# Chapter 9
# Real-Time Hate Speech Detection API: A Scalable Deep Learning Approach

**Vikash Deep Bhaskar**
https://orcid.org/0009-0007-1236-9621
*Madan Mohan Malaviya University of Technology, India*

**Shagufta Shakeel**
https://orcid.org/0009-0002-1820-272X
*Madan Mohan Malaviya University of Technology, India*

## ABSTRACT

*The rise of digital communication has exacerbated the challenge of tackling harmful speech online, particularly hate speech, which dehumanizes individuals or groups based on traits such as race, gender, or ethnicity. Hate speech has been an issue since the start of the Internet, but the advent of social media has brought it to unimaginable heights. To address such an important issue and to improve contextual understanding and classification accuracy, this paper presents a new, scalable, and high-performance method for implementing a real-time hate speech detection API that uses the T5 transformer model to effectively detect highly discussed topics that generate hate speech on twitter. The system uses a deep learning approach, T5EncoderModel, to detect hate speech on annotated tweets. It integrates the T5 model with a dense classification head and is trained using PyTorch. It is deployed via FastAPI to enable real-time classification of social media content. Evaluation demonstrates that our approach outperforms existing models.*

DOI: 10.4018/979-8-3373-3063-1.ch009

## 1. INTRODUCTION

The hateful actions of exploiting social networks on the internet have grown in the same proportion as the massive social interactions using such networks. Among these platforms, Twitter, now rebranded as "X," has drawn particular attention (Boishakhi, Shill, & Alam, 2023). As of 2024, the platform has witnessed a notable surge in hate speech, especially following its acquisition by Elon Musk. Studies reveal that within the first 12 hours of Musk taking over as CEO, more than 4,700 instances of hate speech were documented, averaging 398 incidents per hour. This figure marks nearly a fivefold increase compared to the platform's highest pre-acquisition hourly average, highlighting the urgent need for more robust detection and moderation mechanisms (Reuters, 2024). Because of this big increase, it's now more important than ever to quickly and correctly find hate speech. On Twitter, abusive tweets that include violent behavior toward another person (cyberbullying, politicians, celebrities, products, and so on) or a group of people (a nation, LGBT community, religion, gender, an organization, and so on) are considered hostile tweets. In the investigation of the popular opinion towards a certain group of users in opposition to another, as well as in the prevention of wrongdoing, the detection of hate speech is highly important. Also, it is helpful to filter tweets and then suggest content or train AI chatterbots in the form of tweets (Badjatiya, Gupta, Gupta, & Varma, 2017).

These challenges call for more advanced methods that go beyond traditional techniques. While machine learning algorithms have been widely used for text classification tasks like hate speech detection, they often struggle with capturing complex patterns and contextual relationships in language. Traditional models such as SVMs or logistic regression rely heavily on feature engineering, which can be time-consuming and domain-dependent (Alkomah & Ma, 2022). Moreover, they are limited in handling large-scale unstructured data and fail to generalize well on noisy or informal text like social media posts (Djuric et al., 2015). Intended on the contrary, deep learning models are far superior when it comes to understanding context, semantics, and long-range relationships and automatically finding high-level characteristics in raw data (Rizos, Hemker, & Schuller, 2019).

This paper is used to deploy a real-time model for detecting hate speech based on the T5 transformer. The model is trained on labeled data of tweets and optimized to label hatred speech. In our proposed model, the classification of the tweet is considered in three different classes, namely Class 0, which gives normal tweets, Class 1, which gives offensive tweets and Class 2, which gives hate tweets. It is combined and crafted at a dense classification head and based on PyTorch. Besides the use of deep learning, we also utilized feature engineering on the metadata of tweets, like the length of tweets, the word count and the use of punctuations and special characters. These characteristics were examined to guide the model in dif-