Emetophobia Dataset Collection, NER, Sentiments, and Topic Analysis

Florin-Silviu Dinu - 412

Purpose

Purpose

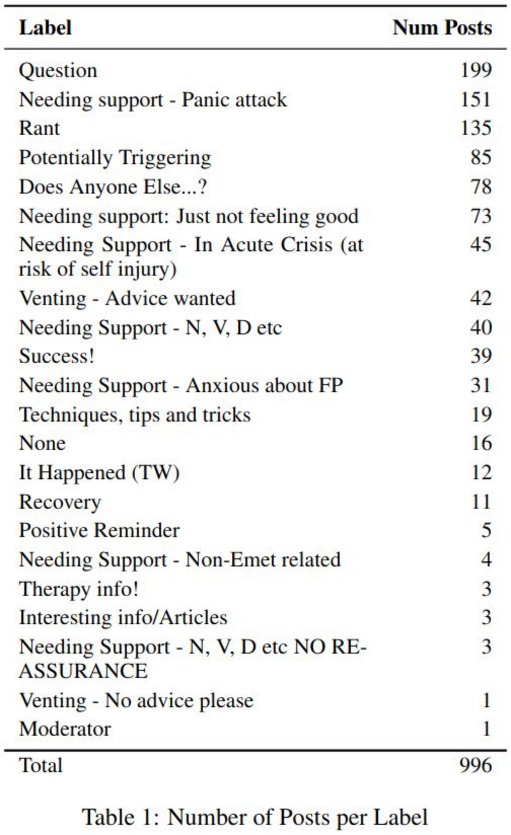
1. Build a dataset
2. Perform sentiment analysis
   a. With and without emojis
   b. Discuss emoji usage
3. Use NER to extract entities
   a. Using scispacy
   b. Using BioBERT
   c. Compare
4. Use RAG
   a. With the original texts
   b. With the original texts and scispacy UmlsEntityLinker’s data
   c. With the original texts and BioBERT’s categories
5. Topic classification
   a. Unsupervised
   b. Supervised

Dataset

Dataset

- Origin: r/emetophobia
- Scraping: Reddit API
- Preprocessing: simple rules to avoid unnecessary text
- Total number of scraped posts: 996
- Total number of kept posts: 986

Dataset



Label	Num Posts
Question	199
Needing support - Panic attack	151
Rant	135
Potentially Triggering	85
Does Anyone Else...?	78
Needing support: Just not feeling good	73
Needing Support - In Acute Crisis (at risk of self injury)	45
Venting - Advice wanted	42
Needing Support - N, V, D etc	40
Success!	39
Needing Support - Anxious about FP	31
Techniques, tips and tricks	19
None	16
It Happened (TW)	12
Recovery	11
Positive Reminder	5
Needing Support - Non-Emet related	4
Therapy info!	3
Interesting info/Articles	3
Needing Support - N, V, D etc NO RE-ASSURANCE	3
Venting - No advice please	1
Moderator	1
Total	996

Table 1: Number of Posts per Label

Sentiment Analysis

Sentiment Analysis

Overall

	Mean Emoji	StDev Emoji	Mean No-Emoji	StDev No-Emoji	p	t
Overall	−0.3116	0.7065	−0.3004	0.7079	0.0017	-3.156

Sentiment Analysis

Top 3 Labels

	Mean Emoji	StDev Emoji	Mean No-Emoji	StDev No-Emoji	p
Questions	−0.1914	0.6835	−0.1781	0.6863	0.07556
Needing support - Panic attack	−0.6125	0.5249	−0.5964	0.5355	0.01213
Rant	−0.4662	0.6447	−0.4505	0.6484	0.1846

Sentiment Analysis

Top 3 Emojis

	Number of apparitions	Mean Emoji	StDev Emoji	Mean No-Emoji	StDev No-Emoji	p
😭	91	−0.724	0.5776	−0.445	0.6646	>0.01
🫶	12	−0.6724	0.5362	−0.5995	0.61467	0.253
❤️	10	−0.2125	0.7806	−0.2123	0.7804	0.0839

NER + RAG

NER + RAG

- Scispacy
    - en_core_sci_md
    - UmlsEntityLinker
  - BioBERT

Top 3 Labels	scispacy	BioBERT
Questions	2581	1578
Needing support - Panic attack	2504	2101
Rant	2616	1657

NER + RAG

- LangChain
- Mistral Nemo
- Query: **“What are the top 5 most common medical or psychological ideas that the documents refer to? Make a numbered list. Respond with just the list, avoid other text”**

NER + RAG

scispacy	**BioBERT**	**RAG No NER**	**RAG scispacy**	**RAG BioBERT**
I-antigen	Sign_symptom	Exposure Therapy	Panic Attacks	Exposure Threapy
Anxiety	Activity	Emetophobia	Nausea/Vomiting	Emetophobia
Ilness (finding)	Detailed_description	OCD	PTSD	Panic Attacks
Daily	Time	Anti-Anxiety Medication	Zofran	Reassurance Seeking (Harmful in Con- text)
Nausea	Medication	CBT-Based Guide to Emetophobia	Mental Suffering/Distress	Anxiety Disorder

Topic Classification

Topic Classification

- Just posts with NERs
- Train-test split: 0.8-0.2
  - Scispacy: 400-101
  - BioBERT: 338-85
- Accepted labels (just top 5): ’Question’, ’Needing support - Panic attack’, ’Rant’, ’Potentially Triggering’, ’Does Anyone Else...?’, ’Needing support: Just not feeling good’

Topic Classification

- Unsupervised
    - DBSCAN
        - Skip-Gram
            - Vector size: 64
            - Window: 8 (scispacy), 10 (BioBERT)
        - Parameters
            - Iterations: 300
    - BERTopic
        - Sentence Transformers: sentence-transformers/all-MiniLM-L6-v2
        - Parameters
            - n-grams 1 and 2
            - topic size 2

Topic Classification

- Unsupervised (silhouette score)
    - DBSCAN
        - Scispacy: 0.6441
        - BioBERT: 0.6348
    - BERTopic
        - Scispacy: 0.0346 (5 topics + noise)
        - BioBERT: 0.0384 (5 topics + noise)

Topic Classification

- Supervised
  - BERTopic
    - Sentence Transformers: sentence-transformers/all-MiniLM-L6-v2
    - SVC
      - Kernel: RBF
  - XGBoost Classifier
    - Scispacy
      - Learning rate: 0.1
      - Maximum depth: 3
      - Estimators: 100
    - BioBERT
      - Learning rate: 0.2
      - Maximum depth: 3
      - Estimators: 200

Topic Classification

- Supervised (Accuracy)
  - BERTopic
    - Scispacy: 0.3564
    - BioBERT: 0.0941
  - XGBoost Classifier
    - Scispacy: 0.2277
    - BioBERT: 0.2

Conclusions

- Sentiments can be clearly delineated
- In-group emoji use can be easily studied
- An improvement in RAG applications can be found if using scispacy UmlsEntityLinker's canonical names and definition
- Usual NER models fare well on the dataset
- Labels can be delineated with a good enough silhouette score through DBSCAN but the rest of the models may lack data or may not fare well with the tried preprocessing methods

Thanks!

Florin-Silviu Dinu