

## ABSTRACT

The pervasive spread of misinformation and fake news on social media platforms has become a significant concern, impacting public opinion, political discourse, and societal trust. Manual content moderation efforts are often insufficient to address the scale and complexity of the problem, prompting the need for automated solutions that can analyze and classify content at scale. In response to this challenge, we propose the development of an automated content moderation system for identifying and classifying tweets as real or fake, with the additional capability of determining the subjectivity of the tweet as a general fact or opinion. The system leverages machine learning algorithms and natural language processing techniques to analyze the content of tweets, extracting relevant features and patterns to facilitate classification. By training a machine learning model on a diverse dataset of labeled tweets, the system learns to distinguish between genuine and deceptive content, enabling it to automatically flag and moderate potentially harmful tweets in real-time. Additionally, the system incorporates algorithms for subjectivity analysis, providing users with insights into the nature of the content being shared on social media platforms.