**Unstructured Data Management  
Group Project Report**

**Topic: Reddit Text Classification: Categorizing Posts and Discovering Insights**

**Table of Contents**

[**Executive Summary** 3](#_Toc147510357)

[**Project background** 4](#_Toc147510358)

[**Data source and descriptions** 5](#_Toc147510359)

[**Data Analytics** 6](#_Toc147510360)

[**Data findings and visulisations and descritions** 11](#_Toc147510361)

[**Project Summary** 14](#_Toc147510362)

[**Implications** 14](#_Toc147510363)

[**Conclusion** 15](#_Toc147510364)

**Executive Summary**

* Project Overview:

This project addresses the challenge of categorizing Reddit posts into specific subreddits, helping users find communities aligned with their interests. We aim to develop a text classification model that predicts the relevant subreddit for a post based on its title and description, benefiting users, and assisting moderators.

* Problem Statement:

Reddit users often struggle to discover subreddits matching their interests amidst the platform's diversity. This project seeks to streamline this process, enhancing the user experience.

* Project Significance:

1. User Benefit: By suggesting relevant subreddits, our model empowers users to access content that aligns with their preferences, improving user engagement and satisfaction.
2. Moderator Assistance: The model aids moderators in identifying irrelevant posts and potential spammers, enhancing subreddit quality and safety.

* Key Objectives:

1. Develop a robust text classification model.
2. Provide personalized subreddit suggestions to users.
3. Assist moderators in content management.
4. Impact:
5. Improved user engagement and content relevance.
6. Enhanced subreddit quality through spam detection.

# **Project background**

In an era defined by digital communities and the sharing of ideas, Reddit stands as one of the most influential and diverse social media platforms. With over 430 million monthly active users and millions of discussions spanning countless topics, Reddit is a hub for discourse, entertainment, and knowledge-sharing. However, the platform's vastness also presents a challenge: how can users efficiently find subreddits that cater to their specific interests?

* **The Challenge:** Reddit's diverse ecosystem encompasses an extensive array of subreddits, each dedicated to a particular topic, hobby, or interest. From movie buffs seeking the latest film discussions to food enthusiasts hunting for mouthwatering recipes, users are often overwhelmed by the sheer volume of choices. Finding the right subreddit manually can be akin to searching for a needle in a digital haystack, leading to frustration and potential disengagement.
* **Project Rationale:** The motivation behind this project is twofold. First, it addresses a pervasive issue faced by Reddit users daily: the difficulty of discovering subreddits that cater to their specific passions. By simplifying this process, we aim to make Reddit more accessible, engaging, and personalized. Second, our project acknowledges the vital role of subreddit moderators.
* **Project Scope:** Our project is focused on developing a robust text classification model that can predict the most suitable subreddit for a given post based on its title and description. This model will be a valuable addition to Reddit's user interface, offering users personalized subreddit suggestions and supporting moderators in content management.

In a digital landscape teeming with information, the ability to efficiently connect users with relevant content is paramount. Our project's objective is to create a solution that bridges the gap between Reddit's vast array of subreddits and individual user interests. By doing so, we aspire to enhance the Reddit experience for all, empowering users to discover their digital communities more easily and assisting moderators in maintaining vibrant, spam-free subreddits.

**Data source and descriptions**

**Data Source**: Reddit API

The Reddit API endpoint https://oauth.reddit.com/r/{subreddit}/new is used to retrieve a list of the newest posts from a specific subreddit. The exact data fields returned by this API endpoint can include a variety of information about each post. However, the specific fields provided in the API response can vary based on the endpoint's parameters and the type of data you request.

**Description of Data Fields:**

A screenshot of a computer

Description automatically generated

* title: The title of the Reddit post.
* author: The username of the Reddit user who posted the post.
* created\_utc: The time the post was created, in Unix time.
* permalink: The URL of the Reddit post.
* score: The total number of upvotes and downvotes the post has received.
* ups: The number of upvotes the post has received.
* downs: The number of downvotes the post has received.
* num\_comments: The number of comments the post has received.
* selftext: The text content of the post.

# **Data Analytics**

1. **Data Collection**

Our data collection process relies on the utilization of the Python Request library to interact with the Reddit API. Specifically, we make dynamic requests to the endpoint https://oauth.reddit.com/r/{subreddit}/new to collect data from various subreddits. This approach ensures adaptability and scalability in data gathering, allowing us to target different subreddits as needed. The choice of the /new endpoint guarantees access to the latest posts, providing up-to-the-minute content for our analysis. Moreover, we've implemented an automated data retrieval process, enabling us to effortlessly acquire a substantial dataset of 25,000 posts efficiently.

We collected 2500 posts from each of the following subreddits using the Reddit API. We also collected ~2500 posts from the /r/all subreddit, which we designated as a catch-all category for posts that don’t belong to any of the other subreddits.

1. Movies
2. Food
3. Technology
4. News
5. Gaming
6. Science
7. Sports
8. Music
9. Books
10. Others
11. **Data Preprocessing**

By **eliminating noise, handling missing values, and transforming data** into a suitable format, data preprocessing enhances the quality and reliability of input data, ultimately improving model accuracy and performance.

* Merged Title and Description as a single text, Convert to lower case, Stop word removal., Lemmatization, Symbols removal

A screenshot of a computer

Description automatically generated

1. **Feature Extraction**

A screenshot of a computer code

Description automatically generated

1. **Model Development**

**(1): Naive Bayes (NB) classifier**

Train accuracy: 0.9826034793041392

Test accuracy: 0.9648140743702519

Classification report:

precision recall f1-score support

books 0.95 1.00 0.97 484

food 0.96 1.00 0.98 525

gaming 0.96 0.97 0.97 501

movies 0.97 0.98 0.98 504

music 0.97 0.98 0.97 484

others 0.94 0.75 0.83 485

science 0.99 0.99 0.99 516

sports 0.97 0.99 0.98 478

technology 0.96 0.99 0.98 493

worldnews 0.97 0.99 0.98 532

accuracy 0.96 5002

macro avg 0.96 0.96 0.96 5002

weighted avg 0.96 0.96 0.96 5002

A diagram of a graph

Description automatically generated with medium confidence

**(2): DecisionTreeClassifier**

Train accuracy: 0.9091681663667267

Test accuracy: 0.9044382247101159

Classification report: precision recall f1-score support

books 1.00 0.97 0.98 484

food 1.00 1.00 1.00 525

gaming 0.99 0.90 0.94 501

movies 0.99 0.98 0.98 504

music 0.98 0.92 0.95 484

others 0.50 0.93 0.66 485

science 0.99 0.88 0.94 516

sports 1.00 0.81 0.90 478

technology 0.99 0.79 0.88 493

worldnews 1.00 0.86 0.92 532

accuracy 0.90 5002

macro avg 0.94 0.90 0.91 5002

weighted avg 0.94 0.90 0.92 5002

A diagram of a graph

Description automatically generated with medium confidence

**(3): RandomForestClassifier**

Train accuracy: 0.945710857828434

Test accuracy: 0.9438224710115953

Classification report:

precision recall f1-score support

books 1.00 1.00 1.00 484

food 0.99 1.00 1.00 525

gaming 0.99 0.95 0.97 501

movies 1.00 0.99 0.99 504

music 1.00 0.94 0.97 484

others 0.64 0.97 0.77 485

science 1.00 0.93 0.96 516

sports 1.00 0.88 0.94 478

technology 1.00 0.86 0.92 493

worldnews 0.99 0.93 0.96 532

accuracy 0.94 5002

macro avg 0.96 0.94 0.95 5002

weighted avg 0.96 0.94 0.95 5002

A diagram of a graph

Description automatically generated with medium confidence

# **Data findings and visulisations and descritions**

Here are the word clouds related to 3 subreddits.

**A close-up of words

Description automatically generated**

It shows the most common words used in posts on that subreddit, with the size of each word representing its frequency. The largest words in the word cloud are "outrage," "Ukraine," "Russia," and "Deny Visa." These words suggest that the subreddit is heavily focused on current events, particularly the war in Ukraine and the rise of China as a global power as climate change, the global economy, and public health.

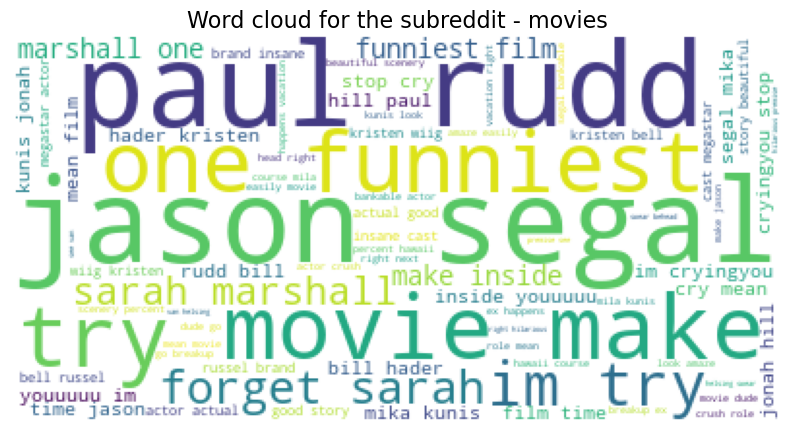
Overall, the word cloud provides a good overview of the topics that are most discussed on the r/worldnews subreddit. It shows that the subreddit is a valuable resource for anyone who wants to stay informed about the latest news and trends from around the world.

**A close-up of words

Description automatically generated**

It shows the most common words used in posts on that subreddit, with the size of each word representing its frequency. The largest words in the word cloud are "sport," "team," "game," and "player." These words suggest that the subreddit is focused on all aspects of sports, including teams, games, and players.

Other prominent words in the word cloud include "league," "NBA," "NFL," and "MLB." These words suggest that the subreddit is particularly interested in professional sports leagues in the United States.

****

It shows the most common words used in posts on that subreddit, with the size of each word representing its frequency. The largest words in the word cloud are "movie", "film", "actor", "actress", and "director". These words suggest that the subreddit is focused on all aspects of movies, including the films themselves, the people who make them, and the people who act in them. Other prominent words in the word cloud include "genre", "review", "release", "box office", and "trailer". These words suggest that the subreddit is also interested in the latest news and trends in the movie industry, as well as the critical and commercial success of individual films. The word cloud also suggests that the subreddit is interested in both classic and new movies. However, there is a particular emphasis on new releases and upcoming movies.

A screenshot of a computer

Description automatically generated

The visualization shows the top 20 most popular terms for each of the 10 topics that were identified in the dataset. The topics are arranged in a circle, with the most similar topics being placed closer together. The size of each topic circle is proportional to the number of posts that are assigned to that topic. The color of each topic circle represents the topic's saliency, which is a measure of how important the topic is to the overall dataset.

The top 20 most popular terms for each topic are listed within the corresponding topic circle. The size of each term is proportional to its frequency within the topic. The color of each term represents the term's relevance to the topic, which is a measure of how well the term discriminates between the topic and other topics.

A screenshot of a computer screen

Description automatically generated

The x-axis of the chart shows the two columns, and the y-axis shows the average word count.

The chart shows that the average word count in the title column is 34.42 words, while the average word count in the description column is 14.15 words. This means that the titles of the posts in the dataset are generally longer than the descriptions. The chart also shows that there is a wider range of word counts in the title column than in the description column. This suggests that the authors of the posts are more careful about the length of their titles than they are about the length of their descriptions.

# **Project Summary**

In the pursuit of enhancing the Reddit user experience through subreddit prediction, this project has made significant strides in addressing the challenges of content discovery and community management on the platform. Through the development of a text classification model and the analysis of extensive data, we have laid the foundation for a more user-centric and efficient Reddit ecosystem.

* Future Directions: As we conclude this project, it is essential to acknowledge that innovation and adaptation are ongoing processes. Future developments and enhancements could include:
* Refinement of the Model: Continuous improvement of the text classification model to achieve even higher accuracy and adaptability to evolving language trends.
* Integration with Reddit: Collaborative efforts with Reddit to integrate this model directly into the platform's interface, offering real-time suggestions to users and assisting moderators in content curation.
* User Feedback Mechanisms: Implementing mechanisms for users to provide feedback on subreddit suggestions, contributing to the model's self-improvement.

In summary, this project represents a step forward in making Reddit more user-centric and efficient. By addressing the fundamental issue of subreddit discovery, we aim to foster a thriving Reddit community where users can easily connect with like-minded individuals and moderators can manage content more effectively. The journey of innovation and user empowerment continues, with the aspiration of shaping a more vibrant digital landscape.

# **Implications**

* Enhanced User Experience: The project's text classification model improves the Reddit user experience by suggesting personalized subreddits, increasing user engagement and satisfaction.
* Moderator Support: The model aids moderators by automating post classification, helping maintain higher subreddit quality, and ensuring a safer environment.
* Scalable and Adaptable: The project's data collection process is scalable, allowing the model to include more subreddits as Reddit grows. Future improvements include model refinement and integration with Reddit's platform.

# **Conclusion**

In conclusion, the Reddit Text Classification Project addresses the fundamental challenge of improving content discovery and community management on the Reddit platform. Its implications span from user-centric benefits to moderator support, and the insights gained from data analysis provide valuable information for both users and content creators. This project sets the stage for ongoing innovation and improvement within the Reddit community and serves as a model for applying data analytics and machine learning to enhance user experiences on social media platforms.