Mental health support chatbot using GPT-2

Yuliya Selevich

Data Science Capstone Project

The problem

Can a chatbot built using GPT-2 provide sensible and relevant mental health advice?

The data

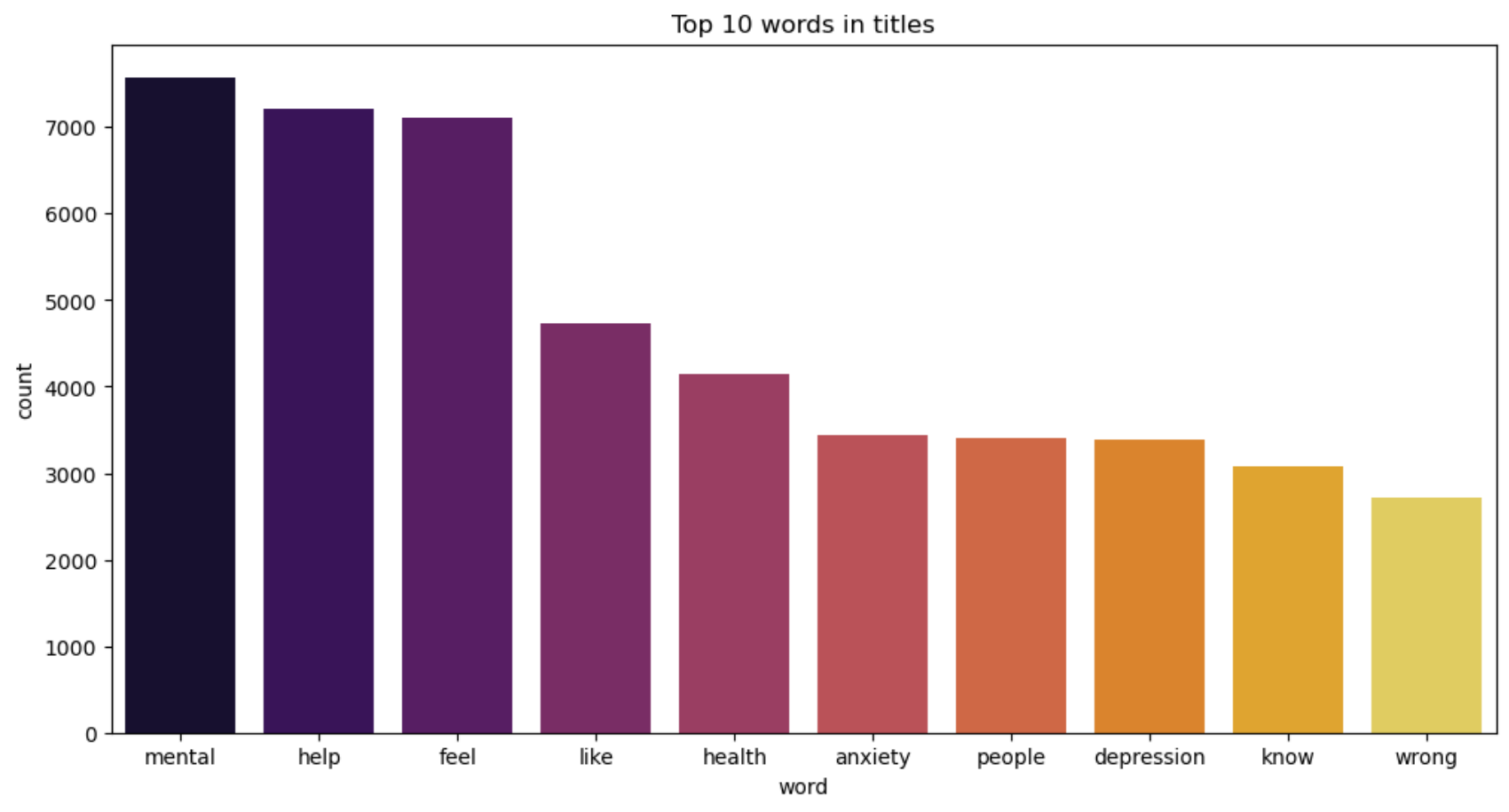
- Data was scraped from mentalhealth and mentalillness subreddits using PSAW (Python Pushshift API Wrapper).
- Scraped data were stored in two DataFrames containing submissions and comments.
- The DataFrame with submissions contained 415290 rows, whereas the DataFrame with comments contained 1556846 rows.

Data Cleaning and Processing

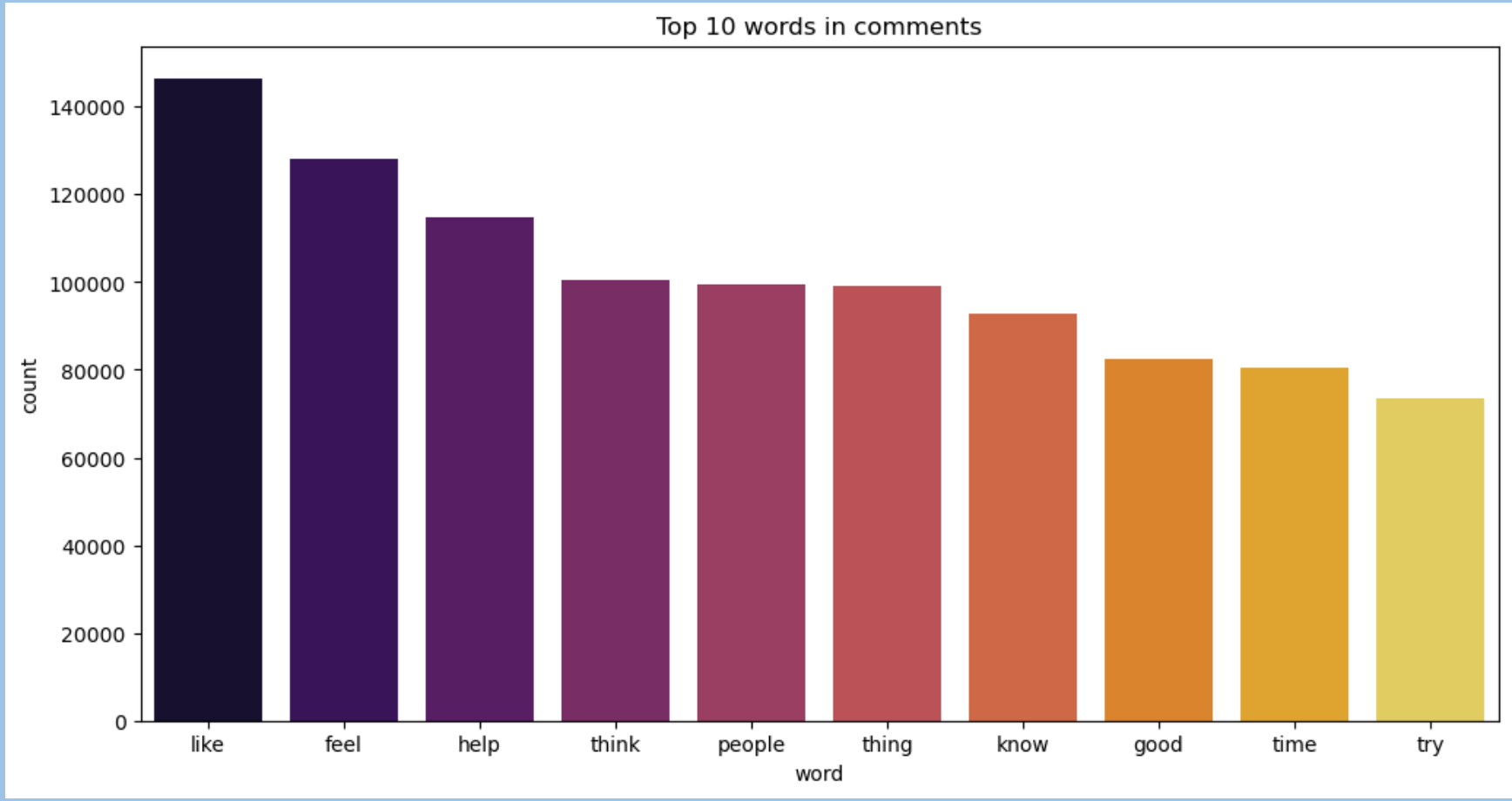
Some of the steps taken:

- Redundant columns were removed
- All text data were converted into lowercase.
- All missing values were removed.
- All duplicated comments, which mostly included automatic responses from a Reddit bot, were removed.
- The submissions were filtered out to remove those that didn’t contain a question.
- Some additional steps were taken to prepare the data for further analysis including contractions expansion, tokenization, and lemmatization, removal of stopwords, punctuation, digits, and links, as well as the removal of extra space between tokens.

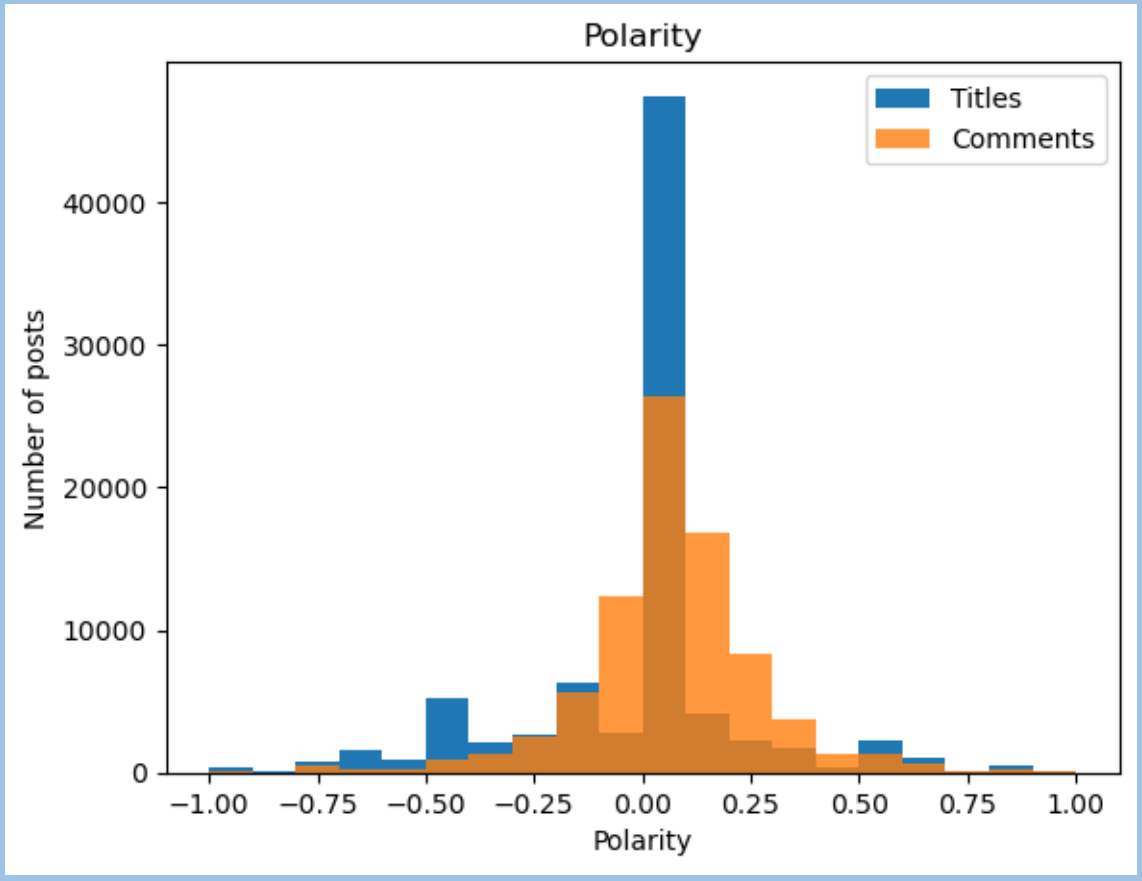
Exploratory Data Analysis



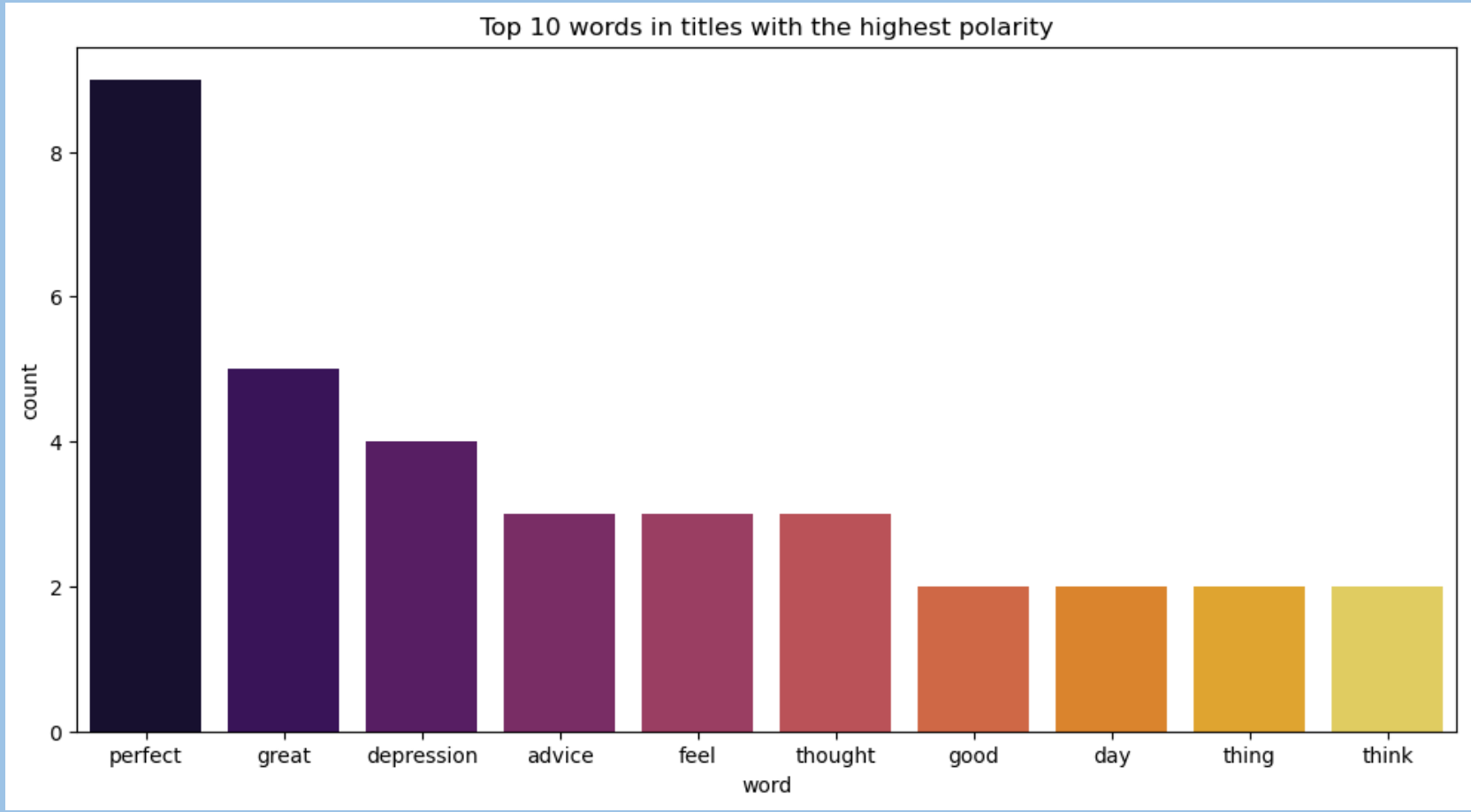
Exploratory Data Analysis



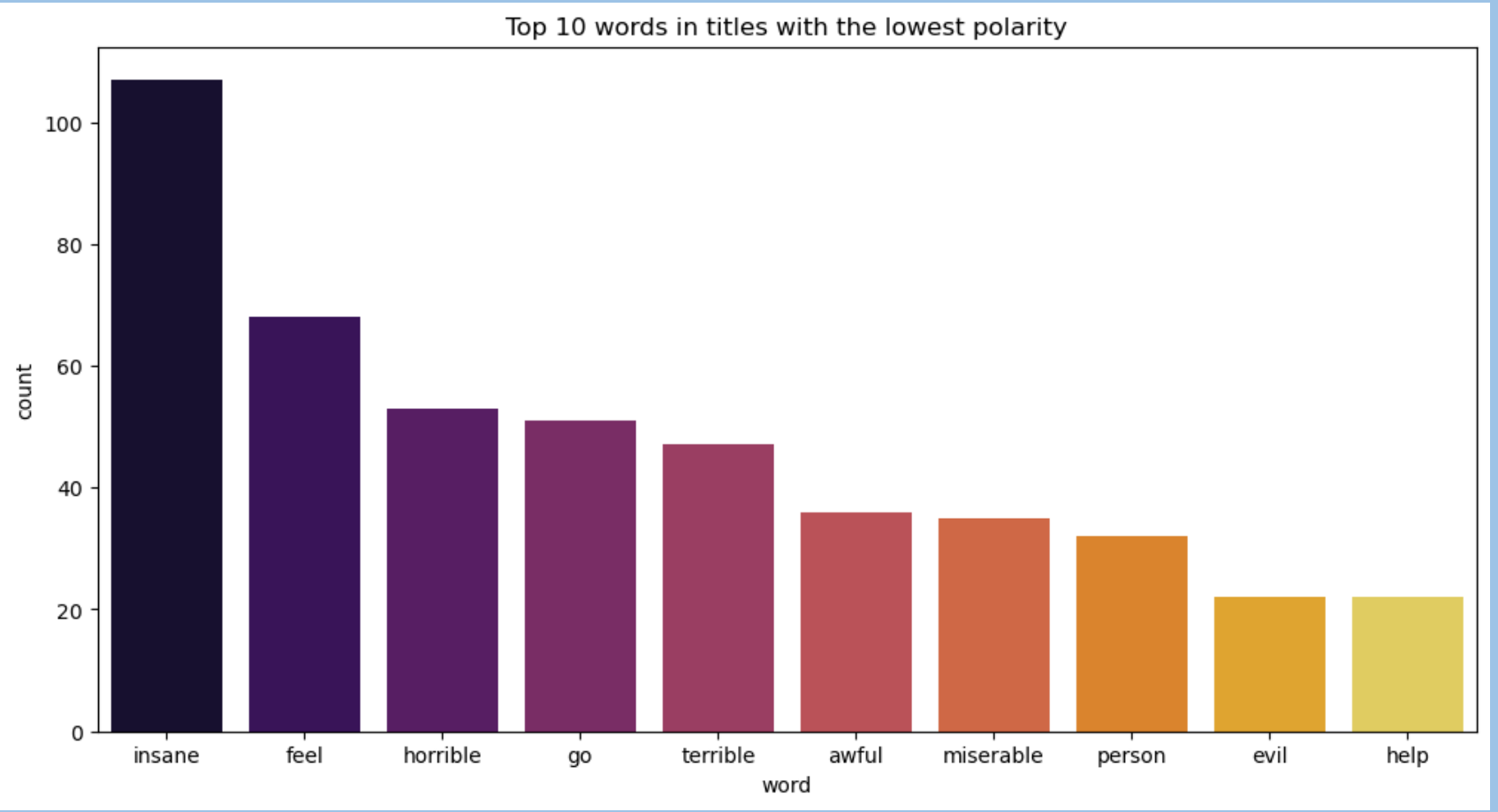
Exploratory Data Analysis



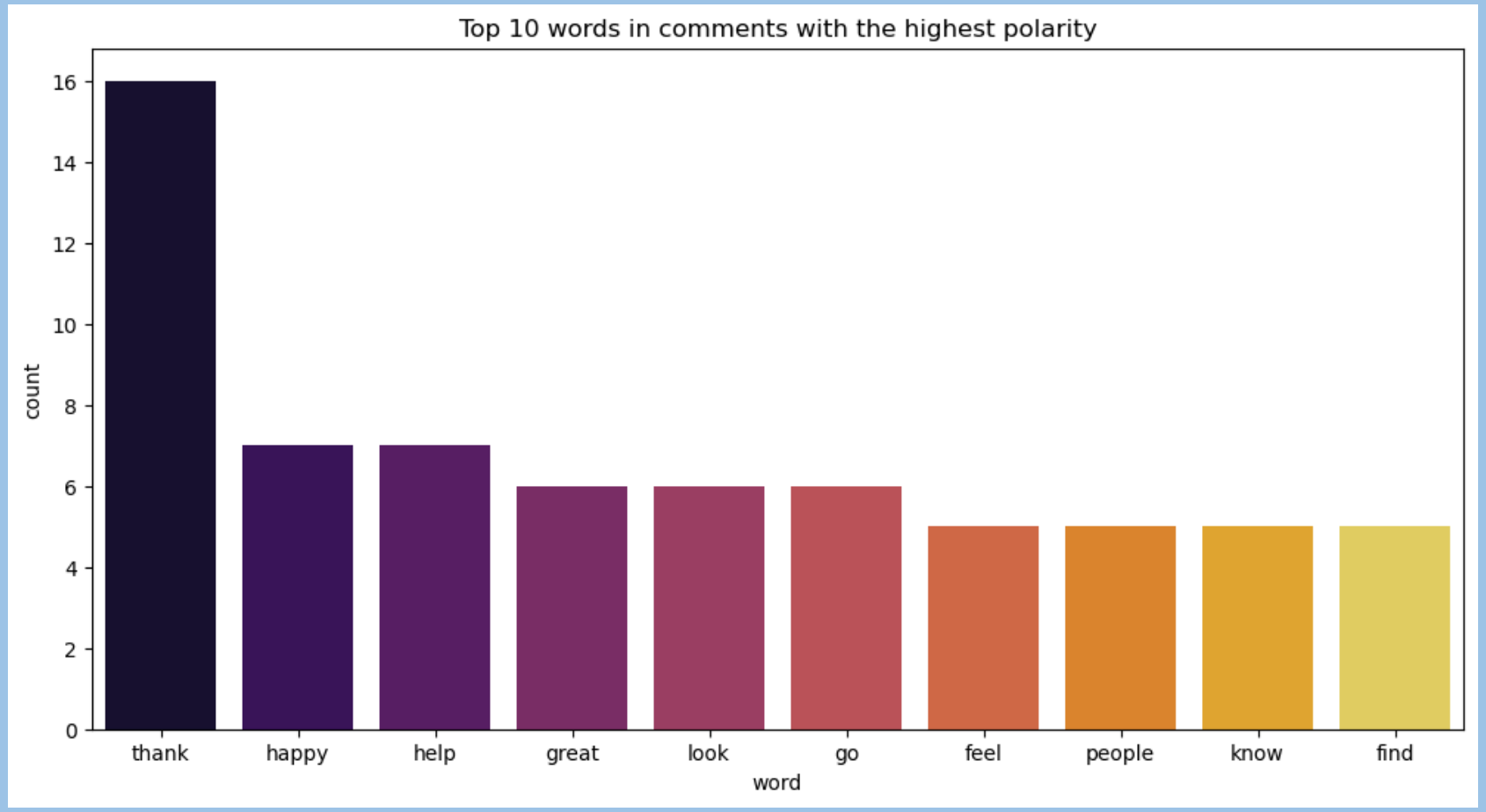
Exploratory Data Analysis



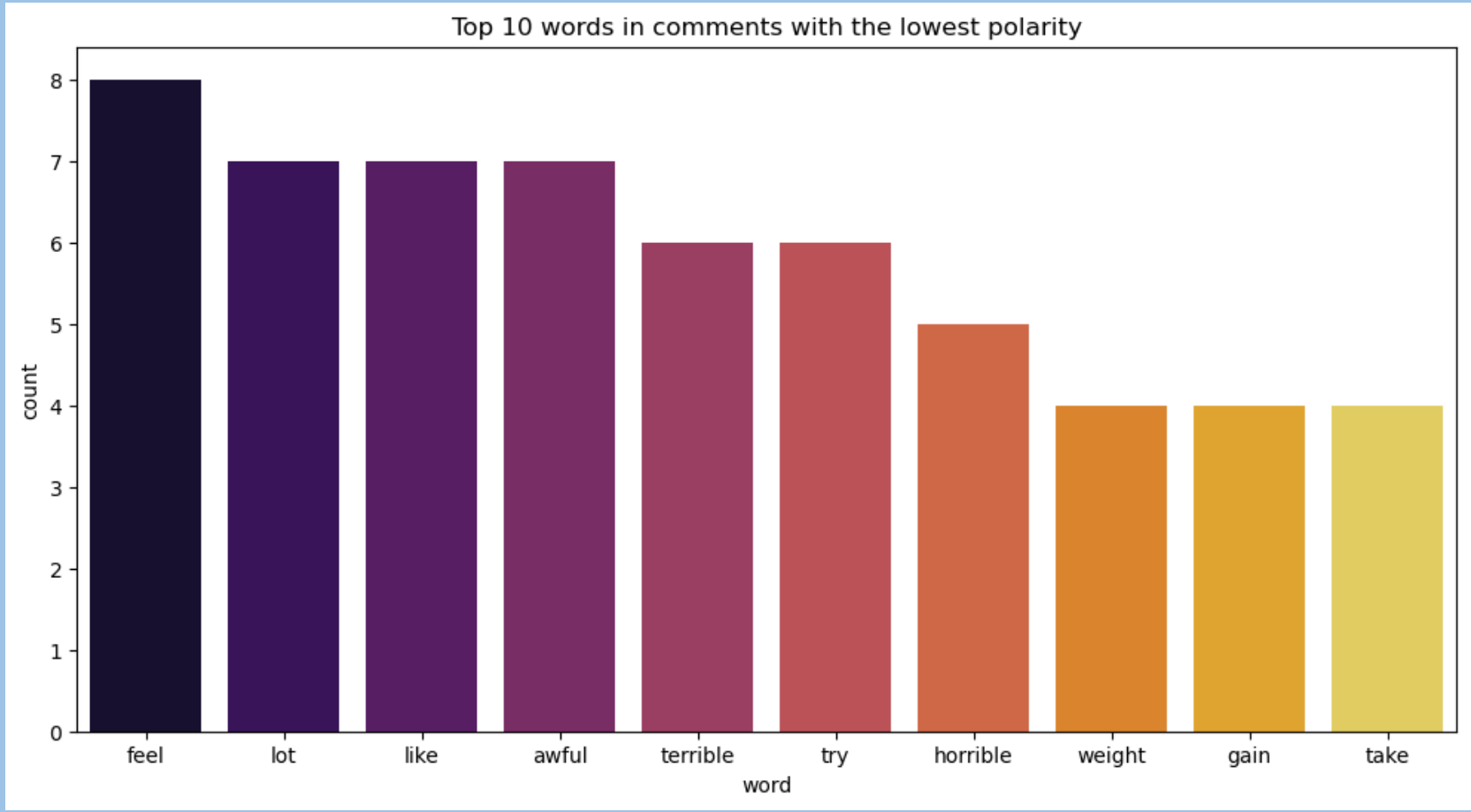
Exploratory Data Analysis



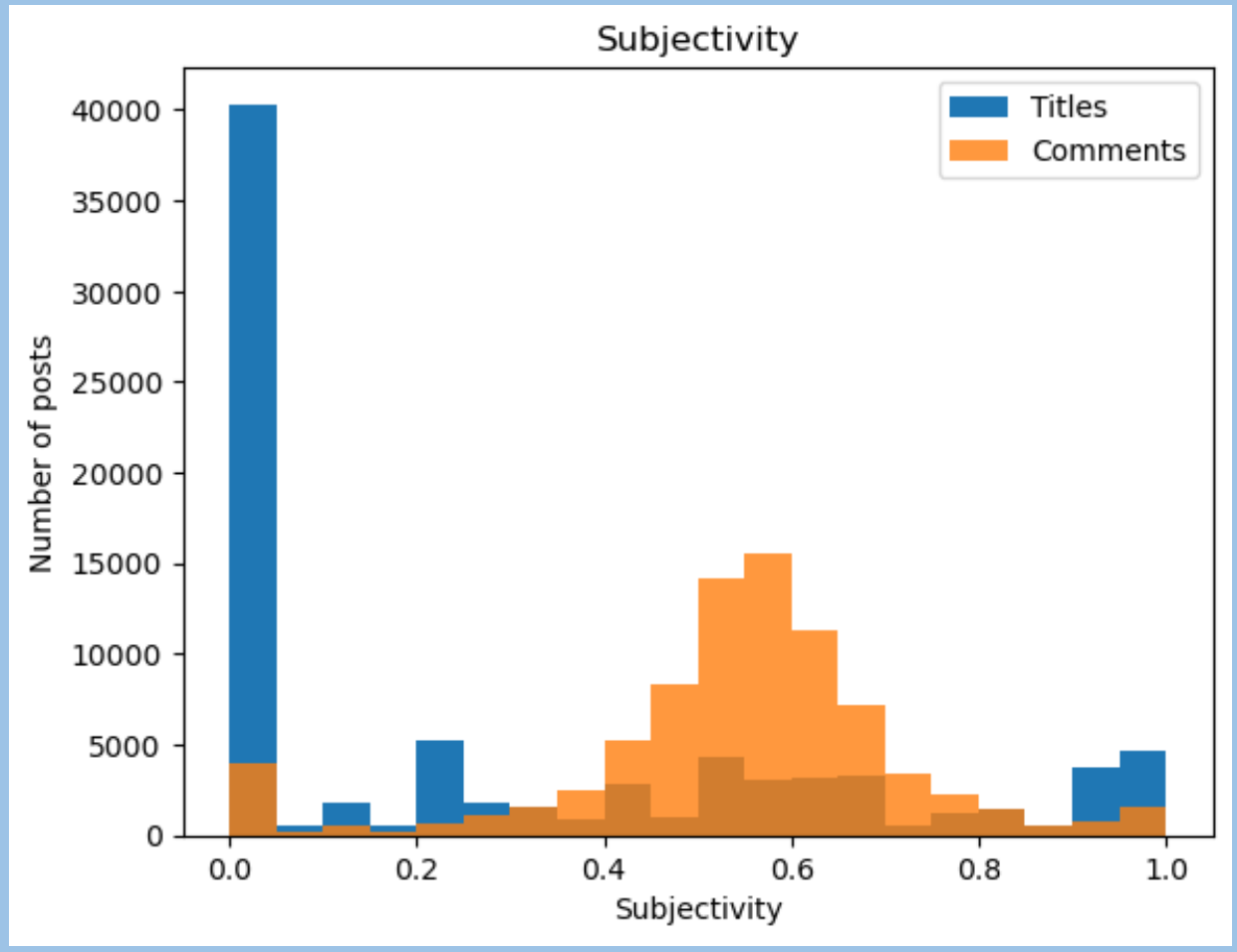
Exploratory Data Analysis



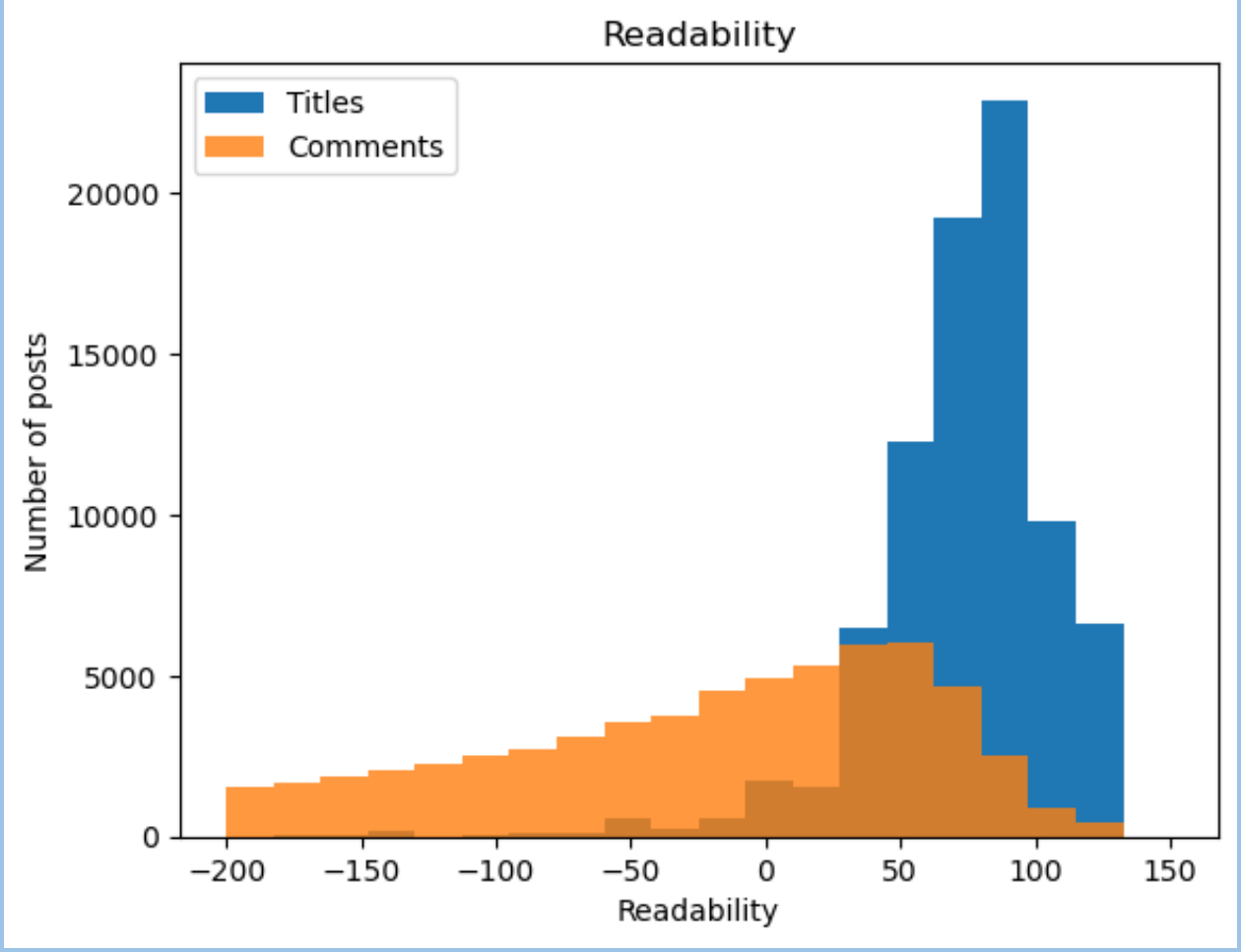
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Modeling

- The GPT-2 pre-trained model and gpt-2-simple package were used for this project.
- The “medium” 355M model with the size of 1.5GB on disk has been selected.
- Fine-tuning has been performed using finetune() method using 2100 steps and printed output samples every 500 steps.
- The fine-tuning took 39 hours and 28 minutes in total.

Results

Examples of a good (sensible) model's outputs:

1) 'The first steps are to accept the problem and then solve it. If you're having trouble, you can try talking about it with a mental health professional'.

2) 'I have trouble sleeping because of my depression. I get so tired that I don't even get up to leave the house.'

3) 'The main thing you want to do is take care of yourself and your health.'

4) 'I'm sorry you have to go through this. There's a lot of advice and information out there about how to deal with depression, what to do and what not to do. If you're not feeling good, you should go to a doctor for a checkup. They may be able to help you. Best of luck!'

Examples of a bad (nonsensical) model's outputs:

1) 'It just so happens my mom is also his grand-nephew.'

2) '…die peacefully due to an accident.'

3) 'I'm 23 and a self-taught neurosurgeon.'

4) 'My mental illnesses helped me a lot in a way that was very uncomfortable and painful.'

5) 'Just because you can't do something doesn't mean you can't do it: and that's okay.'

Conclusion

- Most of the time the model couldn’t generate a good output therefore it cannot be used for deployment without further improvements. Further improvements could include but are not limited to:
  - filtering out bad output (e.g. detecting duplicates in the text, using a BLEU score (Bilingual Evaluation Understudy))
  - fine-tuning the model with more data and steps.
  - performing more text cleaning keeping in mind its effect on the model’s generalization.
- The GPT-3 model may be trained instead as it showed improved performance, especially when given tasks in a specialized area or topic (like mental health).