

# The Topic-Sentiment Analysis and Relevance System (TSAR)

Sai Muktevi, Hasnah Said, Preston Stringham, Andrew Zhou

## OUR TEAM

Our team is composed of 4 incredible UW MSDS students, Sai Muktevi, Hasnah Said, Preston Stringham, and Andrew Zhou. Our backgrounds include software development, mathematics, and information technology.



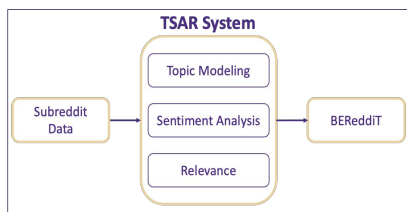
From left to right: Preston Stringham, Andrew Zhou, Hasnah Said, Sai Muktevi

## INTRODUCTION

The TSAR system is a data pipeline that utilizes Natural Language Processing (NLP) techniques to identify conflicting viewpoints on Reddit.com. Reddit.com is an online platform where people can come together to share news, discuss different topics, comment on others posts or comments. Reddit is made up of more than two million communities, each community is known as a subreddit and is dedicated to a specific topic.

To uncover conflicting viewpoints in various subreddits, TSAR leverages the transformer-based machine learning technique Bidirectional Encoder Representations from Transformers (BERT) for topic modeling, sentiment analysis, and relevance scores to measure the similarity between posts and their comments. The findings from our TSAR system will be displayed on a visualization dashboard: BEReDDiT.

## METHODOLOGY



### Dataset: Subreddit Corpus

Our dataset consists of 50 subreddits that were scraped using Asynchronous Python Reddit API Wrapper (AsyncPRAW) with the following configurations due to API limitations:

- > Latest posts categorized as 'hot' by reddit were used
- > Maximum of 1,000 posts per subreddit
- > Maximum of 100 comments per post
- > Maximum of 10,000 comments per subreddit

The raw data is passed into a preprocessor module to transform the data into a format that can be understood and analyzed by the TSAR system. The preprocessor module produces a subreddit dataframe, post dataframe, and a comments dataframe.

### Topic Modeling: What are posts talking about?

> Our topic modeling component uses BERTopic to generate topics for posts and comments. BERTopic is a technique that leverages transformers (BERT embeddings) and a custom class-based TF-IDF to create dense clusters to produce easily interpretable topics.

### Sentiment Analysis: Can we identify conflict with sentiment?

> The sentiment analysis component is implemented using the pretrained Hugging Face Transformers model: 'nlptown/bert-base-multilingual-uncased-sentiment'. This model is fine-tuned for sentiment analysis in six languages. It predicts the sentiment of each post's body text and comments as a rating between 1 (negative) and 5 (positive). Each submission is assigned a sentiment score for each rating and the rating with the highest score is deemed as the sentiment for that submission.

> We believe posts and comments sentiment set the tone for discussion and negative comment sentiment might indicate conflict.

### Relevance Analysis: Are the comments in the discussion relevant to the post?

> The final component in our TSAR system finds how relevant the comments are to the original post. We used cosine similarity and BERT embeddings to find the relevance scores for each comment.

## RESULTS

### BEReDDiT: TSAR Dashboard

Using the Python package Dash, we created an interactive dashboard visualization of the results generated by our pipeline. Our dashboard was deployed on Google Cloud and Heroku, you can visit our app by scanning the QR codes or visiting the URL below the QR codes on your web browser:



<https://beredditi.herokuapp.com/>



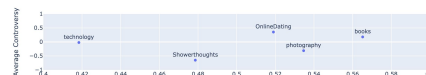
<https://beredditi.herokuapp.com/>



We used posts and comments sentiment scores as well as their relevance score to identify conflicting viewpoints in subreddits. We calculated the controversy score by finding how much the sentiment of a post comments deviates from the sentiment of the post itself:

$$\text{controversy score} = \text{comments sentiment score} - \text{post sentiment score}$$

To illustrate conflicting viewpoints, we produced the following interactive plots for a given subreddit on our dashboard:



This plot visualizes the average comment relevance compared to the average controversy for each subreddit. This gives insight into how focused and polarizing the discussions in each subreddit is likely to be.

## CONCLUSION

In this project, we have successfully implemented the Topic-Sentiment Analysis and Relevance (TSAR) system. TSAR is a complex data engineering pipeline that was developed using powerful Natural Language Processing techniques to identify conflicting viewpoints on the popular social media platform Reddit.com. We utilized BERT models from Hugging Face Transformers to implement our topic modeling, sentiment analysis, relevance scoring components. To better understand the conflicting viewpoints in subreddit, the results from our TSAR system is displayed on our webapp BEReDDiT.