Text data mining: a case study of Reddit

Piotr Sokołowski[1] **Marcin Kosiba**[2] **Michał Pałucki**[3]

[1,2,3]*AGH University of Krakow*
Department of Applied Computer Science
Al. Mickiewicza 30, 30-059 Kraków, Poland
[1]*psokolowski@student.agh.edu.pl,* [2]*mkosiba@student.agh.edu.pl,*
[3]*palucki@student.agh.edu.pl*

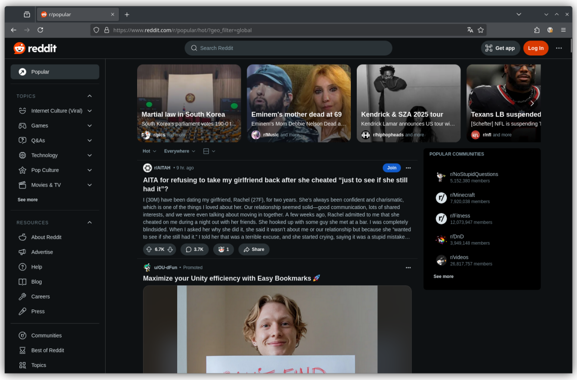
Abstract. *The article describes the progress of a project on text data mining performed by software engineering students. The article presents information about the source of the text data, how it was acquired and cleaned. The process of selecting classifiers for categorizing posts is described, and the results of different models are compared.*

Keywords: *data mining, reddit, classification, clustering, scikit-learn, tensorflow, machine learning*

1. Introduction

Reddit is a social media site that can be described as a collection of forums. Users share their questions, thoughts, and content, while other users can upvote or downvote them. The most popular posts are displayed on the front page, which is the first thing a user sees when they visit the site. Reddit is divided into subreddits [2], which are forums dedicated to specific topics. Users can subscribe to subreddits they are interested in, and posts from these subreddits will appear on their front page, as shown in Fig. 1.

As of September 2024, Reddit has more than 97 million daily active users across more than 100,000 communities [**reddit2024**]. The service contains more than 16 billion posts and comments, which have been posted since June 2005, when the service was founded.



Martial law in South Korea
South Korea's parliament votes 190-0 t...
r/pics and more

Eminem's mother dead at 69
Eminem's Mom Debbie Nelson Dead a...
r/Music and more

Kendrick & SZA 2025 tour
Kendrick Lamar announces US tour wi...
r/hiphopheads and more

Texans LB suspended
[Schefter] NFL is suspending T
r/nfl and more

r/AITAH • 9 hr. ago

AITA for refusing to take my girlfriend back after she cheated “just to see if she still had it”?

I (30M) have been dating my girlfriend, Rachel (27F), for two years. She’s always been confident and charismatic, which is one of the things I loved about her. Our relationship seemed solid—good communication, lots of shared interests, and we were even talking about moving in together. A few weeks ago, Rachel admitted to me that she cheated on me during a night out with her friends. She hooked up with some guy she met at a bar. I was completely blindsided. When I asked her why she did it, she said it wasn’t about me or our relationship but because she “wanted to see if she still had it.” I told her that was a terrible excuse, and she started crying, saying it was a stupid mistake…

u/OU-dFun • Promoted

Maximize your Unity efficiency with Easy Bookmarks 🚀

r/NoStupidQuestions	5,152,380 members
r/Minecraft	7,920,038 members
r/Fitness	12,073,947 members
r/DnD	3,949,148 members
r/videos	26,817,757 members

Figure 1: The appearance of Reddit’s home page, with posts popular among Redditors displayed

2. State of the Art and Related Work

Analyzing social media posts is a frequent topic in academic research. Users’ posts often provide information about world events [**cheong2011social**], as well as their emotions and opinions [1]. Many academic papers focus on analyzing Twitter posts, but Reddit is also a popular source of data.

Reddit is one of the few social networks with such a large number of users, while also offering a free API. In addition, the platform is known for its high diversity of topics, which allows for the analysis of different areas of life. Also important is the fact that Reddit’s structure allows for the categorization of posts, which relieves researchers of the need to manually label data.

To query the API, we intended to use the PRAW[1] library [**praw2024**]. It provides easy access to Reddit’s APIs, as well as convenient data processing tools. Additionally, PRAW handles authentication, which is essential for retrieving data from Reddit. Unfortunately, API limitations proved to be an obstacle during the initial development of the project. When querying to read posts and comments, we were limited to a maximum of 100 results per request. Since we needed at least 5,000 posts to create a representative dataset, we had to find another source.

Nearly 10 years ago, on July 3, 2015, Jason Baumgartner, under the nickname

[1] Python Reddit API Wrapper

“u/Stuck_In_the_Matrix,” made a post in the “r/datasets” subreddit stating that he had collected all publicly available comments from Reddit for research purposes [**datasets2015**]. He collected about 250 GiB of data, which includes 1.7 billion posts and comments published on the platform from October 2007 to May 2015. The posts were grouped by month and year and then saved in JSONL files[2]. A day later, with the help of the community, the collection was made available as a torrent.

References

[1] Anurag P. Jain and Vijay D. Katkar. “Sentiments analysis of Twitter data using data mining”. In: *2015 International Conference on Information Processing (ICIP)*. IEEE, Dec. 2015, pp. 807–810. DOI: 10.1109/infop.2015.7489492. URL: http://dx.doi.org/10.1109/infop.2015.7489492.

[2] Alexey N. Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. “The Anatomy of Reddit: An Overview of Academic Research”. In: *Dynamics On and Of Complex Networks III*. Springer International Publishing, 2019, pp. 183–204. ISBN: 9783030146832. DOI: 10.1007/978-3-030-14683-2_9. URL: http://dx.doi.org/10.1007/978-3-030-14683-2_9.

[2] The file format allows multiple JSON objects to be saved in a single file, with one line corresponding to one JSON document