USING TWITTER DATA FOR PUBLIC HEALTH SURVEILLANCE AND PRECISION DIAGNOSTICS OF AUTISM SPECTRUM DISORDER

Aditi Jaiswal

Dr. Peter Washington, Assistant Professor (ICS, UH Manoa)

MS Plan A , Spring 2023

INFORMATION & COMPUTER SCIENCES
UNIVERSITY of HAWAIʻI at MĀNOA

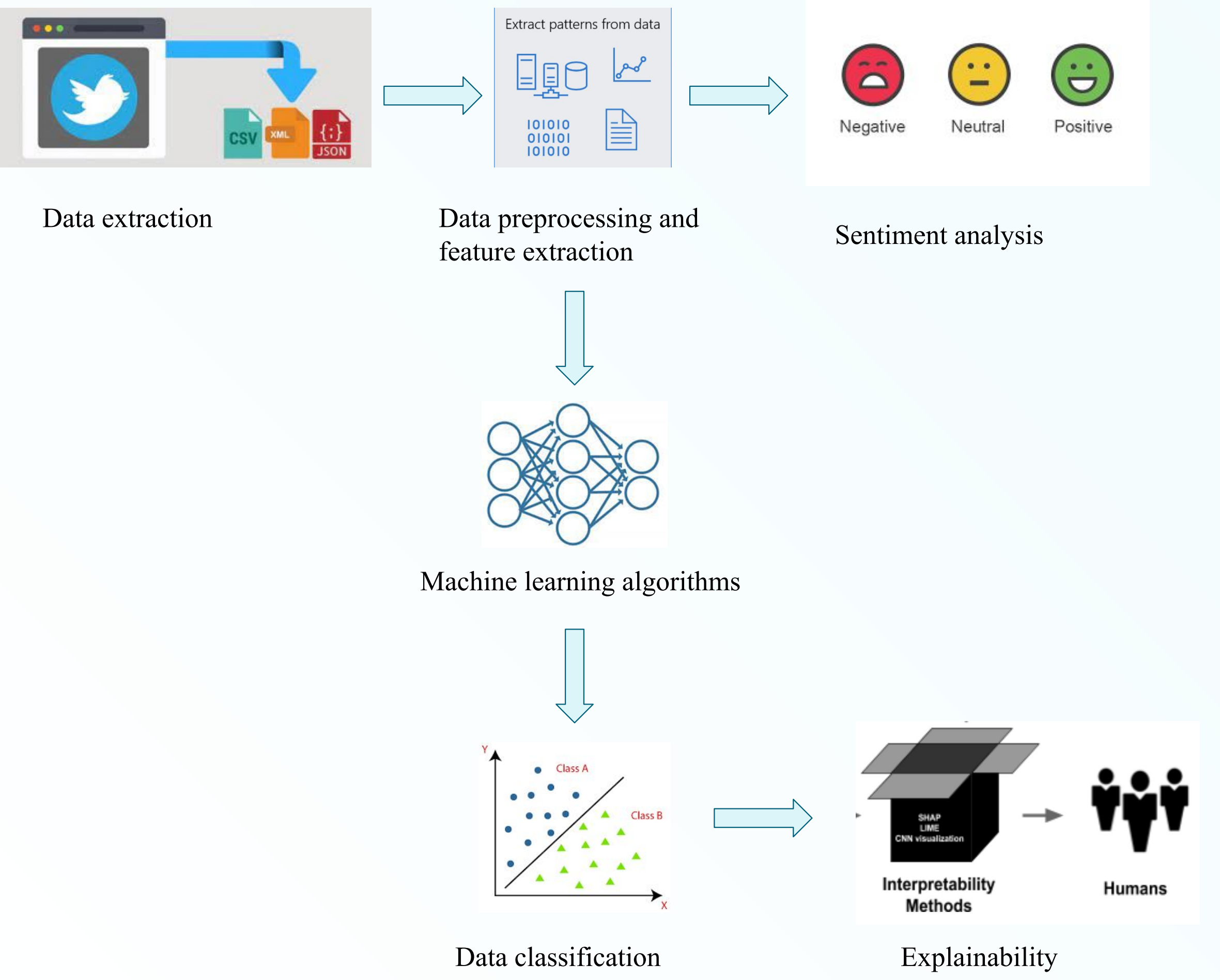
Abstract

Twitter has proven to be an exceptional source of health-related information from both public and health officials, using which researchers and clinicians can conduct studies on diseases and associated symptoms in natural settings by establishing digital phenotypic biomarkers. In this study, we scraped autism related tweets from users self identifying with ASD, using the hashtag "#ActuallyAutistic", and illustrate the usefulness of the curated dataset through simple applications such as: sentiment analysis, text classification and topic modeling. The textual differences in social media communications can help identify various behavioral symptoms, which can be used by the clinicians to distinguish an autistic individual from their typical peers and tailor the treatment plan.

Introduction

- Autism spectrum disorder (ASD): group of developmental disorders causing physical, cognitive and behavioral changes.
- **Challenges:**
  - Delayed diagnosis - autism prevalence has increased 317% since 2000.
  - Misdiagnosis - because of its complex nature its characteristics can often be mistaken for other disorders such as anxiety, OCD, and ADHD.
- **Motivation:** The digital data from wearables, smartphones or social media can serve as a source of observational data - *Digital phenotyping.*
- **Research objective:** using Twitter as the source of observational data:
  - to analyze symptomatic signals and textual differences; frequent topics of discussion and associated sentiments and behavioral patterns
  - using machine learning (ML) for better analytics, extract hidden patterns and build a text classifier.
  - public release of the curated dataset for collaborative scientific research.

Materials



Data extraction

Data preprocessing and feature extraction

Sentiment analysis

Machine learning algorithms

Data classification

Explainability

Methodology

Curating novel dataset:

- Twitter API was used to scrape English tweets using the hashtag "#ActuallyAutistic" from Jan 2014 to Dec 2022
- Focused on user's bio having the keywords "autism", "autistic" or "neurodiverse"
- To build a balanced classifier, random tweets were collected (excluding any ASD user)
- Weakly supervised data labeling

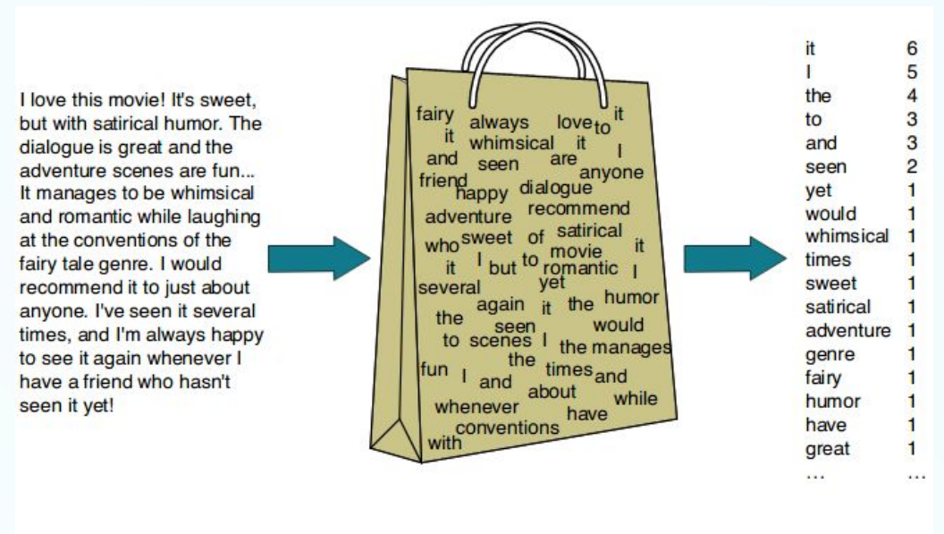
Exploratory Data Analysis:

- Histogram of character counts of the tweets
- Sentiment Analysis: used VADER to analyze the sentiment score of the autism dataset
- Word cloud for data visualization

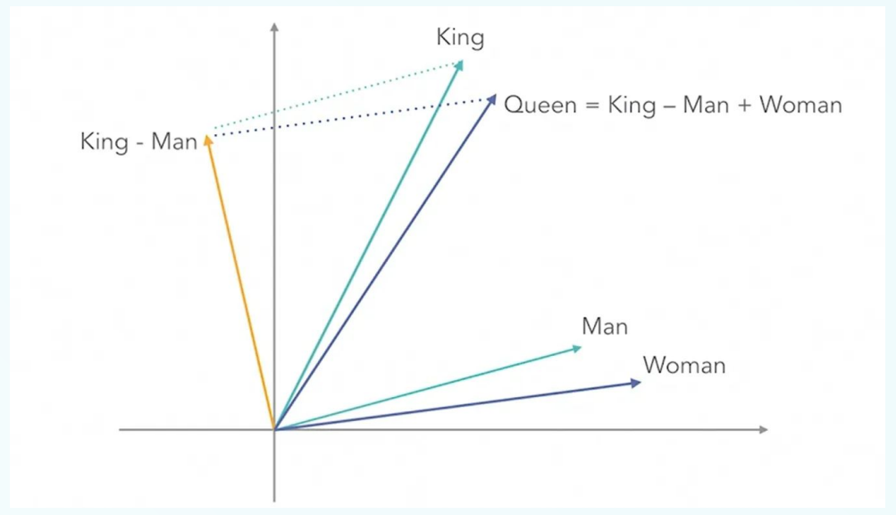
Text classification:

- Data cleaning: text preprocessing to remove profane words, punctuations, stop words, tokenization using NLTK
- Feature extraction:
  - **Bag of words + TF-IDF:** using vocabulary of words rescaled by their frequency of occurrence in all the documents
  - **Word embeddings using Word2Vec:** using dense word-vector representation to capture the meaning of the sentence
- Machine Learning
  - usernames divided into 85:15 ratio of training and test dataset
  - input data: preprocessed tweets from each set of users transformed by tf-idf or word2vec
  - 5 fold cross validation using different algorithms to find the best performing (higher accuracy) and faster algorithm

BoW



Word2Vec

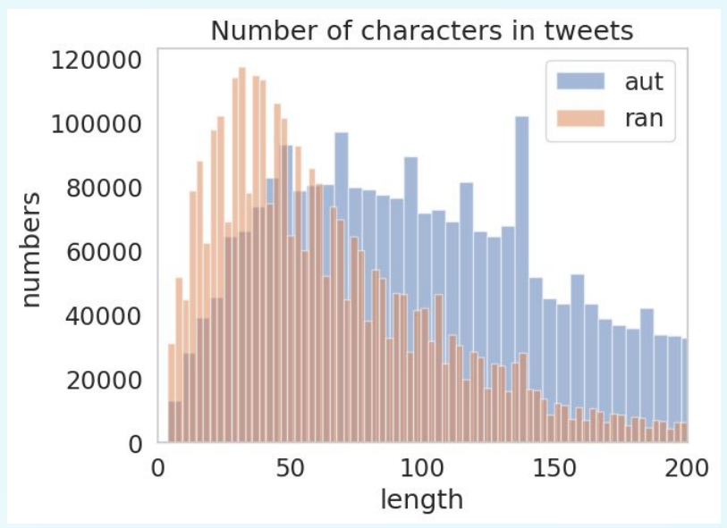


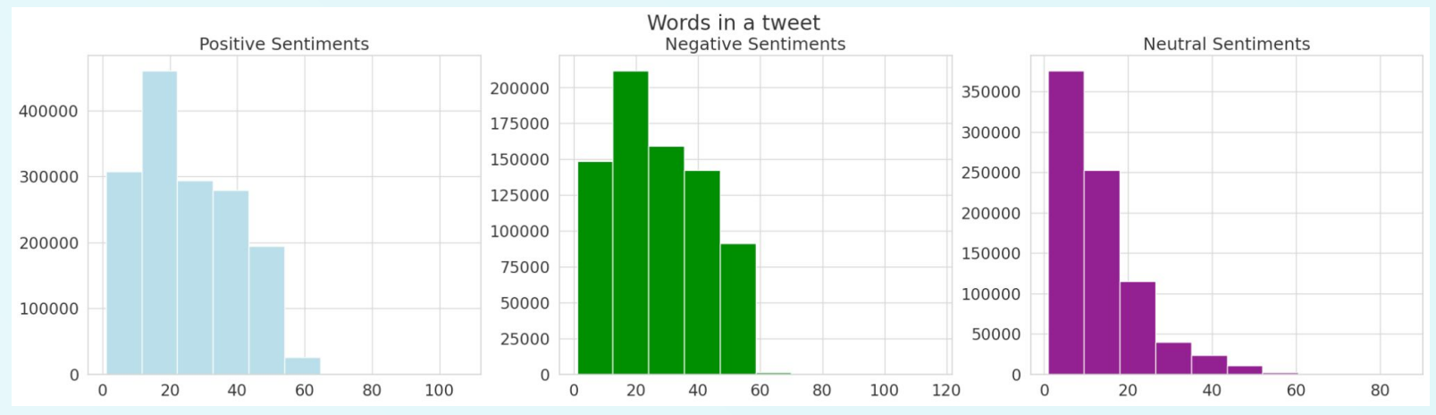
Explainability

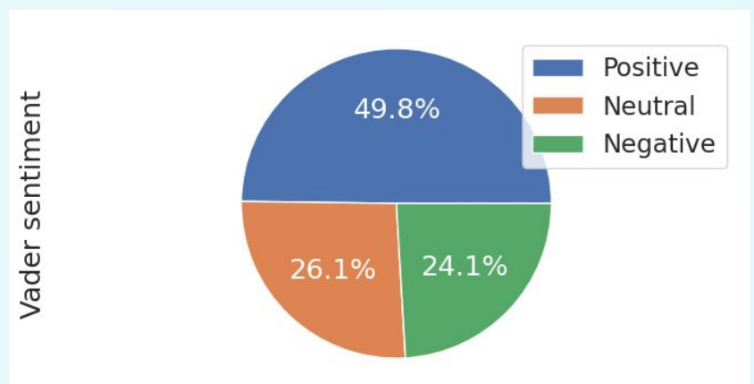
- important to interpret the black-box nature of ML algorithms
- introduce transparency in the model to reduce bias or any ethical/legal concern
- Local Interpretable Model-agnostic Explanations (LIME) was used with TF-IDF feature vectors

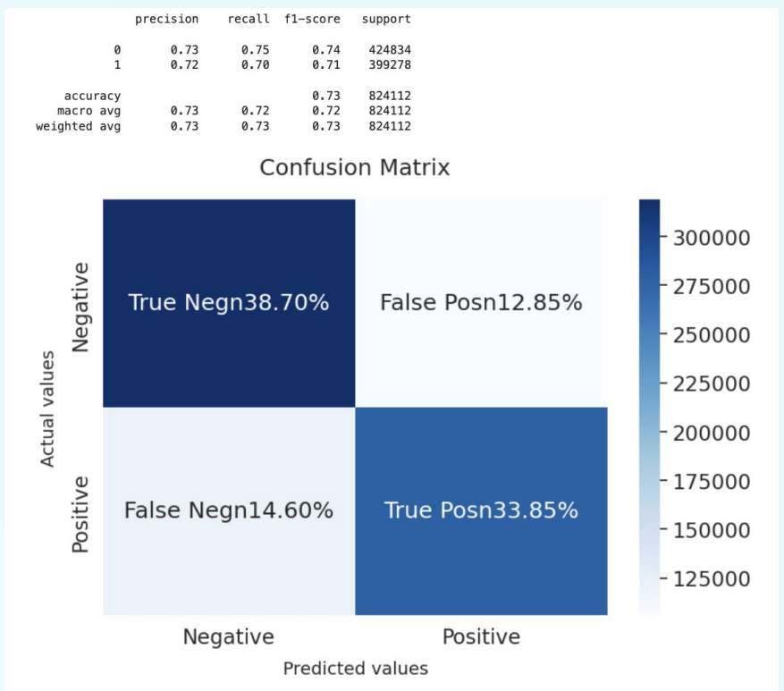
Topic modeling

- unsupervised learning technique to identify hidden topics
- used to study the trends and topics of discussion for better policy making and identify at-risk groups
- Top2Vec algorithm was used
  - clusters semantically similar words in the same topics
  - uses spatial proximity of the words





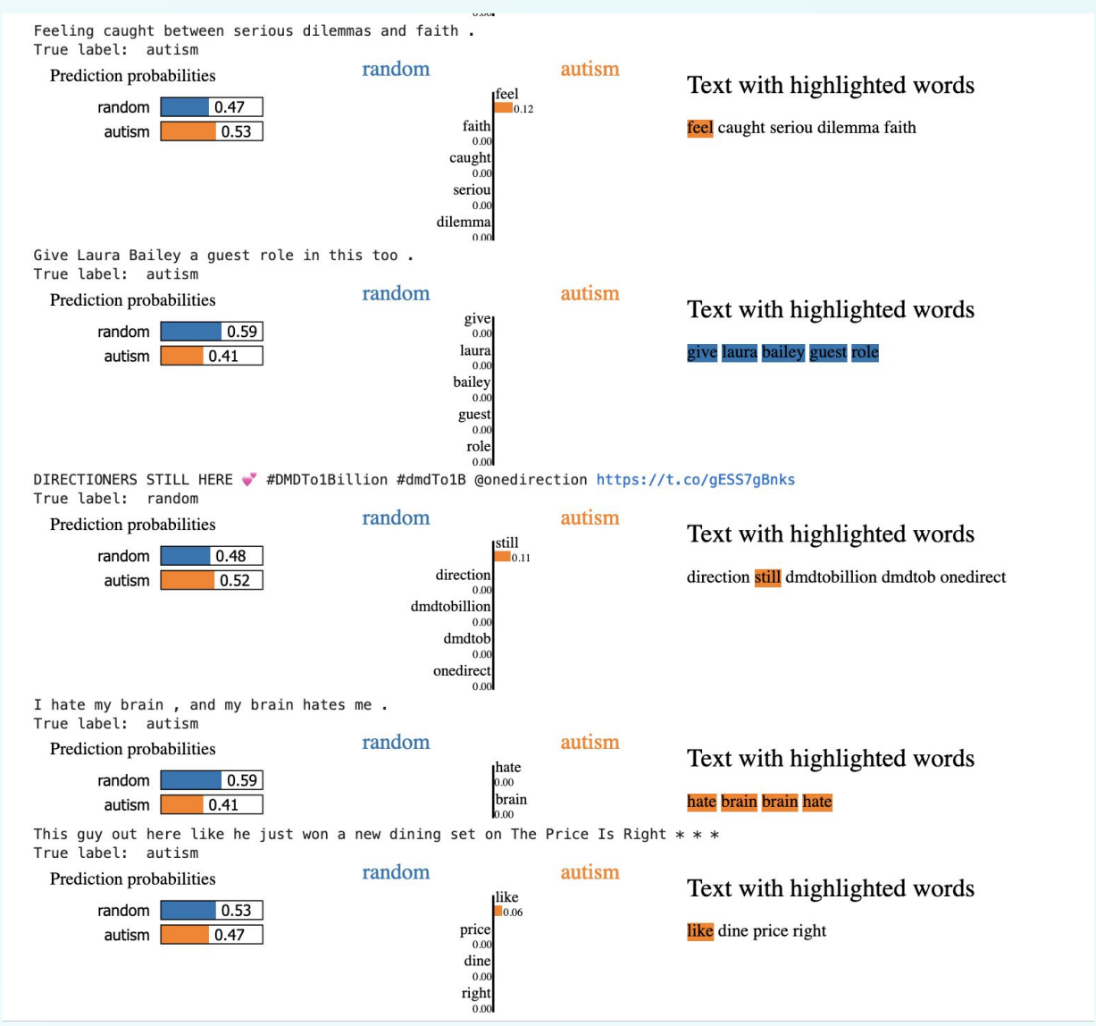




```
              precision    recall  f1-score   support

           0       0.73      0.75      0.74    424834
           1       0.72      0.70      0.71    399278

    accuracy                           0.73    824112
   macro avg       0.73      0.72      0.72    824112
weighted avg       0.73      0.73      0.73    824112
```



```
Feeling caught between serious dilemmas and faith .
True label:  autism
```

Prediction probabilities: random 0.47, autism 0.53 — random / autism — feel 0.12, faith 0.00, caught 0.00, seriou 0.00, dilemma 0.00

Text with highlighted words: feel caught seriou dilemma faith

```
Give Laura Bailey a guest role in this too .
True label:  autism
```

Prediction probabilities: random 0.59, autism 0.41 — random / autism — give 0.00, laura 0.00, bailey 0.00, guest 0.00, role 0.00

Text with highlighted words: give laura bailey guest role

```
DIRECTIONERS STILL HERE 💕 #DMDTo1Billion #dmdTo1B @onedirection https://t.co/gESS7gBnks
True label:  random
```

Prediction probabilities: random 0.48, autism 0.52 — random / autism — still 0.11, direction 0.00, dmdtobillion 0.00, dmdtob 0.00, onedirect 0.00

Text with highlighted words: direction still dmdtobillion dmdtob onedirect

```
I hate my brain , and my brain hates me .
True label:  autism
```

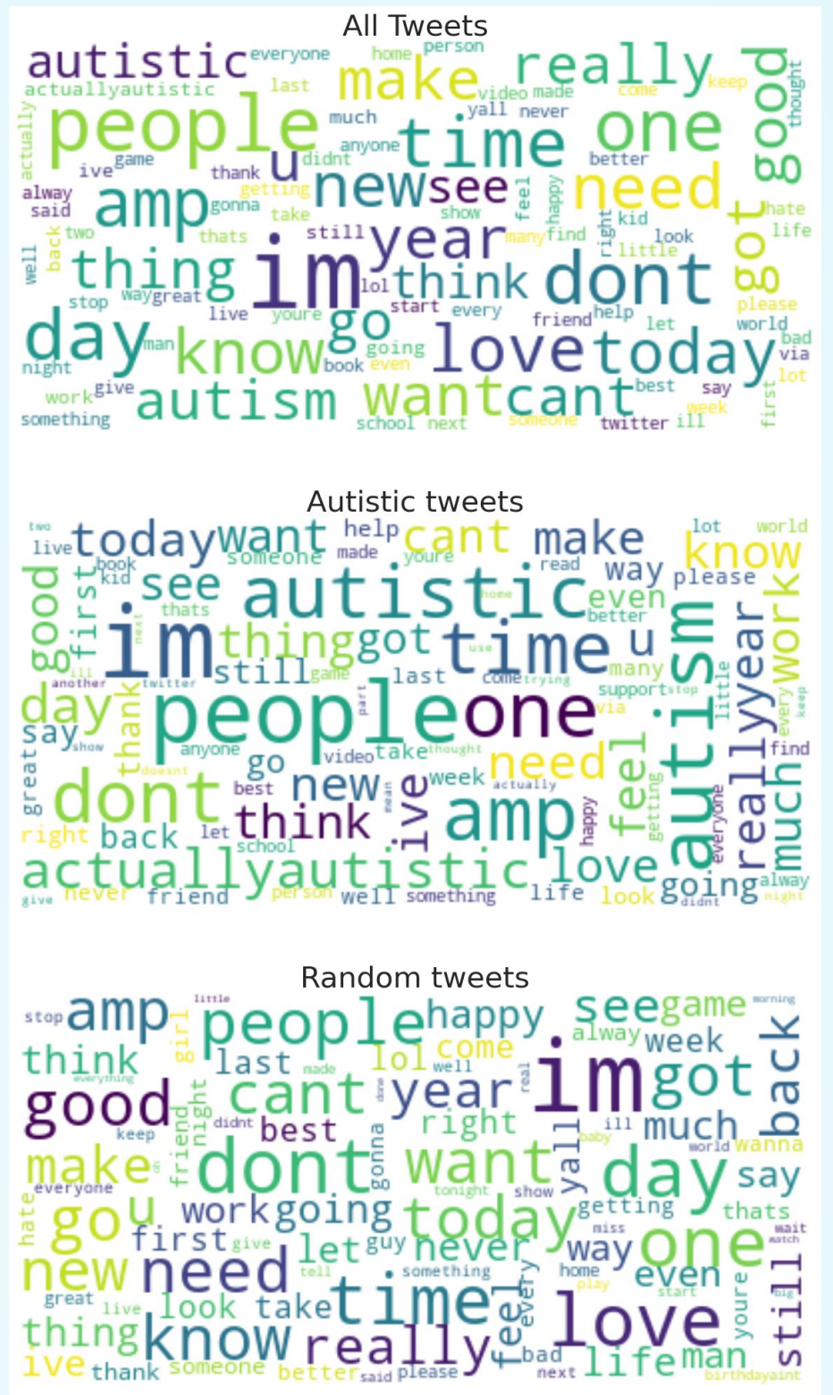
Prediction probabilities: random 0.59, autism 0.41 — random / autism — hate 0.00, brain 0.00

Text with highlighted words: hate brain brain hate

```
This guy out here like he just won a new dining set on The Price Is Right * * *
True label:  autism
```

Prediction probabilities: random 0.53, autism 0.47 — random / autism — like 0.06, price 0.00, dine 0.00, right 0.00

Text with highlighted words: like dine price right



Results

- Total 6,469,994 tweets were collected from approximately 70,000 unique users.
- Logistic regression (LR) was chosen as the best predictor
  - BoW + LR gave an accuracy of 63%
  - Word2Vec + LR gave an **accuracy of 73%**, with better cosine similarity of words
- ASD vs control group tweets (captured by BoW+LIME)
  - autistic individuals use more characters/words
  - observed more positive sentiment tweets with higher word counts, followed by neutral and negative polarities.
  - ASD tweets are more emotion-driven with higher usage of words like "kids" and "school" while control group used more verb-based words sharing daily activities related tweets
- Prominent topics identified:
  - behavioral and emotional symptoms: "hyperactivity", "tourette", "fidgeting", "fear", "jitters" and "anxiety",
  - "vaccine", and "misinformation"
  - "therapy", "diagnosis" and "cats".

Conclusion

- This study shows the textual distinctions between people with and without ASD with an **accuracy of 73%**
  - not a diagnostic tool but the promising results can be used for behavioral interventions and symptomatology,
  - useful for preliminary screening by clinicians or statistical surveys by CDC/NIH
- Consistent results with prior studies done using computer vision models (**facial expression and vocal characteristics achieved an accuracy of 73% and sensitivity of 67%**) published in leading journals (e.g., *Nature* and *Nature Digital Medicine*).
- Future work:
  - combining textual features with other modality data such as audio (conversations) and video (eye gazing, facial expressions) to improve the rigor of autism research,
  - using deep learning and pre-trained language models