What is your ranking? Predicting Social Feedback from Social Media Platform Comments

Problem Statement

Our goal:

To predict the probability of getting a higher score with respect to the scores which other users obtained, via z-score prediction.

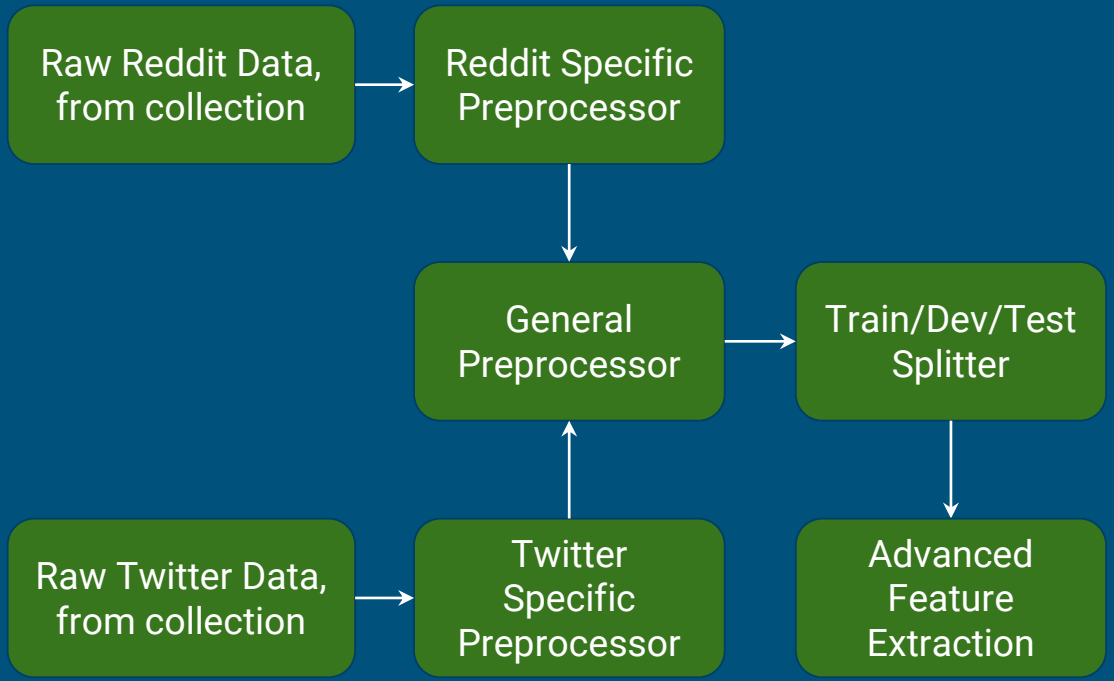
Projected benefits:

- *Allows users to revise their contents and optimize future scores*
- *Applicable to other domains*
- *Help researchers to explore the impacts of the variables such as topics, sentiment score, and readability of contents on the attention online posts get from users and further develop a methodology with better results for social feedback prediction tasks*

Dataset

- Original : Reddit dataset from the Depression subreddit (2015-2017)
- New : Twitter dataset which is crawled based on keyword matching using the real-time streaming API
  - Twitter1 dataset : first-person pronoun
    - Contains tweets that explicitly include keywords such as `I', `my', `me', `myself'.
  - Twitter2 dataset: job, occupation
    - Contains tweets which are closely related to a job and an occupation.

Data Processing Pipeline - High level view



- Data comes from target source in API specific format.
- Data is converted into a unified format with basic fields: score calculation, original text, stopword removed text, etc.
- General features such as time, day of the week are processed.
- Data is split into Train/Dev/Test datasets for future feature extraction, ensuring a fair balance of score examples.
- Advanced feature selection is done: topic modeling, sentiment analysis, readability, etc.

Score calculation

Reddit:

- Upvote and downvote totals are merged internally to a score, and are unavailable through the API.

Twitter:

- Retweets, favorites, replies and quotes are provided through the API. These are summed to create a score.

Basic Features

- The submitted day of the week (0-6), month, year
- The submitted time (1-6, 7-12, 13-18, and 19-24)
- The number of words
- The negative, neutral, positive, and the compound sentiment level
- Readability of posts
- The topic distributions of each post and title

Advanced Features - VADER

- **VADER sentiment analyzer**
  - A lexicon and rule-based sentiment analysis tool customized for social media
  - Uses a combination of lexical features (e.g., words) and emojis which are labelled according to their semantic orientation
  - positive, negative, neutral, and compound sentiment scores for posts

```
sentiment_analyzer_scores("The phone is super cool.")

The phone is super cool---------------- {'neg': 0.0, 'neu': 0.326,
'pos': 0.674, 'compound': 0.7351}
```

Readability of Posts

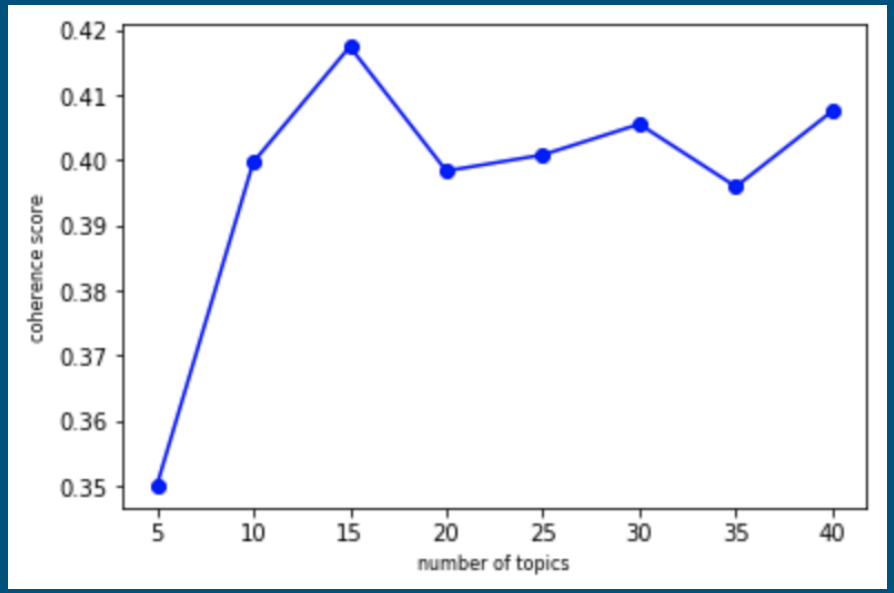
- **Flesch Reading Ease Metrics**
  - The ease of reading a text is estimated based on:
    - # of words per sentence
    - # of syllables per word
  - Higher FRE score → easier to read

$$Flesch\ Reading\ Ease = 206.835 - 1.015 * \frac{total\ words}{total\ sentences} - 84.6 * \frac{total\ syllables}{total\ words}$$

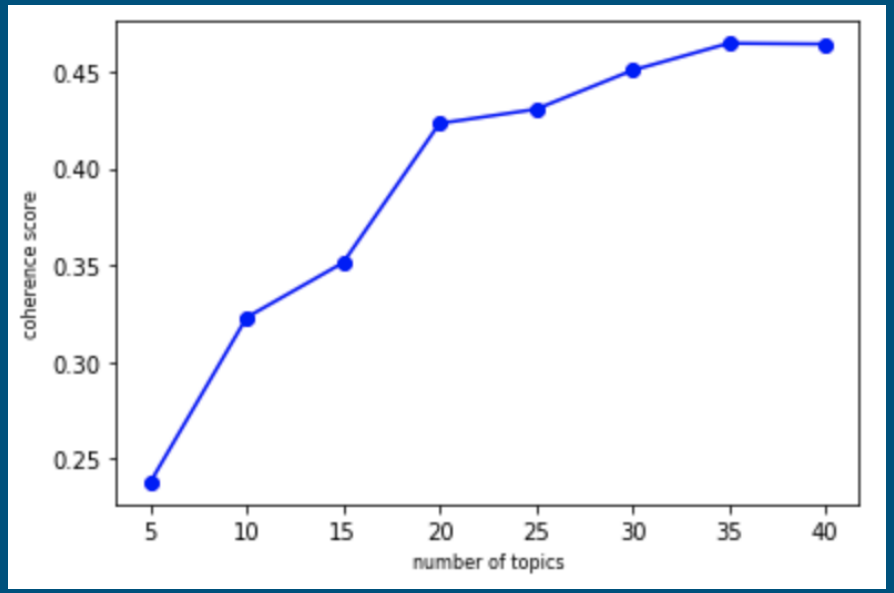
Advanced Features - LDA Topic Modeling

- LDA topic modeling using MALLET
  - An unsupervised learning model that extracts latent topics when a set of documents is given
  -  Removed insignificant words using TF-IDF and over-significant words
  - Found the optimized number of topics via coherence score
  - Performed on both title/content for Reddit dataset and only on the content for Twitter dataset

Advanced Features

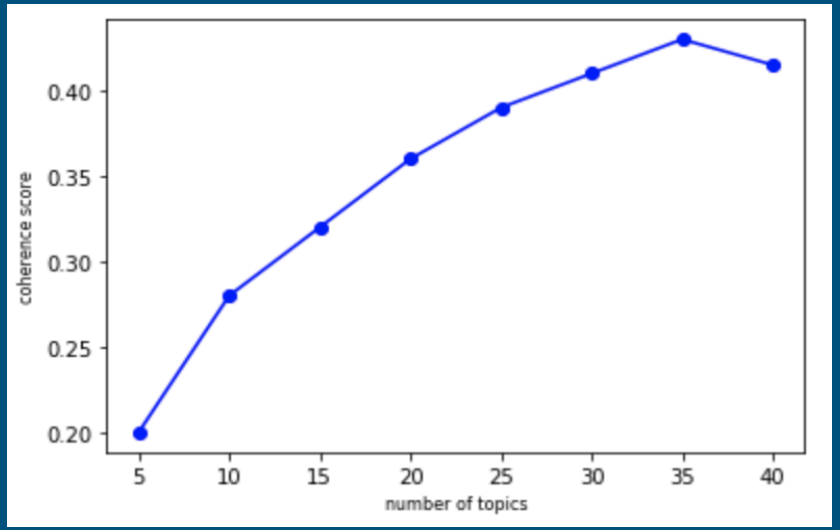


Reddit Content

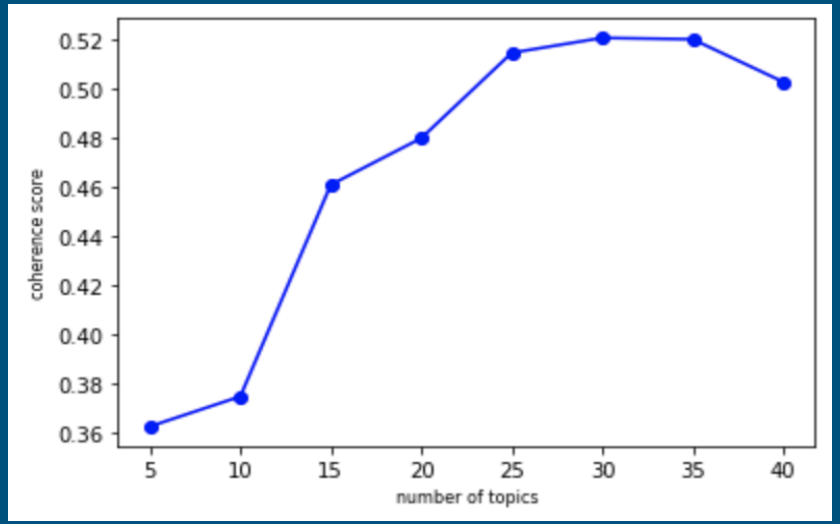


Reddit Title

Features



Twitter1



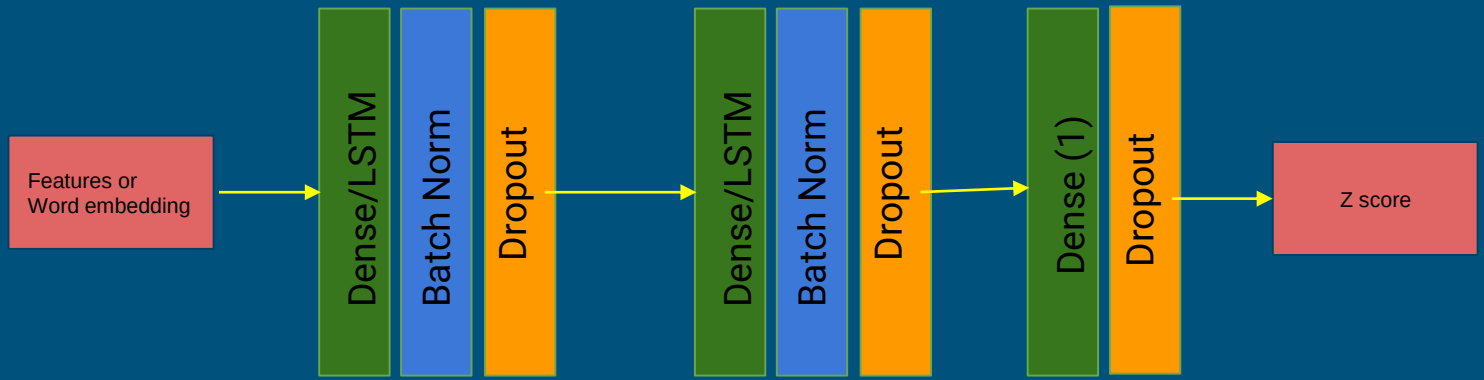
Twitter2

Models

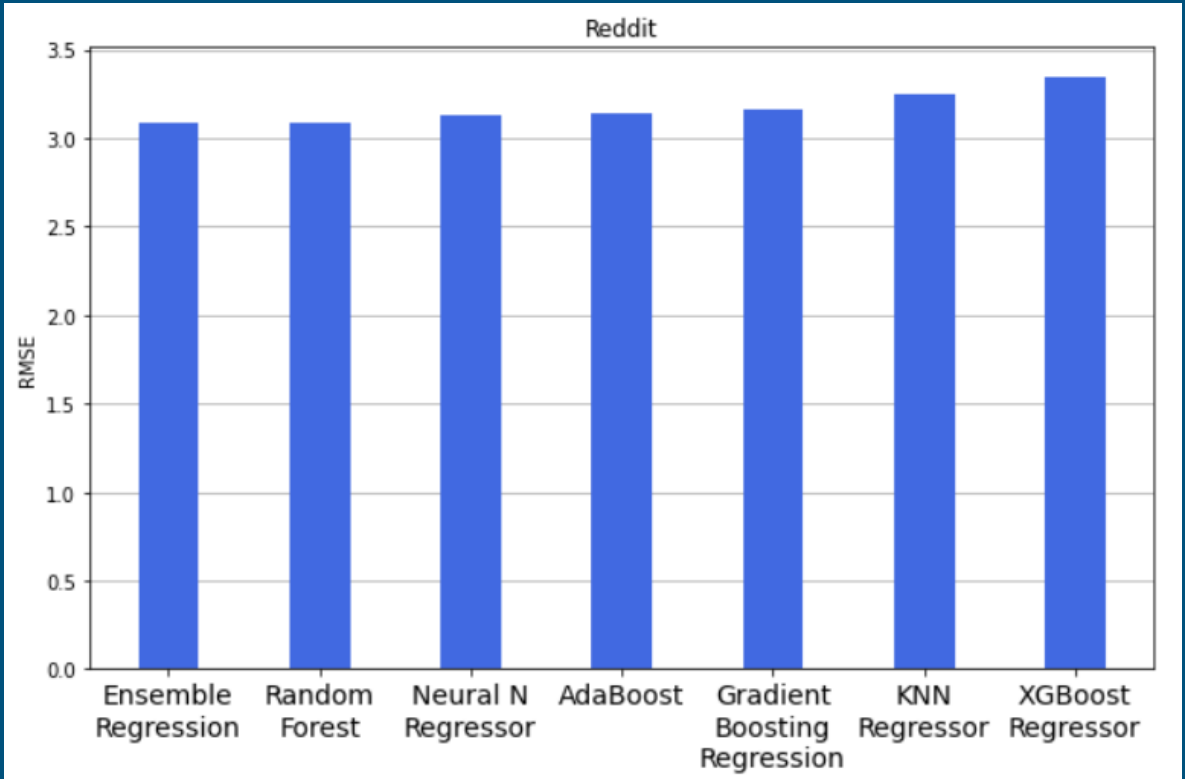
- Adaboost
- Ensemble regressor
- Gradient boosting regressor
- KNN
- Neural network regressor
- Random forest
- XGBoost regressor

Neural Network model

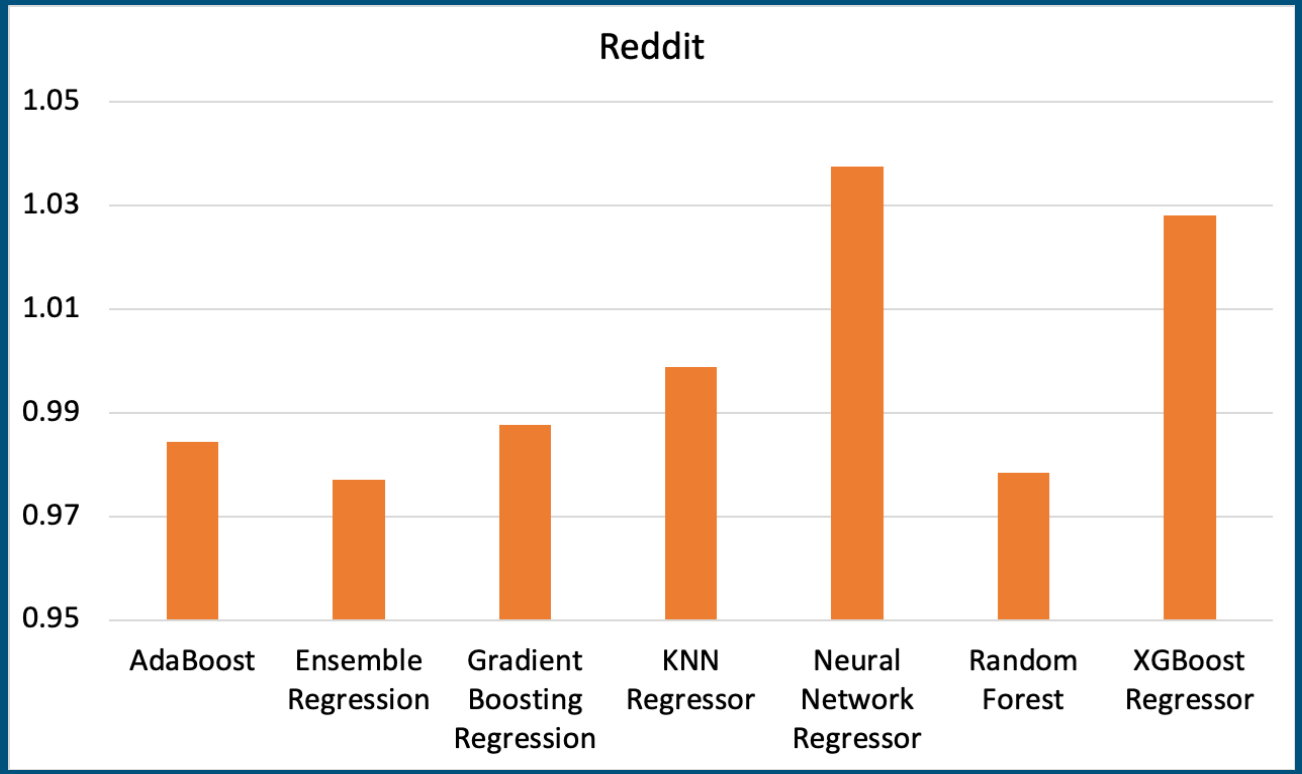
1. Manual feature based (80 reddit, 41 twitter)
2. LSTM based (Post, Title)



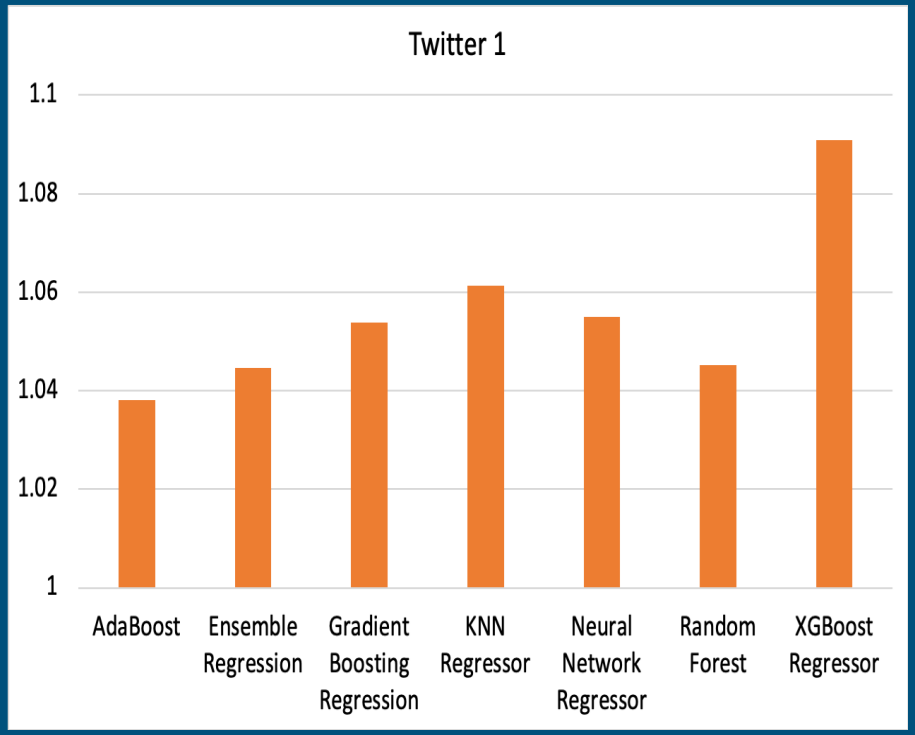
Preliminary results : RMSE (20+ to 3.09)

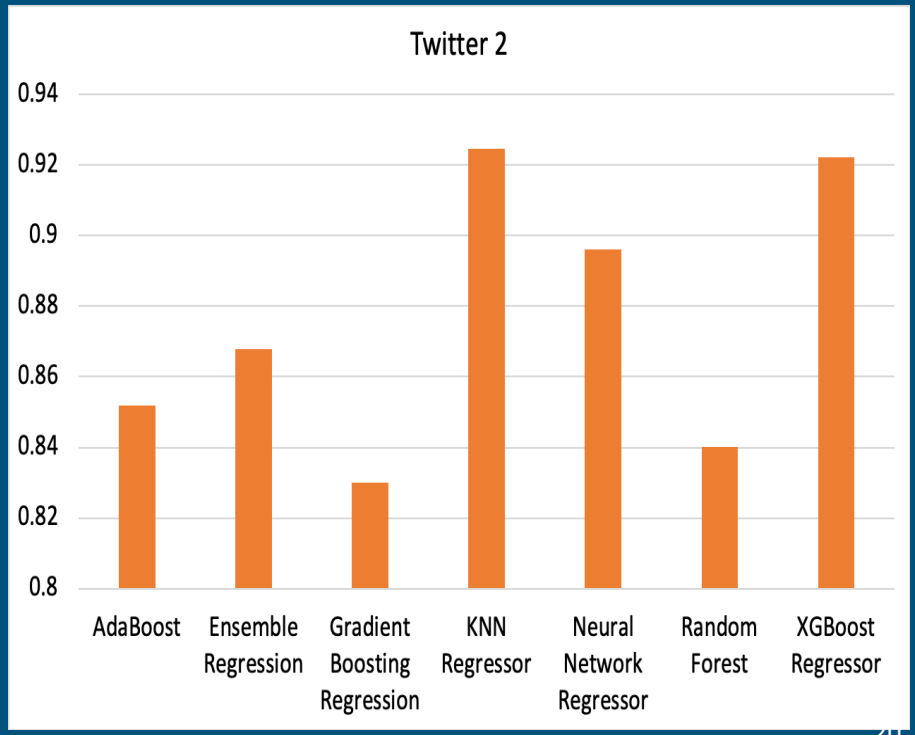


Final results : RMSE



Final results : RMSE





Feature Importance

Reddit

Feature	Importance
Title Topic 5	8.88%
Content Topic 14	8.50%
Content Topic 4	6.80%
# of Words in Content	6.22%
Compound Sentiment of Title	6.18%
Content Topic 10	6.11%
Positive Sentiment of Content	5.89%
Content Topic 5	3.87%
Compound Sentiment of Content	3.78%
Negative Sentiment of Title	3.46%

- Sentiment of title and content is important in predicting how much attention a post will receive;
  → *Reddit data were collected from **Depression** subreddit*

- **Title Topic 5:** *shit, useless, stupid, terrible, shitty, crap, dumb, miss, lonely, worry, bore, miserable …*
  → *swear words and **negative emotion-related** words*

- **Content Topic 14**: *Friend, hang, social, people, play, meet, group, talk, close, game, conversation, invite, joke …*
  → *social interaction*

- **Content Topic 4**: *day, sleep, night, bed, home, wake, morning, cry, spend, tonight, fall asleep, late, weekend …*
  → *sleep, night*

Feature Importance

Twitter 1

Feature	Importance
Topic 10	30.46%
Neutral Sentiment	14.85%
Negative Sentiment	13%
Content Readability	7.03%
Topic 32	4.82%
Topic 18	3.87%
Topic 29	3.37%
Topic 13	2.62%
Topic 25	2.26%
Topic 23	2.01%

Twitter 2

Feature	Importance
Topic 29	8.83%
Topic 5	7.89%
Topic 23	7.56%
Content Readability	5.89%
Topic 34	5.88%
# of Words in Content	5.88%
Topic 3	5.8%
Topic 25	4.5%
Topic 13	4.03%
Topic 4	3.96%

- **Content Readability** is relatively more important in Twitter
- Sentiment of a post is less important in Twitter

Top 5 Important Features

Twitter 1

Feature	Importance
Topic 10	30.46%
Neutral Sentiment	14.85%
Negative Sentiment	13%
Content Readability	7.03%
Topic 32	4.82%
Topic 18	3.87%
Topic 29	3.37%
Topic 13	2.62%
Topic 25	2.26%
Topic 23	2.01%

- **Topic 10:** *hate, care, break, heart, beautiful, friend, adore, funny, disgust, ….:* emotion-related words
- Neutral / Negative sentiment embedded in Twitter 1 (first-pronoun dataset);
  → *might be related to how people express **empathy** by responding to others' posts*

- Topic 32: *friend, family, baby, girlfriend, boyfriend, brother, crush, wife, husband, trust...*
  → *social relationship*

Feature Importance

Twitter 2

Feature	Importance
Topic 29	8.83%
Topic 5	7.89%
Topic 23	7.56%
Content Readability	5.89%
Topic 34	5.88%
# of Words in Content	5.88%
Topic 3	5.8%
Topic 25	4.5%
Topic 13	4.03%
Topic 4	3.96%

- Topic 29: *interview, program, first weeks employment, pick, fair, train, question, review, hire, ...*
  → employment-related words
- Topic 5: *great, team, win, proud, congratulations, fit, interest, congrats, join, awesome, earn, finally, …*
  → work accomplishment
- Topic 23: *engineer, technical, development, software, technology, software engineer, electrical, design, product, ericsson, ...*
  → software engineer

Discussion

- Advancement from the previous stage:
  Removing insignificant words using TF-IDF and over-significant words before extracting topics improves the performance of a model to a great extent
- Readability is critical in raising attention from others in Twitter
- RMSE values lower than 1 → the suggested approach is applicable in predicting how much attention a given post will receive from other users with respect to others' posts when adjusted to a certain platform or a domain

Thanks