Generalizable Natural Language Processing Framework for Migraine Reporting from Social Media

Yuting Guo[1], **Swati Rajwal**[1], Sahithi Lakamana[1], Chia-Chun Chiang[2], Paul C. Menell[2,] Adnan H. Shahid[2], Yi-Chieh Chen[3], Nikita Chhabra[4], Wan-Ju Chao[5], Chieh-Ju Chao[6], Todd J. Schwedt[4], Imon Banerjee[7,8] and Abeed Sarker[1]

[1] Biomedical Informatics, Emory University, GA, [2] Neurology dept., Mayo Clinic, MN, [3] Pharmacy Services Dept., Mayo Clinic, MN, [4] Neurology Dept., Mayo Clinic, AZ, [5] Psychology Dept., University of North Texas, TX, [6] Cardiology Dept., Mayo Clinic, MN, [7] Radiology Dept., Mayo Clinic, AZ, [8] CS Dept., Arizona State University, AZ

Abstract

- Migraine is a highly prevalent & disabling disorder.
- Our Tasks:
  - Verify existence of self-reported migraine chatter on social media
  - Develop supervised text classifier for detecting self-reported migraine posts
  - Assess the utility of social media for studying cohort-specific challenges.
- **Manually** annotated **5750** Twitter & **302** Reddit posts
- Trained & evaluated supervised ML models.
- Best system F_1 score: **0.90** (Twitter) & **0.93** (Reddit)
- Analysis show sentiment trends associated with migraine medications

Background

- EHRs capture health info, not daily **habits/interests:** captured in **patient-generated social media data**[1]
- Social media analysis using NLP improve patient-centered outcomes in cohort studies (breast cancer, substance use) [2,3]
- Studies[4] investigated migraine info. on social media.
- It's unclear if those methods are portable to other social media platforms

Table 1: Classification results of different transformer-based models.

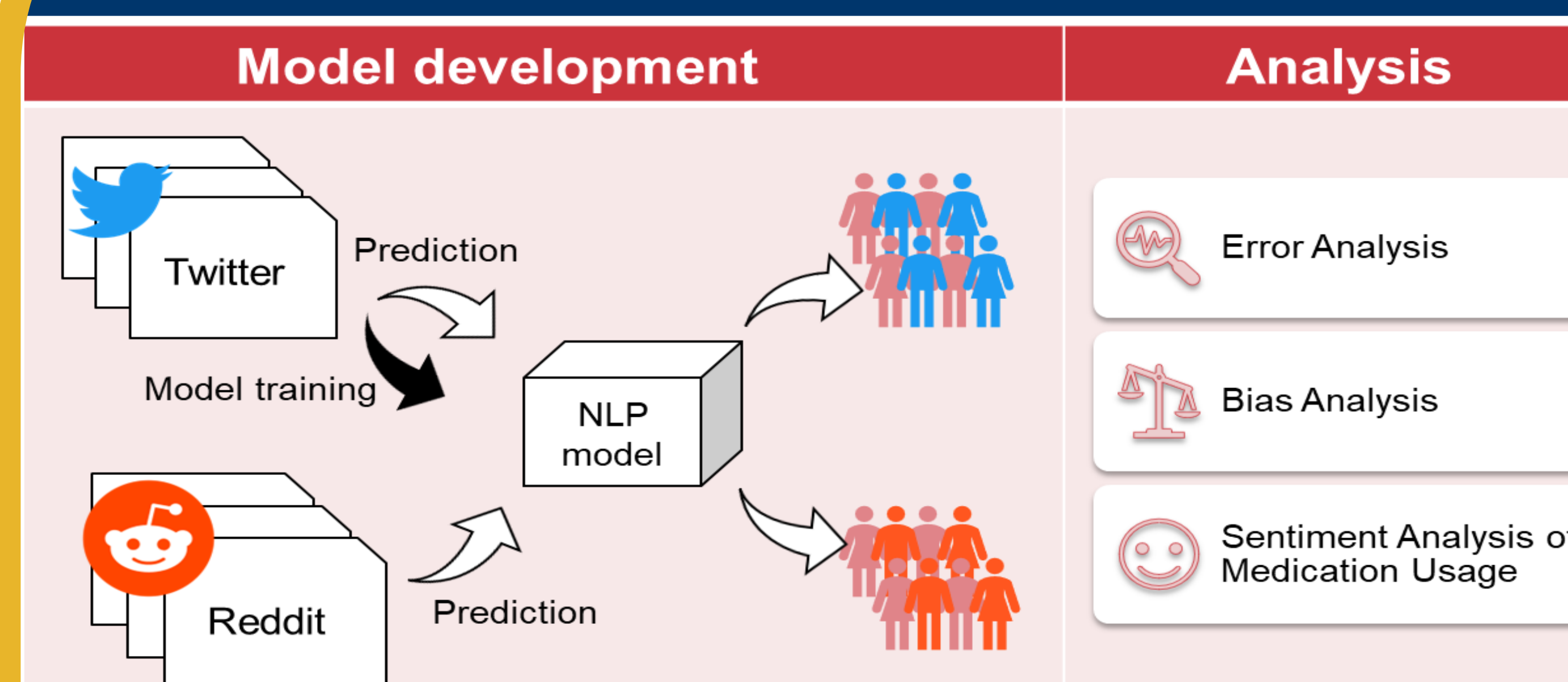| Model	Precision	Recall	F_1-score (95% CI)
Twitter Data			
RoBERTa	0.84	**0.95**	0.89 (0.87-0.91)
SciBERT	0.87	0.89	0.88 (0.85-0.90)
BioBERT	**0.88**	0.89	0.88 (0.86-0.91)
BioClinicalBERT	0.85	0.91	0.88 (0.86-0.91)
BERTweet	**0.88**	0.91	**0.90 (0.87-0.92)**
Clinical_KB_BERT	0.86	0.91	0.88 (0.85-0.90)
External: Reddit data			
RoBERTa	0.91	**0.95**	**0.93 (0.91-0.95)**
BERTweet	0.89	0.90	0.90 (0.87-0.93)

Methodology

Figure 1: The development framework of system.

- (Ref. fig. 1) Data collected from:
  - **Twitter API** (keywords "migraine" & generic/brand names of migraine-specific medications): **N = 5654**
  - **Reddit API** (four subreddits: r/migraine, r/NDPH, r/headache, r/headaches): **N = 302**
- 37% of Twitter, 75% of Reddit data was self-reported
- **Transformer-based models** evaluated to construct migraine self-report classifier (Ref. Table 1)
- **Error Analysis**: Analyze the contents of **false positives**
- **Bias Analysis**: impact of gender/identity word changes on model predictions.
- **Sentiment Analysis**: used **VADER** for tweets/posts

Results

- **Twitter**: **RoBERTa** achieved best recall (**0.95**)
- Optimal models were evaluated on **Reddit** dataset.
- **Reddit**: **RoBERTa** achieved best F_1 score (**0.93**)
- **Error analysis** show lack of context, ambiguous reference to word "migraine" as primary false positives
- Hard to spot such errors, **even for human annotator**
- Manual **Bias analysis** on 5% of all tweets in test set
- Changes in gender/identity words slightly affected word importance distributions; didn't alter classification results

Sentiment Analysis

- On **Twitter**, sentiment scores of **onabotulinumtoxinA**, **triptans**, **topiramate**, **beta-blockers**, and **tricyclic antidepressants** are neutral (0 mean score)
- Sentiment distributions of **CGRP monoclonal antibodies** and **gepants** tend to be more **positive**.
- In **Reddit** posts, **beta-blockers** have a **positive** sentiment while **topiramate** posts have a mean **negative** sentiment.
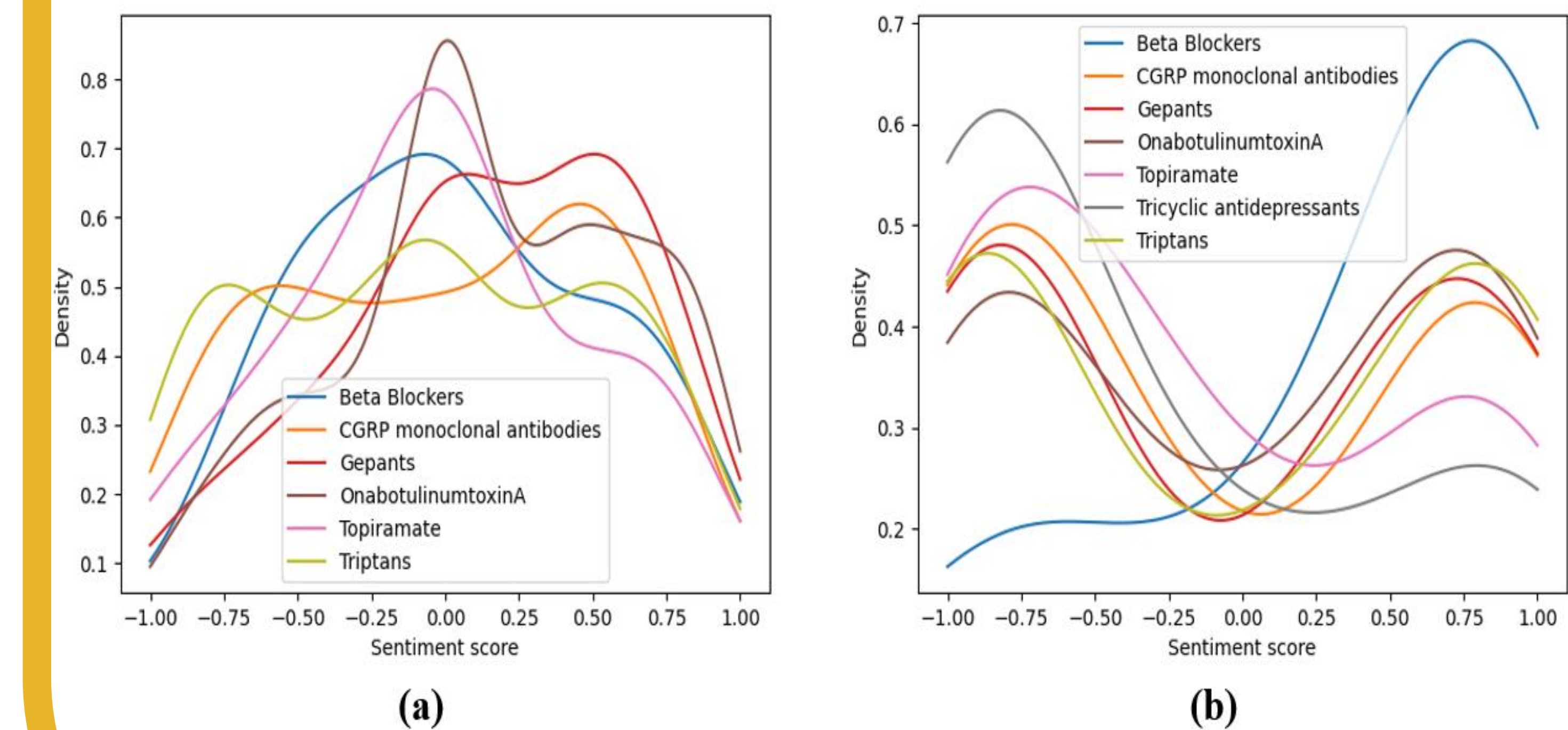

Figure 2. Normalized sentiment distributions (a) Twitter and (b) Reddit.

References

1. Nittas V, Lun P, Ehrler F, Puhan MA, Mütsch M. Electronic Patient-Generated Health Data to Facilitate Disease Prevention and Health Promotion: Scoping Review. *J Med Internet Res.* 2019;21(10):e13320. doi:10.2196/13320.
2. Al-Garadi MA, Yang YC, Cai H, et al. Text Classification Models for the Automatic Detection of Nonmedical Prescription Medication Use From Social Media. *BMC Medical Informatics and Decision Making.* 2021;21(1):1–13. doi:10.1186/s12911-021-01394-0.
3. Ghosh S, Misra J, Ghosh S, Podder S. Utilizing Social Media for Identifying Drug Addiction and Recovery Intervention. *IEEE International Conference on Big Data (Big Data)* 2020:3413–3422. doi:10.1109/BigData50022.2020.9378092.
4. Deng H, Wang Q, Turner DP, et al. Sentiment analysis of real-world migraine tweets for population research. *Cephalalgia Reports.* 2020;3:2515816319898867.

Georgia Tech

Wallace H. Coulter Department of Biomedical Engineering
at Georgia Tech and Emory University

EMORY UNIVERSITY