ABSTRACT

Today, the internet and social media are essential tools for communicating and consuming information. About half of individuals now use social media to share their views and opinions, which indicates that it has undergone significant development over time. Over the past 10 years, communication between people has undergone significant transformation, in part due to the ubiquitous development of social media. It has enabled a more connected and knowledgeable world, but it has also paved the path for a brand-new phenomenon: toxic speech. Due to the open platform for content production, debate, and sharing, some fairly opportunistic individuals have participated in toxic speech and generally negative remarks. This pattern served as the basis for our project.

We want to build a highly accurate classifier on hazardous speech using the Random Forest Classifier method in order to efficiently detect the existence of toxic speech in comments and texts. Words like "Obscene," "Toxic," "Severe Toxic," "Threat," "Insult," and "Identity Hate" have been categorized as "Toxic" speech material since they are commonly used together. As a result, it is essential to recognize and naturally eliminate harmful speech from online social media networks. As a consequence, when it comes to classifying provided comments into different levels of toxicity, our model now has higher precision, recall, and accuracy scores.