicated by precision), as having a large proportion of false positives could negatively impact user experience with the application. We have decided to prioritise recall over precision slightly, as misclassified malignancy can be manually reviewed, reported and rectified by the social media community later on. A summary statistic (namely the average recall) is used, given that the recall of one metric and the precision of the other are directly related. A confusion matrix will also be used to present the findings.

3 Description of the raw data

3.1 Data Source

This dataset is obtained from a Kaggle competition *Toxic Comment Classification Challenge*[1] that aims to identify and classify different types of toxic comments made on Wikipedia, generalisable to those made on other social media platforms.

3.2 Dataset Description

The dataset contains approximately 220K samples of text or comments posted by Wikipedia users labelled based on their type of toxicity. There are six possible labels for each sample: 'toxic', 'severe_toxic', 'obscene', 'threat', 'insult', and 'identity_hate', where a '1' under each column indicates that the sample is labelled under that category and a '0' indicates otherwise.(Figure 1) For clarity, we will use the term 'malignant' to refer to samples that are labelled under at least one of these categories, i.e. samples that have at least one label as '1'.

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

Figure 1: Sample of raw data

Through preliminary analysis, we have also discovered that the data is highly imbalanced, with non-malignant comments outnumbering malignant comments roughly 9 to 1 (Figure 2).

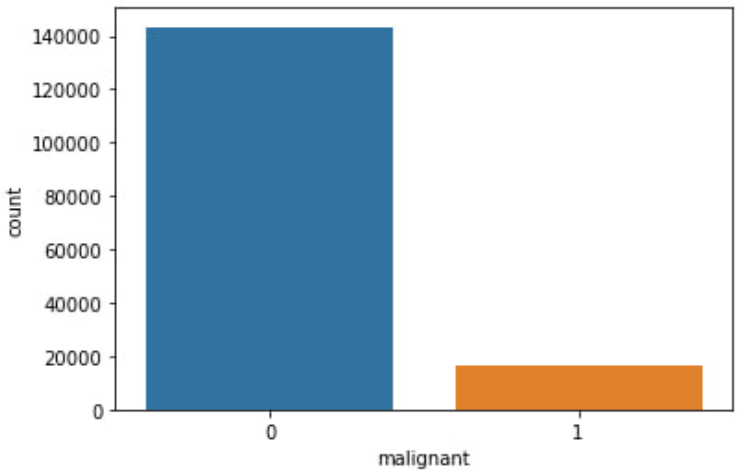

Figure 2: Number of malignant vs non-malignant comments in the raw dataset

4 Preliminary model

The objective is to derive a dataset and model that extracts the most important features from samples, therefore allowing us to build an accurate classification model to effectively classify comments into malignant and non-malignant. To achieve this, we combined and simplified the labels as follows.

First, we note that all comments labelled as 'severe_toxic' are also labelled 'toxic'. However, the rest of the labels are independent of one another, and each sample may have multiple labels. As our aim is to distinguish

[1]https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data

between malignant and non-malignant comments only, and not to differentiate them based on different degrees of toxicity or different types of malignancy, we combined the labels into a new label named 'malignant' following our definition of 'malignant' (Figure 3).

	id	comment_text	malignant
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0

Figure 3: Sample of data with combined labels

To account for the imbalance in the dataset, we then performed random undersampling by splitting the data into their respective categories, and randomly selecting 16,225 samples (the size of the minority group) from the majority group. This is combined with the original minority group data, giving us a dataset with an even distribution of class labels.

However, this method of population correction has a possible drawback in that it may discard samples that could be useful in our ML algorithm. Therefore, we also attempted to train a kNN and DT model on the original, unbalanced dataset. Other methods to mitigate the imbalance in the dataset will be attempted in the future.

For the two aforementioned datasets (balanced and unbalanced), basic preprocessing on text data was applied. The unnesessary parts of strings (i.e. punctuation, stop words) were eliminated, and the documents were tokenized into separate words. Then, the tokenized words were then lemmatized using python libraries (NLTK) Next, the original and the balanced datasets were tested on one label - namely the class 'malignant' - using the default kNN and DT algorithm implemented using the *scikit-learn* library. The results are given in the table below. While the Precision and Recall of the kNN model were unsatisfactory regardless of the dataset used, the DT model had high scores for Precision and Recall for both datasets (Table 5). This is promising given that the models used are very simple for such a complex problem.

Table 1: KNN on Original Dataset

	Actual +	Actual -
Predicted +	56494	1271
Predicted -	4125	2118

Table 2: KNN on Balanced Dataset

	Actual +	Actual -
Predicted +	4842	52893
Predicted -	83	6160

Table 3: Decision Tree on Original Dataset

	Actual +	Actual -
Predicted +	52471	5264
Predicted -	1235	5008

Table 4: Decision Tree on Balanced Dataset

	Actual +	Actual -
Predicted +	44571	13164
Predicted -	570	5673

Model	Dataset	Balanced Accuracy	Precision	Recall
KNN	Original Dataset	0.6586	'0': 0.93	'0': 0.98
			'1': 0.62	'1': 0.34
KNN	Balanced Dataset	0.5353	'0': 0.98	'0': 0.08
			'1': 0.10	'1': 0.99
Decision Tree	Original Dataset	0.8555	'0': 0.98	'0': 0.91
			'1': 0.49	'1': 0.80
Decision Tree	Balanced Dataset	0.8403	'0': 0.99	'0': 0.77
			'1': 0.30	'1': 0.91

Table 5: Summary of results

5 Plans for future work

Our preliminary experiment using a DT model on the dataset shows satisfactory Precision and Recall scores. However, we note that there are a few areas that require improvements, such as the precision for malignant data points in the DT model. Moreover, the Recall scores for both categories can be further improved.

5.1 Feature Engineering

For our preliminary experiments, only some basic steps of preprocessing text were attempted. In the future, we may try to extract more features, especially features that are semantically meaningful but are hard to extract. (e.g., email address etc.) While we have removed all punctuation for simplicity, some punctuation might be related to text sentiment. Further research will be conducted to extract such punctuation. Additionally, our current method of data preprocessing fails to account for the sequence of words in a sentence. Therefore, we would also experiment with word embeddings, which is a way to represent words and whole sentences in a numerical manner.

5.2 Imbalanced Dataset

To mitigate the imbalanced dataset, only the random undersampling method has been attempted, and using performance metrics that are resistant to imbalanced datasets. (i.e. Balanced Accuracy). This is because normal resampling methods such as undersampling and SMOTE (Synthetic Minority Oversampling Technique) are limited to fundamentally numerical values, and are poorly suited to text dataset applications (although they have been transformed into numerical representation). Therefore, in the future, methods such as Data Augmentation, for example backtranslation or using synonym replacement, will be applied.

5.3 More Complex Models

In this stage, we have only utilised simple models such as kNN and DT. In the further stages, we would like to attempt more complex state-of-the-art models that may be better suited to perform sentiment analysis. These complex models include LSTM (Long Short-Term Memory) and Transformers (i.e., BERT - Bidirectional Encoder Representations from Transformers). Specifically, for LSTM, pre-trained embeddings such as Word2Vec (developed by Google) or GloVe (developed by Stanford) will be applied. We acknowledge that to train LSTM or Transformers, we have to create a new dataset completely different from the one created above. However, we will be able to compare the performances of these models and the preliminary models, as they are created under the same goal and use the same raw dataset. It would be easier to imagine each intermediate dataset created to train the models as a part of our model creation.

6 Conclusion

In conclusion, we aim to improve the user experience on social media platforms by providing an effective application for detecting and filtering malignant comments. We performed random undersampling to address the high imbalance in the raw data, used kNN and DT algorithms to construct our models, and evaluated the performance using Precision and Recall. Our initial tests yielded promising results, however we plan to continue to refine the model by exploring other more advanced methods like LSTM. We also intend to experiment with further feature pre-processing and engineering, as well as use other methods to mitigate imbalanced datasets.