NLPeace

Name	Student ID
Fatima El Fouladi	40108832
Jeff Wilgus	29206345
Mira Aji	40041473
Raya Maria Lahoud	40129965
Anum Siddiqui	40129811
Adam Qamar	40175980
Shabia Saeed	40154081
David Lemme	40157270
Nelly Bozorgzad	40189770
Joshua-James Nantel-Ouimet	40131733

Project Description

Social media continues to play an ever increasing role in people's lives. In addition to facilitating connections between people over vast geographical, demographic, and cultural distances, a number of negative outcomes have been observed. Chief among them are a preponderance of disinformation, hate speech, and other harmful language. A major challenge in combating this issue is the great effort required to monitor the massive amount of activity on these platforms to identify instances of deleterious content. Previous solutions surely attempted to automate the process to the highest degree possible, however, a large staff of dedicated personnel is still necessary to carry out much of the work involved. This is clearly demonstrated by the recent changes at X (formerly Twitter), where a concomitant increase in fake news stories and harassment followed the dismissal of a significant portion of the moderation team [1].

Emerging artificial intelligence (AI) technologies, particularly in natural language processing (NLP) and large language models exemplified by ChatGPT, offer new ways for overcoming the previously discussed problem. Specifically for automating the identification and removal of problematic language in a large corpus. To the extent that this may be done without human intervention, a social media platform using the proposed solution should enjoy lower costs and a greater freedom to innovate and dedicate resources towards activities that add value for their users.

We propose to build a social media platform that offers a similar experience to X, but that utilizes NLP to mitigate online hate speech and spam. We intend to use familiarity to minimize the mental burden of prospective users to adapt to the platform. However, the distinguishing feature of the product will be its relative freedom from offensive content, which will enable users to have a peaceful and pleasant experience . Following are the primary features we aim to implement:

- NLP feature to moderate hate speech and spam.
- Make word posts of a predefined number of characters, including videos, and photos.
- Reporting posts.
- Reposting a post and adding commentary.
- Liking posts (and potentially disliking or saving posts).
- Leaving comments on posts.
- Follow or block other users.
- Direct Messaging, sending pictures, gifs, etc.
- Bot detection.
- Recommend users or posts depending on the hashtags a user frequently uses.

Risk

In order to ensure a successful delivery of the <NAME > application, we must consider the potential risks associated with the project and identify mitigation strategies to reduce their impact.

- **Inaccurate Moderation:**
  Harmless content might be inaccurately flagged as hate speech or might fail to detect actual harmful speech which could lead to user frustration.
  Risk Level: High
  Mitigation Strategy: Regularly update and test the system using diverse examples, including false positives and negatives. Additionally, provide users with the ability to appeal content decisions. We will also add admin users that can perform QA on the moderation.

- **Avoidance Techniques:**
  Users might try to bypass moderation by using coded language or through media to convey harmful speech.
  Risk Level: Medium
  Mitigation Strategy: Update the moderation rules regularly to adapt to new avoidance techniques. Additionally, implement keyword filtering to detect evasive content.

- **User Backlash and Public Relations Issues:**
  Users may perceive the moderation efforts as either too strict or not strict enough, leading to public backlash and negative publicity for the app. Moreover, controversial content moderation decisions could spark outrage on social media.
  Risk Level: Medium
  Mitigation Strategy: Clearly communicate the app's moderation policies and guidelines to users and the public to manage expectations and minimize misunderstandings. Develop a crisis communication plan to manage potential PR issues and controversies effectively.

- **Potential technological obsolescence**
  The AI algorithms and models are quickly evolving; newer, more advanced models and techniques emerge rapidly. This can result in the app's AI moderation system becoming less effective over time.
  Risk Level: High
  Mitigation Strategy: Adopt an agile development approach to allow quick iterations and updates. This flexibility will let us integrate newer AI models and techniques as they become available. Our AI models will continuously

learn as our user-base grows and as more content is posted, it will be added into our NLP pipeline to train our models.

- **Security Vulnerabilities:**
  Bad coding practices can lead to security vulnerabilities in the code of the application. When this happens attackers or even regular users may be able to gain access to privileged data or functions.
  Risk Level: High
  Mitigation Strategy: All developers should brush up on good coding practices and all code should be peer reviewed.

Competition

Search terms: Social Networks , Twitter Competitors , Social Networking platforms, Facebook Competitors, X Competitors, What is Facebook?, What is Twitter?, Bluesky, Mastodon, What is Mastodon?, What is Bluesky?
Number of pages examined: see Annex A

Competitor 1: Mastodon

Mastodon offers the option for users to make posts, share pictures and connect with other like-minded users. However, Mastodon is a decentralized social network. Having a decentralized social network can bring many benefits such as lower server downtime and enhanced user privacy, but it comes with its own set of potential drawbacks. Such networks bring more freedom to the user, which in of itself can augment the risk for hate speech. Mastodon was confronted with this problem in 2019, when Gab, a social network known for its widely far-right userbase, migrated to it. The CEO himself stated *"You have to understand it's not actually possible to do anything platform-wide because it's decentralized,* [...] *I don't have the control*." [2] For this reason, our product would not be decentralized, which would make it possible for us to implement an effective measure against hate speech. Not only will we be able to monitor what users share on our network, but also ensure that no one will be able to use our product to share harmful content.

Competitor 2: Bluesky

Bluesky is similar to X (formally Twitter) where users can post photos, add posts as well as share other people's posts. However, its limitations include the inability to currently upload videos and doesn't have direct messaging [3]. Bluesky is also a decentralized social network which gives users more control over their social media interaction. Although decentralized networks have many benefits, they can also have negative outcomes such as the spread of hate speech. Our product will not be decentralized and it will allow users to upload videos and send direct messages to each other.

Competitor 3: X

X is a well-known platform used worldwide for real-time micro-blogging, keeping its users informed and educated about people, information, and news. X enables its users to exchange messages and post photos, videos, links, and text. Similar to X, our product will not be decentralized, making it possible to control and limit hate speech. X moderates harmful speech by relying on reports made by its users on the harmful content which will then be reviewed by X before taking the necessary actions. On the other hand, our product will use natural language processing (NLP) to moderate and counter hate speech. The use of NLP to counter hate speech makes the process automatic and efficient in terms of reducing the workload and stress on moderators, as well as the amount of time required for an action to be taken.

Description of Customer and Company

Describe the company in max 1 paragraph. Provide a link to the company website.

SwiftConnect is a company that builds access control infrastructure. SwiftConnect powers physical access credentials and permissions in real-time, anywhere. Their software can interface with one's Apple Wallet. They build software to manage access and space inside buildings. They do this by connecting third party access management systems together in order to make conversion to their software seamless.

Describe the customer's expertise in max 1 paragraph (you can also attach a CV or link to the customer's profile.)

Victor Deleau has been working at SwiftConnect for nearly one year. He has work experience as a backend developer and as an AI developer. More details can be found on his linkedin profile provided below.
https://www.linkedin.com/in/victordeleau/

Describe why the customer is interested in max 1 paragraph.

The customer is interested because the company is looking into using NLP for their product and has tasked the customer with determining if this technology can be useful for their product. Therefore, this product would be used as research to determine the validity of this technology for the company's purposes.

Annex A : Links Examined for Competitors

1. 5 Social Networks That Are Alternatives to Twitter
2. The Top 10 Social Media Sites & Platforms
3. Who Are Facebook's (Meta's) Main Competitors?
4. 19 New Social Media Apps & Platforms in 2023
5. The 15 Biggest Social Media Sites and Apps [2023] - Dreamgrow
6. What Is Mastodon and Why Is Everyone Going There?
7. 8 Twitter/X alternatives for if you want to get off Elon Musk's wild ride
8. Twitter alternative SPILL takes off online
9. How Does Mastodon Work And How Do You Get Started?.
10. 8 Twitter/X alternatives for if you want to get off Elon Musk's wild ride
11. What is Bluesky? Everything to know about the app trying to replace Twitter | TechCrunch
12. Bluesky, the Twitter replacement that AOC and Chrissy Teigen just joined, explained - Vox
13. What Bluesky Tells Us About the Future of Social Media | The New YorkerDetecting and Monitoring Hate Speech in Twitter - PMC (nih.gov)
14. What Is Bluesky? The Twitter Alternative With Promising Networking Technology - Decrypt
15. https://www.forbes.com/sites/mattnovak/2023/07/01/bluesky-suffers-outage-as-people-flee-twitter-in-wake-of-new-restrictions/?sh=5086f29a73bc
16. Detecting and Monitoring Hate Speech in Twitter - PMC (nih.gov)
17. Twitter | Company Overview & News (forbes.com)
18. New user FAQ (twitter.com)
19. Twitter's policy on hateful conduct | Twitter Help
20. Countering Online Hate Speech: An NLP Perspective – arXiv Vanity (arxiv-vanity.com)
21. X seemed to throttle some competitors and news sites for more than a week - The Verge
22. Tired of Elon Musk? Here are the Twitter alternatives you should know about | CNN Business
23. The eight top social media sites you should prioritize in 2023
24. Facebook: What is Facebook?
25. The Now: What is Reddit?

References

[1] S. Frenkel and K. Conger, “Hate speech’s rise on Twitter is unprecedented, researchers find,” The New York Times, https://www.nytimes.com/2022/12/02/technology/twitter-hate-speech.html (accessed Sep. 8, 2023).

[2] A. Robertson, “How the biggest decentralized social network is dealing with its Nazi problem,” The Verge, https://www.theverge.com/2019/7/12/20691957/mastodon-decentralized-social-network-gab-migration-fediverse-app-blocking (accessed Sep. 2, 2023).

[3] M. Novak, “Bluesky suffers outage as people flee Twitter in wake of new restrictions,” Forbes, 01-Jul-2023. [Online]. Available: https://www.forbes.com/sites/mattnovak/2023/07/01/bluesky-suffers-outage-as-people-flee-twitter-in-wake-of-new-restrictions/?sh=5086f29a73bc. [accessed: 09-Sep-2023].