Application of Ensemble Methods to extract the optimal threshold value for permissible amount of profane words in Communal posts

Michael *Santa Clara University*

Geethika *Santa Clara University*

Shaunak *Santa Clara University*

Jack *Santa Clara University*

Milson *Santa Clara University*

Akshat *Santa Clara University*

Editor: Our Team

Abstract

With the influx of textual analysis in different forms, this paper presents the ability of Machine Learning models to conduct a large schema based analysis for communal posts, which as such forms a base line platform of our project to provide a positive reinforcement space for teenagers who are suffering with depression or mild state of similar problems. The primary purpose of this paper is to convey the approach of finding out the numeric threshold metric that is used for classifying a post between abusive and non-abusive categories so the people who are trying to detoxify from their mental health problems are not further nudged down the same road they are willing to ebb away from.

Keywords: Machine Learning Models, Ensemble Methods, Filtering Techniques, Sentiment Analysis

1. Introduction

The use of online platforms has become a ubiquitous part of modern-day communication. These platforms offer individuals an avenue to share their experiences, connect with others, and seek support in times of distress. One such platform, targeted towards youth dealing with depression, aims to provide a safe space for individuals to share their stories, connect with others, and find motivation to overcome their struggles. However, the nature of online communication opens up the possibility of inappropriate language and behavior, which can create a negative experience for the platform's users. Although, there has been considerable analysis of Social media done in the past [Chow and Liu (1968)], but none of that has taken this route towards an exclusively optimistic style of representing content.

One of the primary concerns for the administrators of the platform is the use of curse words in posts. While some use of profanity may be acceptable in certain contexts, excessive and inappropriate use of curse words can create a hostile and unwelcoming environment for users seeking support. The challenge for the platform administrators is to find the most optimal threshold for the number of curse words allowed in a post that would balance the need for freedom of expression with the maintenance of a positive and supportive community.

In this research paper, we investigate the use of machine learning models to determine the optimal threshold for the number of curse words allowed in a post on the platform. Our study aims to provide a quantitative analysis of the relationship between the frequency of curse words in posts and user engagement, as well as the impact of moderation strategies on user behavior. The results of our study can inform the platform administrators on the best practices for maintaining a positive and supportive community for youth dealing with depression.

2. Project

Our project is a website that aims to provide a safe space for individuals to share their stories, find motivation, and connect with others who have gone through similar experiences. The website caters to individuals who are looking for inspiration and motivation to overcome their struggles and seek support for mental health issues.

The website provides various resources for individuals seeking help, including a forum for discussions, a directory of mental health professionals, and a library of self-help materials. The forum allows users to share their experiences and connect with others who have gone through similar struggles. The directory of mental health professionals provides users with information on qualified and licensed professionals who can offer clinical support. The library of self-help materials includes articles, videos, and other resources on mental health and coping strategies.

In the future prospects, our website also aims to provide a private messaging feature that allows users to connect with each other one-on-one to seek advice and support. This feature shall intend to engender a more personalized approach to support and help individuals feel heard and understood.

The project team has prioritized the safety and security of users in the development of the website. We have implemented strict guidelines and moderation strategies to prevent the use of inappropriate language or behavior that can create a negative environment for users seeking support. Additionally, the website has a feature that allows users to report inappropriate content or behavior, which enables the moderation team to take swift action and ensure the safety of the community.

Overall, our project aims to provide a safe and supportive community for individuals seeking inspiration, motivation, and support for mental health issues. We hope that our website can make a positive impact in the lives of those who are struggling and help them find the resources they need to overcome their challenges.

3. Involvement of Novel Theoretical Concept

While there have been various instances of filtering out social media posts in the past based upon the occurrence of certain negatively oriented buzzwords in them[Khanday et al. (2022)], the application of classifying algorithms in those approaches are restrained to an all pervasive structure of profane words existing as part of the content, and hardly does any algorithm deals with the nature of filtering out only the posts which might be anti-thetical towards creating a better society for tackling depression and other pressing problems. The approaches mentioned above largely rely on the overall quantifiable proportion of negative words and thus can be contrary to their purpose of instilling an inclusive decision wherein one gets the freedom to overtly describe their experiences.

Hence, taking that into consideration, we came across the idea of running ensemble methods on a large dataset of online tweets to better boil down the proportion of swear words allowed in such content in order to accurately predict the category of a particular post as offensive or safe.

Another important distinction in our paradigm is also based upon the inference that we try to draw based upon how large a content is, for any particular post. For instance, the threshold value of the proportion of admissibly negative words for a post with a length of 10 can be entirely different from one with length of 100 or 2500 words. Which is why, it's imperative for us to take this distinction into account, before our models run into a high bias or end up underfitting the data.

4. Technical Methods Undertaken

Our project was developed by utilizing various machine learning models, including Random Forest, Logistic Regression, and Linear Discriminant Analysis. We implemented ensemble methods to enhance the accuracy of the models and preprocessed the Tweets dataset to improve the quality of the data. We also created the threshold percentage value for all posts to ensure that the posts were within the acceptable limit of curse words. Our approach involved evaluating the performance of multiple models using packages like PyCaret to deploy and compare the best model.

We considered various evaluation metrics, including precision, recall, F score, and more to measure the performance of each model accurately. Our use of machine learning algorithms, ensemble methods, and careful preprocessing of the dataset resulted in the development of a robust and effective model that can accurately predict the optimal threshold for the number of curse words allowed in a post on our website.

Acknowledgments

We would like to acknowledge support for this project from the National Science Foundation (NSF grant IIS-9988642) and the Multidisciplinary Research Program of the Department of Defense (MURI N00014-00-1-0637).

References

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, 1968.

Akib Mohi Ud Din Khanday, Syed Tanzeel Rabani, Qamar Rayees Khan, and Showkat Hassan Malik. Detecting twitter hate speech in covid-19 era using machine learning and ensemble learning techniques. *International Journal of Information Management Data Insights*, 2(2):100120, 2022. ISSN 2667-0968. doi: https://doi.org/10.1016/j.jjimei.2022.100120. URL `https://www.sciencedirect.com/science/article/pii/S2667096822000635`.