Generative LLM oversampling for the class imbalance problem in hate speech detection

Nicolas Antonio Cloutier and Nathalie Japkowicz[1]

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 [21]. 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 [21], online hate speech also played a roll in disseminating extremist anti-Rohingya voices leading to violence in Myanmar [10], and has provided motivation for several perpetrators of offline violent hate crimes [21]. 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 [23], and governmental bodies, that are seeking to regulate or prevent the spread of hate speech [2]. 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 [17].

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 [23]. 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 [23]. Chandra et al. [4] used a combination of image and text processing and classification algorithms to classify images and text on social media sites Twitter and Gab [4]. 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 [22], affecting numerous domains that use ML classification. Many ML algorithms are inadequately prepared to handle the class imbalance problem [22], leading many to look to other solutions, including resampling methods, that in some way change the training data by adding or removing data points in order to allay the effects of the class imbalance problem [12].

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

Antisemitism detection, as well as hate speech detection generally, has been the subject of much research. Martins et al. [16] used ML models to analyze and classify antisemitism in posts on social media sites Gab and 4chan, and González-Pizarro and Zanettou [9] used large language models (LLMs) to do the same. Chandra et al. [4] introduced a new dataset of labeled antisemitic posts, including text and images. These posts were labeled not only for whether or not they are antisemitic, but also their type of antisemitism, with the researchers grouping antisemitic posts into political, economic, religious, and racial antisemitism, and trained models to classify both antisemitic status and type of antisemitism. This Twitter dataset is imbalanced, with the non-antisemitic class far outnumbering the antisemitic class, and with political and racial antisemitism being more common than religious and economic antisemitism.

Imbalance is a common issue in ML classification problems. The class imbalance problem can severely negatively affect model predictive power, with the less common class (termed the "minority class") is often misclassified due to its low prevalence in the dataset relative to the larger class (the "majority class") [1]. This problem has affected fields as distinct as medicine, fraudulent call detection, and risk management [22]. Due to the severity with which this problem can limit model performance and its widespread nature, being seen in numerous distinct fields, it has become a focus of researchers as an area of improvement [1], with numerous techniques being created to allay the effects of the problem.

One such technique is resampling, altering the training data by adding new data points or taking existing data points away in order to improve model performance. Generally, there are two types of resampling: oversampling, which describes the process of adding synthetic data points to a dataset, and undersampling, which describes the process of removing existing data points from a dataset [20]. Resampling techniques can improve the performance of ML models when trained on imbalanced data [15] [13], advancing the state of the field of ML in problems with unbalanced datasets.

One method for oversampling is the use of generative models, including generative adversarial networks (GANs) and autoencoders to generate synthetic data. These methods seek to match the distribution of the original dataset and create synthetic datapoints in accordance with that distribution [11], with the goal of creating authentic synthetic examples for model training. They have achieved success in their applications to primarily computer vision and tabular data [11] [6] [3] [5]. Applications of these generative models to NLP do occur [18], they are generally less common than applications to other areas. Natural language pro-

cessing (NLP) tends to use more traditional statistical methods for oversampling such as SMOTE and random oversampling [24] [8].

While these advances in resampling have been occuring, similar advances have been made in generative LLMs, such as the GPT series of models from OpenAI, which have the ability to generate human-like text [7] and been applied to patent claim generation [14], healthcare education [19]. These LLMs, despite their ability of producing authentic examples given prompts, have not been widely investigated as a method for oversampling text data.

3 Methodology

4 Results

5 Discussion

6 Conclusions

References

1. Abd Elrahman, S. M., Abraham, A: A review of class imbalance problem. Journal of Network and Innovative Computing **1**, 332–340 (2013)
2. Banks, J.: Regulating hate speech online. International Review of Law, Computers and Technology **24**(3), 233–239 (2010)
3. Bellinger, C., Japkowicz, N., Drummond, C.: Synthetic oversampling for advanced radioactive threat detection. IEEE CONFERENCE ON MACHINE LEARNING AN APPLICATIONS 2015.
4. 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.
5. Dai, W., Ng, K., Severson, K., Huan, W., Anderson, F., Stultz, C.: Generative oversampling with a contrastive variational autoencoder. IEEE INTERNATIONAL CONFERENCE ON DATA MINING 2019, pp. 101-109.
6. Engelmann, J., Lessmann, S.: Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. Expert Systems with Applications **174**, (2021)
7. Floridi, L., Chiriatti, M.: GPT-3: Its nature, scope, limits, and consequences. Minds and Machines **30**, 681–694 (2020)
8. Glazkova, A.: A comparison of synthetic oversampling methods for multi-class text classification. *arXiv preprint arXiv:2008.04636t*, (2020)
9. González-Pizarro, F., Zannettou, S.: Understanding and detecting hateful content using contrastive learning. *arXiv preprint arXiv:2201.08387*, (2022)
10. Green, P., MacManus, T., De la Cour Venning, A.: Countdown to Annihilation: Genocide in Myanmar. International State of Crime Initiative, London (2015)
11. Hao, J., Wang, C., Zhang, H., Yang, G.: Annealing genetic GAN for minority oversampling. *arXiv preprint*, (2020)

12. Japkowicz, N., Shaju, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis **6**(5), 429–449 (2002)
13. Khushi, M., Shaukat, K., Alam, T. M., Hammed, I. A., Uddin, S., Luo, S., Yang, X., Consuelo Reyes, M.: A Comparitive Performance Analysis of Data Resampling Methods on Imbalance Medical Data. IEEE Access **9**, 109960–109975 (2021)
14. Lee, J. S., Hsiang, J.: Patent Claim Generation by Fine-Tuning OpenAI GPT-2. World Patent Information **62**, (2020)
15. Lee, P. H.: Resampling methods improve the predictive power of modeling in class-imbalanced datasets. International Journal of Environmental Research and Public Health **11**(9), 9776–9789.
16. Martins, R., Gomes, M., Almeida, J. J., Novais, P., Henriques, P.: Hate speech classification in social media using emotional analysis. BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS 2018, pp. 61-66
17. 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)
18. Phung, N. M., Mimura, M.: Evaluation of a cGAN Model and Random Seed Oversampling on Imbalanced JavaScript Datasets. Journal of Information Processing **30**, 591–600 (2022)
19. Sallam, M.: The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *medRxiv preprint*, (2023)
20. Shelke, M. S., Deshmukh, P. R., Shandilya, V. K.: A review on imbalanced data handling using undersampling and oversampling technique. International Journal of Recent Trends in Engineering Research **3**(4), 444–449 (2017).
21. Siegel, A.: Online Hate Speech. Social media and democracy: The state of the field, prospects for reform, 56–88 (2020).
22. 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)
23. Ullmann, S., Tomalin, M.: Quarantining online hate speech: technical and ethical perspectives. Ethics and Information Technology **22**, 69–80 (2020)
24. Wijaya, I. D., Putrada, A. G., Oktaria, D.: Overcoming Data Imbalance Problems in Sexual Harassment Classification with SMOTE. International Journal on Information and Communication Technology **8**(1), 20–29 (2022).