Conferences > 2024 15th International Confe...

The impact of employing sampling-based strategies to address the issue of an imbalanced dataset in hate speech and offensive language categorization

Publisher: IEEE Cite This PDF

Anju ; Inderdeep Kaur Aulakh ; Raj Kumari **All Authors**

33
Full
Text Views

Authors

Figures

References

Keywords

Metrics

More Like This

Abstract:

The improper usage of social media platforms has led to global disturbances as individuals often engage in trolling, judging, and criticizing others, including public figures. Hate speech is a serious problem since it is defined as harsh, hurtful and abusive language directed at groups because of characteristics like race, religion, age, gender, or ethnicity. The vast amount of content generated daily necessitates the development of automatic methods to identify and manage such harmful content. This study proposes a comparative analysis of classifiers combining Natural Language Processing (NLP) with Machine Learning (ML) techniques to detect hate and offensive language in a publicly accessible Twitter dataset. Sampling-based strategies were used to balance the imbalanced dataset. Data preprocessing including tokenization, stemming, and lemmatization, was employed to transform raw data into a clean and structured format. The technique for feature extraction employed was Term Frequency Inverse Document Frequency (TF-IDF) to generate a set of features from dataset. The metrices employed to evaluate the results were F1 score, Accuracy, Precision, and Recall. The imbalanced dataset was first subjected to three machine learning classifiers Logistic Regression (LR), Support Vector Machine (SVM) and Decision Tree (DT), without the use of a sampling technique, resulting in low performance across the metrices. Subsequently, Random Over Sampling (ROS) was used to balance the dataset, and the classifiers were reapplied, showing improvements in performance. Additional sampling approaches, namely Random Under Sampling (RUS), Synthetic Minority Over Sampling Technique (SMOTE), and Adaptive Synthetic Sampling Method (ADASYN), were employed to repeat this procedure. Among all classifiers and sampling methods, the Decision Tree with ROS achieved the highest accuracy of 96%, outperforming all other combinations. The studydemonstrates that the combination of ROS and Decision Tree i...

Show More

Published in: 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)

Date of Conference: 24-28 June 2024

Date Added to IEEE *Xplore*: 04 November 2024

ISBN Information:

ISSN Information:

Electronic ISSN: 2473-7674

Print on Demand(PoD) ISSN: 2162-7665

DOI: 10.1109/ICCCNT61001.2024.10726250

Publisher: IEEE

Conference Location: Kamand, India

Sign in to Continue Reading

Authors

Figures

References

Keywords

Metrics

IEEE Personal Account

CHANGE USERNAME/PASSWORD

Purchase Details

PAYMENT OPTIONS

VIEW PURCHASED DOCUMENTS

Profile Information

COMMUNICATIONS PREFERENCES

PROFESSION AND EDUCATION

TECHNICAL INTERESTS

Need Help?

US & CANADA: +1 800 678 4333

WORLDWIDE: +1 732 981 0060

CONTACT & SUPPORT

Follow

About IEEE *Xplore* | Contact Us | Help | Accessibility | Terms of Use | Nondiscrimination Policy | IEEE Ethics Reporting | Sitemap | IEEE Privacy Policy

Anju