Sentiment Analysis of User Comments on Machine Learning Posts: A Data-driven Approach Using R*

Joseph Chung

Apr 7, 2023

Abstract

In this study, we present a comprehensive data analysis paper that utilize a sentiment model in R to analyze user comments from a dataset containing post_ids and comments related to the topic of Machine Learning. We aim to investigate sentiment polarity of user comments and its distribution across different posts, identifying trends and patterns that may provide valuable insights into the Machine Learning community's perception and engagement.

Table of contents

*Code and data are available at: https://github.com/UtopianYoungChung/Sentiment-Analysis-Using-R.git

List of Figures

List of Tables

1 Introduction

The rapid growth of online forums and communities discussing various topics, including Machine Learning(ML), has led to an abundance of user-generated content in the form of text data. Analyzing this data can reveal valuable insights into user perceptions, opinions, and sentiment towards specific subjects. Sentiment analysis, or opinion mining, is a natural language processing technique used to determine the sentiment expressed in a piece of text, such as positive, negative, or neutral. This paper focuses on employing sentiment analysis techniques to understand the sentiment polarity of user comments related to the topic of Machine Learning.

1.1 Background and motivation

Machine Learning has become a popular field of study and research due to its vast applications in various industries. As a result, online communities have emerged where people discuss ML topics, share knowledge, and express their opinions. Analyzing the sentiment of these user comments can provide insights into the perception of the ML community, identify common issues or concerns, and uncover emerging trends. This understanding could be valuable for researchers, educators, and industry professionals who aim to improve their work based on community feedback.

1.2 Research questions

This study aims to address the following research questions:

1. What is the overall sentiment polarity (positive, negative, or neutral) of user comments related to Machine Learning?
2. How is the sentiment polarity distributed across different posts in the dataset?
3. Are there any correlations between post attributes (e.g., post length, engagement) and sentiment polarity?
4. Can any trends or patterns in sentiment polarity be identified over time or in in relation to specific subtopics within Machine Learning?

1.3 Scope and limitations

The scope of this study is limited to the analysis of a dataset containing post IDs and comments related to the topic of Machine Learning. The results and interpretations are based solely on the data provided and may not be generalizable to other online communities or topics. The sentiment analysis model employed in this study is subject to the limitations inherent in natural language processing techniques, such as context ambiguity and handling of sarcasm or irony. Additionally, the study does not account for demographic information or biases present in the dataset, which could potentially influence the sentiment analysis results.

2 Literature Review

2.1 Sentiment analysis in online communities

Over the past decade, sentiment analysis has become increasingly popular as a tool to mine user opinions and emotions from online platforms, such as social media, blogs, and forums (**citePang?**). Numerous studies have focused on the application of sentiment analysis in various domains, such as politics (**citeTumasjan?**), finance (Bollen, J., Mao, H., & Zeng, X 2011), and consumer reviews (Dave, Kushal and Lawrence, Steve and Pennock, David M. 2003). These studies demonstrate the potential of sentiment analysis as a valuable technique for understanding user-generated content and informing decision-making processes.

2.2 Sentiment analysis techniques in R

R is a widely used programming language for statistical analysis and data visualization, with several libraries and packages available for natural language processing and sentiment analysis (R Core Team 2020). Some of the most popular sentiment analysis packages in R include tidytext (**citeSilge?**), syuzhet (**citeJockers?**), and sentiment (**citeRinker?**). These packages provide various algorithms and techniques for sentiment

analysis, such as lexicon-based approaches, machine learning models, and deep learning methods. Researchers have employed these techniques to analyze sentiment in diverse contexts, ranging from Twitter data (**citeWijeratne?**) to news articles (**citeKearney?**)

2.3 Applications of sentiment analysis in Machine Learning research

Sentiment analysis has been employed in Machine Learning research to understand user perceptions, opinions, and trends related to the field. For example, Aliza (Sarlan, Aliza and Nadam, Chayanit and Basri, Shuib 2014) analyzed the sentiment of tweets related to Machine Learning conferences, revealing insights into the community's responses to specific presentations and events. In another study, Wang (**citeWang?**) investigated the sentiment of user comments on popular Machine Learning platforms, such as GitHub and Stack Overflow, to identify trends and patterns in the community's engagement with specific algorithms and techniques. These studies highlight the potential of sentiment analysis to uncover valuable insights into the Machine Learning community and inform future research directions.

The literature demonstrates the growing interest in sentiment analysis as a valuable tool for understanding user-generated content in various domains. R has emerged as a popular platform sentiment analysis due to its extensive libraries and packages. However, there remains a need for more comprehensive studies that investigate the sentiment of user comments related to Machine Learning, particularly in the context of online communities where users share their knowledge and opinions on the topic. This study aims to address this gap by employing sentiment analysis techniques in R to analyze a dataset of post IDs and comments related to Machine Learning.

3 Data Collection and Preprocessing

This section describes the process of data collection, including web scraping and API requests, and the subsequent preprocessing steps applied to the dataset to prepare it for sentiment analysis.

3.1 Data Collection

The dataset used in this study was collected from Reddit, a popular online platform where users can share and discuss a wide range of topics. The data was obtained by performing web scraping and API requests on the topics of Machine Learning, Data Science, and Artificial Intelligence over the last 10 years (2013 to 2023) using Python. The Reddit API (https://www.reddit.com/dev/api/) was utilized to access and retrieve relevant posts and associated comments from the specified subreddits (e.g., r/MachineLearning, r/datascience, r/artificial).

To perform the web scraping and API requests, the Python package PRAW (Python Reddit API Wrapper) was used, which provides a convenient interface to interact with the Reddit API. The data collection process involved:

1. Authenticating with the Reddit API using a registered application's credentials.
2. Querying the API for posts related to the specified topics (Machine Learning, Data Science, and Artificial Intelligence) within the targeted subreddits.
3. Retrieving the following attributes:
    1. post_id
    2. subreddit
    3. created_utc
    4. selftext

5. post_url
6. post_title
7. link_flair_text
8. score
9. num_coments
10. upvote_ratio

4. Combining the collected data into a structured dataset for further analysis.

Table 1: 12 observations from dataset of Reddit API

ID	Subreddit	Score	Upvote
gh1dj9	MachineLearning	7803	0.99
oisl3e	datascience	3197	0.98
ia2aob	MachineLearning	1941	0.99
w6kj9y	MachineLearning	1728	0.99
i5yres	MachineLearning	1204	0.98
7b7ghl	MachineLearning	1054	0.95
cgwvds	datascience	824	0.98
jboe91	datascience	721	0.91
mdldtt	MachineLearning	639	0.95
n62qhn	MachineLearning	584	0.99
mhh5zu	datascience	527	0.96
fedkoa	datascience	502	0.97

3.2 Data cleaning and preprocessing

place holder

3.2.1 Handling missing values

place holder

3.2.2 Text normalization

place holder

3.2.3 Tokenization

place holder

3.2.4 Stopword removal

place holder

3.2.5 Stemming/Lemmatization

place holder

4 Methodology

place holder

4.1 Sentiment analysis model selection

place holder

4.2 Model training and validation

place holder

4.3 Model implementation in R

place holder

4.4 Evaluation metrics

place holder

5 Results

place holder

5.1 Sentiment polarity distribution

place holder

5.2 Correlation between post attributes and sentiment

place holder

5.3 Identification of trends and patterns

place holder

5.4 Comparison with previous studies

place holder

6 Discussion

place holder

6.1 Interpretation of results

place holder

6.2 Implications for Machine Learning community

place holder

6.3 Limitations of the study

place holder

6.4 Future research directions

place holder

7 Conclusion

place holder

References

Bollen, J., Mao, H., & Zeng, X. 2011. "Twitter Mood Predicts the Stock Market." *Journal of Computer Science* 2 (1): 1–8. https://doi.org/https://doi.org/10.1016/j.jocs.2010.12.007.

Dave, Kushal and Lawrence, Steve and Pennock, David M. 2003. "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews." In *Proceedings of the 12th International Conference on World Wide Web*, 519–28. WWW '03. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/775152.775226.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sarlan, Aliza and Nadam, Chayanit and Basri, Shuib. 2014. "Twitter Sentiment Analysis." In *Proceedings of the 6th International Conference on Information Technology and Multimedia*, 212–16. https://doi.org/10.1109/ICIMU.2014.7066632.