DATA417

# The Trustworthy Data Scientist

# Group Project

# Contents

[1 Contents 2](#_Toc166699892)

[2 Table of Figures 2](#_Toc166699893)

[3 Tables 2](#_Toc166699894)

[4 Introduction 4](#_Toc166699895)

[5 Ethical Issues of the existing method 5](#_Toc166699896)

[5.1 Content Moderators 5](#_Toc166699897)

[5.2 Ethical Issue 5](#_Toc166699898)

[5.3 Real world example 6](#_Toc166699899)

[5.3.1 Phuc Dat Bich Incident 6](#_Toc166699900)

[6 Methodology 7](#_Toc166699901)

[7 Algorithm 8](#_Toc166699902)

[7.1 Grab a topic 8](#_Toc166699903)

[7.2 Run match () 8](#_Toc166699904)

[7.3 Highlight () 8](#_Toc166699905)

[7.4 Display the highlighted text with definitions 8](#_Toc166699906)

[7.5 Run sentiment analysis to get the scores 8](#_Toc166699907)

# Table of Figures

[Figure 1: Screenshot of Phuc Dat Bich Incident 6](#_Toc166699908)

# Introduction

# Ethical Issues of the existing method

## Content Moderators

Content moderators play a crucial role in maintaining the integrity and safety of social media platforms. They work behind the scenes to filter and review content, ensuring it adheres to the platform's specific policies and community standards. These moderators handle violations flagged by AI algorithms, which are designed to detect potential breaches of content guidelines, as well as issues reported directly by users. The method combines automated technologies and human judgement, with moderators making crucial decisions that software alone may be incapable of.

## Ethical Issue

Since content moderators frequently handle content from a variety of diverse sources, they are crucial for ensuring the accuracy of decisions made on social media platforms. Typically, moderators are allocated to review content in languages they are proficient in, which helps in understanding the subtleties and cultural contexts unique to those languages. Major languages such as English, Spanish, French, and Mandarin, which are widely spoken across different regions, can have variations in meaning and interpretation. This complexity requires moderators to have a deep understanding of not just the language but also the specific dialects and cultural complexities associated with each region where the language is spoken.

Due to the complexities of language and culture, content moderators can sometimes do mistakes during their review of reported content. This may lead to the wrongful removal of content that does not breach any policies, or the oversight of content that does violate community standards. Both situations present ethical issues, as they directly impact users. Incorrectly removing content can suppress legitimate expression and unfairly penalize individuals, while failing to remove harmful content can expose users to potentially damaging material, undermining the platform’s safety and integrity. These scenarios highlight the significance of precise and fair content moderation processes, and this research tries to modify the present moderation system to prevent such mistakes.

Moderators would make mistakes due to the following reasons,

* Language usage and slang can vary significantly across regions.
* Subtleties in language can lead to misinterpretation of content.
* Importance of understanding local dialects and expressions.
* Moderators’ decisions may be influenced by their own cultural context.

## Real world example

### Phuc Dat Bich Incident

Phuc Dat Bich, a 23-year-old of Vietnamese descent living in Australia, faced repeated issues with Facebook’s identity verification system. His name, which is pronounced approximately as "Phoo Da Bic" in Vietnamese, was flagged by Facebook’s automated systems as being potentially fake or humorous due to its pronunciation and spelling in English.

Figure 1: Screenshot of Phuc Dat Bich Incident

After having his Facebook account repeatedly shut down due to what the platform's administrators suspected to be a stage name, Phuc Dat Bich took the step of posting a photograph of his passport online. This was intended as proof that he was indeed using his real and legal name. His passport photo clearly showed his name, aiming to clear up any misunderstanding with Facebook’s administration and to address the broader issue of cultural insensitivity and misunderstanding in automated systems (Dow, 2015).

The figure 1 shows a Facebook post which was shared by the particular user, and this is a perfect example of misinterpretation of words by the content moderators.

# Methodology

# Algorithm

The new algorithm can be described in the following steps.

## Grab a topic

The system starts by identifying a topic of interest. This could be triggered by trending topics, user complaints, or predefined filters targeting specific subjects. The identification process might utilize keyword extraction technologies to scan large volumes of content efficiently.

## Run match ()

Once a topic is identified, the system scans available content sources to find relevant mentions of this topic. This step employs search algorithms and pattern matching to sift through text, focusing on content that includes the identified keywords or topics.

## **Highlight ()**

Relevant content segments are highlighted for further review. This function not only marks the areas of interest but also prepares the content for deeper analysis, such as sentiment assessment or context evaluation. This step is crucial for focusing the moderation effort on specific content sections that are most pertinent to the topic.

## Display the highlighted text with definitions

Highlighted content is then presented to the content moderators, accompanied by definitions and additional context. This could involve linking to external knowledge bases, dictionaries, or internal guidelines to help moderators understand the nuances of the topic and the reasons it was flagged.

## Run sentiment analysis to get the scores

Sentiment analysis methods determine the emotional tone of highlighted information. This phase is crucial for understanding how subjects are discussed, is the sentiment mostly negative, favourable, or neutral? Such insights are essential in determining the possible benefit or damage of the information.

The process involves calculating the total negative sentiments from content and setting a predefined threshold. If the cumulative negative sentiment exceeds this threshold, an alert is triggered to notify content moderators that there has been a significant accumulation of negative sentiments.

# References

Dow, A. (2015, November 21). *Phuc Dat Bich, the Australian man with a name so awkward no one believes him*. Retrieved from The Sunday Morning Herald: https://www.smh.com.au/technology/phuc-dat-bich-the-australian-man-with-a-name-so-awkward-no-one-believes-him-20151120-gl4egu.html

OpenAI. (2024, March 01). *ChatGTP*. Retrieved from ChatGTP: https://chat.openai.com/

QuillBot. (2024, March 01). *QuillBot*. Retrieved from QuillBot: https://quillbot.com/