Data Glacier Internship

Data Science: Natural Language Processing

Hate Speech Detection: Transformer Based Neural Networks

Siobhan Hwang, USA
siobhan.hwang@outlook.com

Andrew O'Drain, USA
andrewodrain@outlook.com

September 18, 2023

Contents

1 Problem Statement **2**

2 Business Understanding **2**
2.1 Stakeholders 2
2.2 Why Hate Speech Needs to be Addressed 3
2.3 Goals of Project 3
2.4 Benefits of Project 3
2.5 Drawbacks of Project 3

3 Project Life-cycle & Timeline **4**

1 Problem Statement

Hate speech has a plethora of definitions, all of which create different implications in different spheres. Most definitions agree that some kind of harm is being done to an individual or group directly or indirectly from the ideas or particular words used. The harm caused can originate from instigating violence, intimidation, discrimination, inciting hatred or vilification, degradation, and insult toward a specific group based on particular features they may share such as race, nationality, religion, gender, sexual orientation, ethnicity, etc.

Focusing solely on the hate speech encountered online, the first hate-centered forum website, Stormfront, laid the groundwork for white nationalist social networking in 1995. Explicit slurs and symbolism were banned in order to foster a more subtle mode of discussing ideals that aligned with the Ku Klux Klan and Neo-Nazism. According to a piece published by Southern Poverty Law Center, violent hate crimes, including almost 100 murders, have been committed by members of Stormfront.

Now almost 30 years later, though Stormfront was removed in 2017, online hate speech and its offline consequences are still prevalent in global discussion. While flat-out censorship can be tricky in countries like the USA, of which one of its pillars is free speech, many government organizations and brands are in favor of preventing and removing hate speech in order to garner a safer and more attractive environment for online users to express themselves. For example, Meta, an American conglomerate, hopes to create a space where people do not feel attacked and can connect freely with others by following a three-tier system of categorizing and handling posts that go against their Community Standards in the form of hate speech.

By developing ways to identify and eliminate hate speech online, a task of scale nearly impossible for human moderators to handle on platforms as big as those under the Meta umbrella, for example, the very real and sometimes deadly offline implications of such discourse can be prevented.

Our project aims to support this cause by leveraging deep learning via transformer model, trained on 31,962 Tweets from the platform *Twitter*, now known as *X,* to accurately identify hate speech in the form of sentiment analysis.

2 Business Understanding

In the ever-evolving landscape of online communication, the issue of hate speech stands as a significant challenge. To address this complex problem, it is essential to first establish a clear understanding of the key aspects and stakeholders involved.

2.1 Stakeholders

Online Platforms: Social media giants, forums, and online communities have a direct interest in detecting and mitigating hate speech on their platforms. They seek to provide safe and inclusive spaces for users to express themselves.

Government Organizations: Governments worldwide have a stake in regulating online content, especially when it pertains to hate speech, which can incite violence, discrimination, and social unrest.

Brands and Advertisers: Companies that advertise on online platforms want their products and services to be associated with content that aligns with their values. Hate speech can harm a brand's image and impact advertising strategies.

Users: The most important stakeholders are the users themselves. They deserve a safe online environment where they can freely express their views without fear of harassment or harm.

2.2 Why Hate Speech Needs to be Addressed

Violence and Harm: Hate speech can lead to real-world violence, discrimination, and harm against targeted groups.

Social Division: It can deepen societal divides, promoting hostility among different communities.

Undermining Free Expression: Excessive hate speech can discourage individuals from participating in online discussions, inhibiting the free exchange of ideas.

2.3 Goals of Project

Accurate Identification: Develop a deep learning model, based on transformer architecture, to accurately identify hate speech in online content, specifically Twitter data.

Mitigation and Prevention: Contribute to the prevention of hate speech by enabling timely content moderation on digital platforms.

Enhance User Experience: Create a safer and more inclusive online environment where users can engage in discussions without fear of harassment.

2.4 Benefits of Project

Safer Online Spaces: The project aims to make online platforms safer for users, fostering a more positive digital experience.

Compliance: It helps online platforms comply with regulations and community standards related to hate speech.

Brand Reputation: Brands and advertisers can benefit from a cleaner online environment, protecting their reputation.

2.5 Drawbacks of Project

False Positives: Overly aggressive hate speech detection may lead to false positives, potentially restricting free expression.

Algorithmic Bias: There's a risk of algorithmic bias, where the model may inadvertently target certain groups unfairly.

Political Implications: Instantiating hate speech detection may en-flame political tensions within many countries.

3 Project Life-cycle & Timeline

Exploratory Data Analysis (EDA): This crucial phase involves data collection, preprocessing, and initial analysis. The estimated date of completion is within a week

Feature Extraction: Identifying relevant features and transforming data for model training, expected to be completed within two weeks.

Model Selection: A deep learning model based on transformer architecture will be developed and fine-tuned over a period of a week or two.

Validation and Testing: Rigorous testing and validation will be conducted for another week.

Deployment: The deployment phase, including integration with online platforms, will take approximately a week.

Monitoring and Maintenance: Ongoing monitoring and model updates will be part of the project to adapt to evolving trends and emerging threats.

References

[1] Anderson, Luvell and Barnes, Michael. "Hate Speech." In: *The Stanford Encyclopedia of Philosophy*. Ed. Edward N. Zalta. 2022. https://plato.stanford.edu/archives/spr2022/entries/hate-speech/.

[2] Southern Poverty Law Center. *Stormfront*. https://www.splcenter.org/fighting-hate/extremist-files/group/stormfront.

[3] Transparency Center. 2023. *Hate Speech Policy*. Meta. https://transparency.fb.com/policies/community-standards/hate-speech/.