# Generative LLM resampling for the class imbalance problem with hate speech detection

Nicolas Antonio Cloutier and Nathalie Japkowicz[1]

American University, Washington, D.C., USA

**Abstract.**

## 1 Introduction

In recent years, hate speech has become increasingly mainstream and common within social media sites [6]. This rise has not only caused online spaces to become less hospitable, but also had several offline effects. On top of having severe psychological effects on the recipient [6], online hate speech also played a roll in disseminating extremist anti-Rohingya voices leading to violence in Myanmar [3], and has provided motivation for several perpetrators of offline violent hate crimes [6]. These events have motivated responses from numerous parties, including the social media sites themselves, that are increasingly looking to stop the spread of these messages [8], and governmental bodies, that are seeking to regulate or prevent the spread of hate speech [1]. At the same time as hate speech has been becoming more common, social media sites have been generating more and more content, with popular social media site Twitter generating an average of 500 million tweets per day in 2019 [5].

With these developments and the large amount of content on social media sites, these sites have been increasingly looking to automatic detection methods for hate speech [8]. These methods use Machine Learning (ML) to automatically detect and classify hateful speech, removing some of the work done by moderators, whose primary job is to respond reactively to user reports of hate speech [8]. Chandra et al. (2021) used a combination of image and text processing and classification algorithms to classify images and text on social media sites Twitter and Gab [2]. We use their Twitter dataset to further analyze the presence of hate speech on Twitter and investigate new algorithms for classification.

One difficulty with this dataset is that it is imbalanced, meaning one classification is far more common in the dataset than another. Imbalanced data can negatively impact the performance of ML classification algorithms [7], affecting numerous domains that use ML classification. Many ML algorithms are inadequately prepared to handle the class imbalance problem [7], leading many to look to other solutions, including resampling methods, that in some way change the training data in order to allay the effects of the class imbalance problem [4].

With this in mind, there are two crucial research questions this paper seeks to answer. First: how can automatic ML methods be improved in the domain of text classification for hate speech detection? Second: how can the class imbalance

problem be dealt with for text data? These are the questions we seek to provide answers to, with the hope that they may inform future research and hate speech detection systems.

## 2 Previous work

## 3 Methodology

## 4 Results

## 5 Discussion

## 6 Conclusions

## References

1. Banks, J.: Regulating hate speech online. International Review of Law, Computers and Technology **24**(3), 233–239 (2010)
2. Chandra, M., Pailla, D., Bhatia, H., Sanchawala, A., Gupta, M., Shrivastava, M., Kumaraguru, P.: "Subverting the Jewtocracy": Online Antisemitism Detection Using Multimodal Deep Learning. ACM WEB SCIENCE CONFERENCE 2021, pp. 148-157.
3. Green, P., MacManus, T., De la Cour Venning, A.: Countdown to Annihilation: Genocide in Myanmar. International State of Crime Initiative, London (2015)
4. Japkowicz, N., Shaju, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis **6**(5), 429–449 (2002)
5. Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, M.: Detecting and Monitoring Hate Speech in Twitter. Sensors **19**(21), 4654–4691 (2019)
6. Siegel, A.: Online Hate Speech. Social media and democracy: The state of the field, prospects for reform, 56–88 (2020).
7. Sun, Y., Wong, A., Kamel, M.: Classification of Imbalanced Data: A Review. International Journal of Pattern Recognition and Artificial Intelligence **23**(4), 687–719 (2009)
8. Ullmann, S., Tomalin, M.: Quarantining online hate speech: technical and ethical perspectives. Ethics and Information Technology **22**, 69–80 (2020)