“CIVIL COMMENT” CASE: A STUDY ON CLASS IMBALANCE PROBLEM AND TOXICITY DETECTION

Ezgi GÜNBATAR

Table of Contents

“CIVIL COMMENT” CASE: A STUDY ON CLASS IMBALANCE PROBLEM AND TOXICITY DETECTION

1. Introduction

Today, more and more people can interact with each other thanks to social media or other online platforms. Through these platforms, people with all kinds of diversity such as race, gender, religion, sexuality, etc. can meet more than they would ever encounter in daily life. Everyone feels free to write their comments about any topic. Unfortunately, while doing this, some people forget the fact that the person they are dealing with is a real human being. This leads us to think about the "toxicity" issue. According to Risch & Krestel (2019), “a toxic comment is defined as a rude, disrespectful, or unreasonable comment that is likely to make other users leave a discussion”. If we think about the volume of messages sent online every day, it isn’t possible to prevent the spread of toxic messages and remove them manually. Therefore, toxicity detection methods are of great importance. (Bonetti et al., 2023)

In this study, user comments on an online platform were used as data. I focused on 2 main topics of toxicity detection: Class imbalance problem and detecting toxic comments. To ride of class imbalance problem two sampling methods were used: Random undersampling and oversampling with SMOTE. After this step, two machine learning (ML) models (Logistic regression and Stochastic gradient descent classifier) were applied to the original and resampled data to detect toxic comments. In the end, performance evaluations and comparisons were made and reported.

2. Literature Review

Toxicity detection is a difficult problem because toxicity comes in many forms, such as obscene language, insults, threats, hate speech, and many other forms. (Bonetti et al., 2023) For this purpose, many studies and different approaches were conducted. Brassard-Gourdeau & Khoury (2019), used sentiment information to improve the toxicity detection task and found that there is a clear correlation between sentiment and toxicity. Kiilu et al. (2018), compared different sentiment classifiers, including Logistic Regression, Linear SVC and Naïve Bayes for detecting and classifying hate speech on Twitter. Miró-Llinares (2018), created an algorithm for detecting online hate speech through the application of the Random Forest (RF) technique. Almatarneh (2019), compared different ML classifiers to identify the hate speech focused on two targets (women and immigrants) in English and Spanish tweets. The paper showed that Support Vector Machine (SVM), Complement Naive Bayes and Random Forest are clearly by far the best performers compared to the others. (Bonetti et al., 2023)

3. Data & Methodology

3.1.Data

For this study, "Jigsaw Unintended Bias in Toxicity Classification" data from Kaggle was used. This data contains public comments on the "Civil Comment" platform which is an online commenting platform and aims to reduce toxic comments and maintain a respectful atmosphere in online communities. The variable descriptions are listed below:

- *"comment_text":* the text of each comment is in the column.
- *"target":* shows the toxicity label of the *"comment_text".* It's the target variable of the study and on a scale between 0-1. If *target* >= 0.5, this indicates the situation of toxicity.
- Other toxicity subtype and identity attributes: Models don't need to predict these variables; they carry just additional information for the study. *"severe_toxicity", "obscene", "threat", "insult", "Asian", "atheist", "bisexual" … etc.*

Before starting the analysis, the target variable was labeled as toxic (equal and higher than 0.5) and non-toxic (lower than 0.5). The distribution of labels showed that this was highly imbalanced data: 657 comments with the label "toxic" and 14617 comments with the label "non-toxic". This situation addressed that, it's necessary to find a cure for class imbalance in the first step.

3.2. Resampling

When a dataset shows an imbalance between its classes, in other words, some classes in data are underrepresented, the classification task becomes more challenging. Traditional classification algorithms tend to favor majority over minority class elements due to their incorrect implicit assumption of an equal class representation during learning. As a result, recognition of minority examples is hindered. Since minority classes are usually the ones of interest, custom techniques are required to deal with such data skewness (Vluymans, 2018). Especially, in matters related to detecting something, dataset is more prone to imbalance problem. Toxicity detection is an example of this area. Resampling methods are one of the remedies for the problem.

Resampling methods aim to modify the dataset to reduce the discrepancy among the sizes of the classes. Some of these methods are: Random oversampling, random undersampling, SMOTE oversampling, ADASYN oversampling etc. In this study, undersampling and oversampling with SMOTE methods were used before the classification step.

Random undersampling is a basic sampling method that is based on eliminating instances from the majority class randomly. The disadvantage of this method is that it causes observation and information loss when the data is not large enough. **Synthetic Minority Over-sampling Technique (SMOTE)** is an oversampling method based on creating synthetic instances for the minority classes. The algorithm takes each minority class sample and introduces synthetic samples along the line joining the current instance and some of its k-nearest neighbors from the same class (Padurariu & Breban, 2019). The disadvantage of the SMOTE is that it is too dependent on the first sample it selects, so if that sample is an observation close to the majority class, it might produce examples too deeply in the majority class space.

3.3.Preprocessing

When working with text, in order to make the text data usable for the learning models, it needs to be converted into numerical vectors. **Term Frequency-Inverse Document Frequency (TF-IDF)** one of the vectorization techniques was used to extract these vectors in the study, so that they can be used by ML models. TF-IDF relates the number of times a term appears in a document to the number of documents in which it appears and it performs a weighting to promote the least common terms and discount the most frequent terms (Bonetti et al., 2023).

Before starting analysis, data was split into train (%70) and test (%30) sets.

3.4.Classification

When it comes to detecting toxicity in the texts it can be mentioned about supervised and unsupervised learning algorithms. Without labeling the data, it can only be used for unsupervised learning, such as clustering or dimensionality reduction. Supervised learning approaches require labeled data. Since the data was labeled in this study, supervised learning methods Logistic

Regression and Stochastic Gradient Descent Classifier were used for detecting toxicity in comments.

Logistic regression (LR) is a powerful supervised ML algorithm used for binary classification problems which is a transformation of a linear regression using the nonlinear sigmoid function. It is easy to implement, has no assumptions on distributions of classes, easily extends to multiple classes, and has a natural probabilistic view of class predictions. (ScienceDirect, 2023)

Stochastic Gradient Descent (SGD) is a simple and easy-to-apply but very efficient approach to fitting linear classifiers and regressors under convex loss functions (like SVM and LR). Even though SGD has been around in the ML community for a long time, it has received a considerable amount of attention just recently in the context of large-scale learning. (Scikit-learn, 2023)

These two ML methods were applied on original, undersampled and oversampled (SMOTE) training datasets then model results were compared in terms of performance metrics.

4. Results

Because the datasets are imbalanced, accuracy isn't used but precision, recall and (weighted) macro F1-score. Weighted F1-score focuses on the classification of the minority class by emphasizing the respective penalty for misclassification (Risch & Krestel, 2019). Performance metrics for the majority class (non-toxic) is between 0.96-1 for all method combinations. Also, this situation misleads the accuracy. Since the goal of the study isn't detecting non-toxic comments, majority class performance metrics and accuracy don't carry importance. Therefore, tables represent precision, recall and F1 score for toxic comments and macro F1 score.

When we look at Table 1, for both ML model performed poorly on the original (imbalanced) dataset.

Table 1

Without Resampling

Model	Precision	Recall	F1	Macro F1
LR	1	0.01	0.01	0.49
SGD	0.92	0.06	0.11	0.54

Note. Precision, recall and F1 score for toxic comments

For undersampled data, there are better results than the imbalanced dataset. When we look at Table 2, LR and SGD show similar performances to each other, but it still can't be said good. LR model can correctly identify 69% of the actual toxic comments. In other words, when there is a toxic comment, the model successfully flagged it as such 69% of the time. However, both methods have poor precision scores. This means that out of all the instances predicted as positive by the model, only 10% of them are truly positive. Precision shows how reliable the model is when it flags a comment as toxic. In other terms, if the model predicts a comment as toxic, there is a relatively high chance (90%) that the prediction is a false positive.

Table 2

Undersampling

Model	Precision	Recall	F1	Macro F1
LR	0.10	0.69	0.18	0.51
SGD	0.10	0.70	0.18	0.51

Note. Precision, recall and F1 score for toxic comments

In Table 3, SMOTE oversampling method supplies the best performance for macro F1 metrics. Recall values decreased but precision values increased. In other words, while the ability to capture toxic comments got worse, the reliability of the models got better.

Table 3

Oversampling with SMOTE

Model	Precision	Recall	F1	Macro F1
LR	0.45	0.44	0.45	0.71
SGD	0.62	0.39	0.48	0.73

Note. Precision, recall and F1 score for toxic comments

5. Discussion and Conclusion

In the study, public comments on the “Civil Comment” platform were used for toxicity detection. Since there was a high imbalance between toxic and non-toxic comments, two resampling methods were used: undersampling and oversampling with SMOTE. After fixing the class imbalance problem, text data is converted to vectors with the TF-IDF method then two ML models were applied: Logistic regression and stochastic gradient descent.

Performance metrics showed that the worst case is using the original imbalanced data. In this case, models can’t learn toxic units from the data due to insufficient numbers. It’s obvious that, resampling methods provide better results than original data. If we compare between undersampling and SMOTE, SMOTE has better F1 and macro F1 scores. It would be preferable but still has not perfect performance. In undersampled data, although the success of the models in detecting toxic comments is good (69%- 70%), the reliability of the model is very poor (%10). No significant performance difference is observed between ML models.

The most important reasons affecting the success of the study are the small size of the dataset and the imbalance of the classes. The class of interest contains few observations. Although we use resampling to solve this problem, it does not give as good results as having real information. In the undersampling method, data and information loss occur when removing observations from the majority class. SMOTE oversampling method is highly dependent on the first sample it selects, so if that sample is an observation close to the majority class, it might produce examples like the majority class. At the same time, since synthetic data is produced based on existing observations, it can never provide as much information to the models as real data.

References

- Risch , J., & Krestel, R. (2020). Deep Learning-Based Approaches for Sentiment Analysis. Toxic Comment Detection in Online Discussions. Springer. DOI: 10.1007/978-981-15-1216-2_4
- Bonetti, A., Martínez-Sober, M., Torres, J. C., Vega, J. M., Pellerin, S., & Vila-Francés, J. (2023). Comparison Between Machine Learning and Deep Learning Approaches for the Detection of Toxic Comments on Social Networks. Applied Sciences, 13(10), 6038.
- Vluymans, S. (2018). Learning from Imbalanced Data. Dealing with Imbalanced and Weakly Labelled Data in Machine Learning using Fuzzy and Rough Set Methods. (pp. 81-110) Springer Cham.
- Padurariu, C., Breban,M. E. (2019). Dealing with Data Imbalance in Text Classification. Procedia Computer Science. Elsevier. https://doi.org/10.1016/j.procs.2019.09.229.
- Brassard-Gourdeau, E.; Khoury, R. (2019). Subversive Toxicity Detection using Sentiment Information. In Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, 1–2 August 2019. (pp. 1–10)
- Kiilu, K.K., Okeyo, G., Rimiru, R., Ogada K., (2018). Using Naïve Bayes Algorithm in detection of Hate Tweets. International Journal of Scientific and Research Publications. DOI: 10.29322/IJSRP.8.3.2018.p7517.
- Miró-Llinares, F., Moneva, A., Esteve, M. (2018) Hate is in the air! However, where? Introducing an algorithm to detect hate speech in digital microenvironments. Crime Science. DOI: 10.1186/s40163-018-0089-1
- Almatarneh, S., Gamallo, P., Pena, F.J.R., Alexeev, A. (2019). Supervised Classifiers to Identify Hate Speech on English and Spanish Tweets. Digit. Libraries at the Crossroads Digital Information for the Future. DOI: 10.1007/978-3-030-34058-2_3
- Kaggle (2023, 13 December). https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data?select=train.csv
- Science Direct (2023, 14 December). https://www.sciencedirect.com/topics/computer-science/logisticregression#:~:text=Logistic%20regression%20is%20a%20process,%2Fno%2C%20and%20so%20on