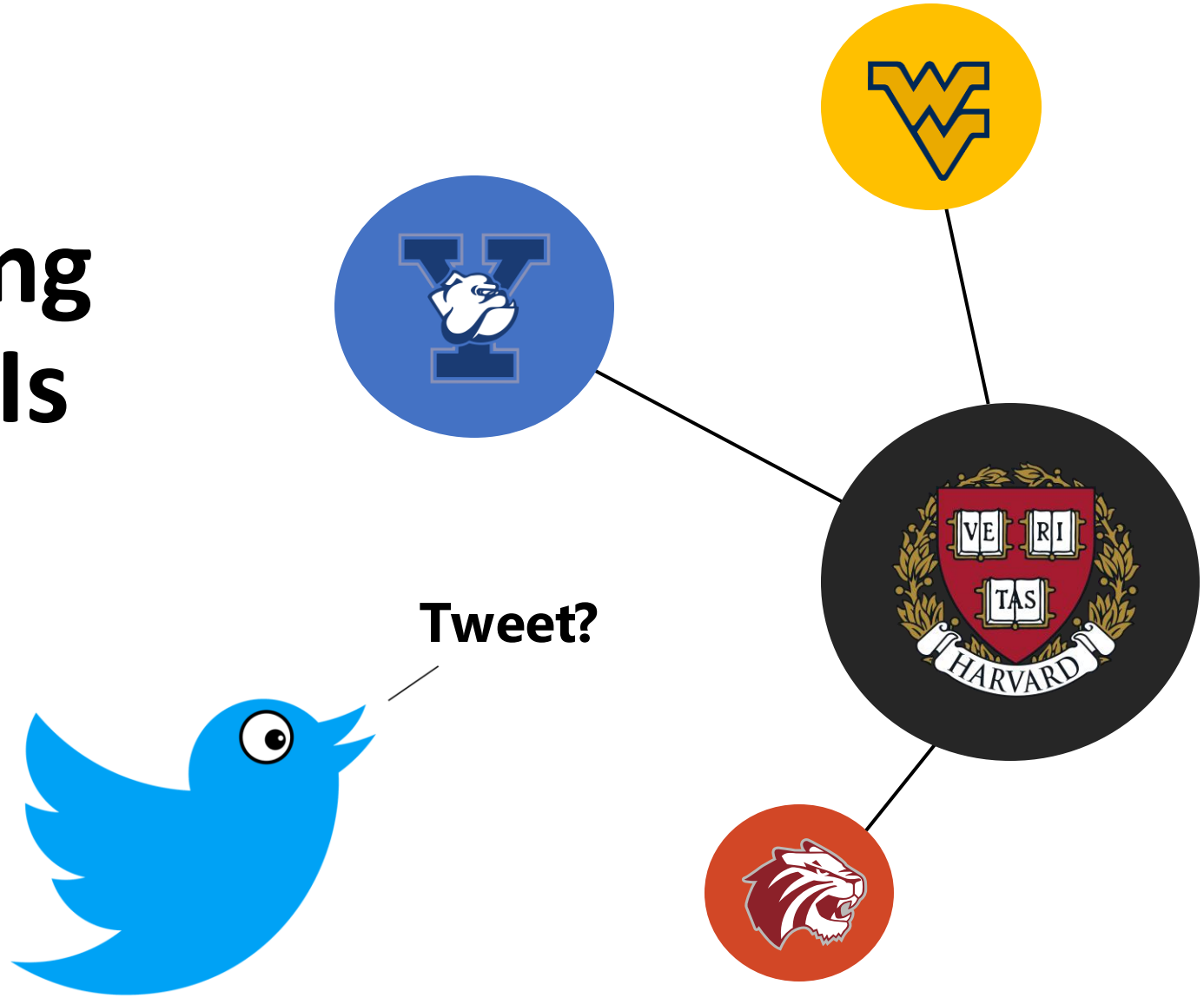
U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

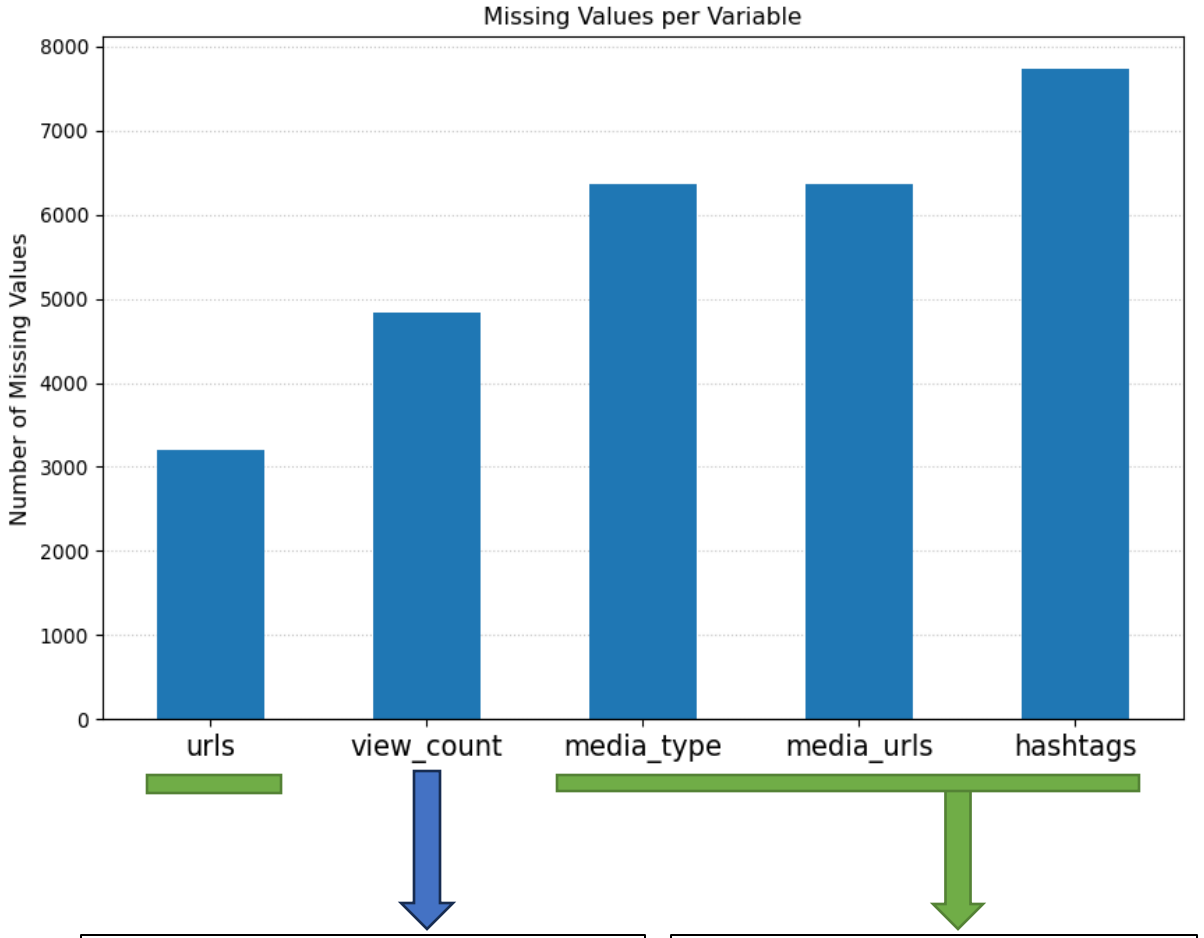
Analysis of Posting Strategies for HEIs



Data Mining II Project by:

- Adriano Chessa
- Carlos Vilela
- Gabriel Guimarães
- Pedro Leite

Part 3A: Data Pre-processing and Analysis



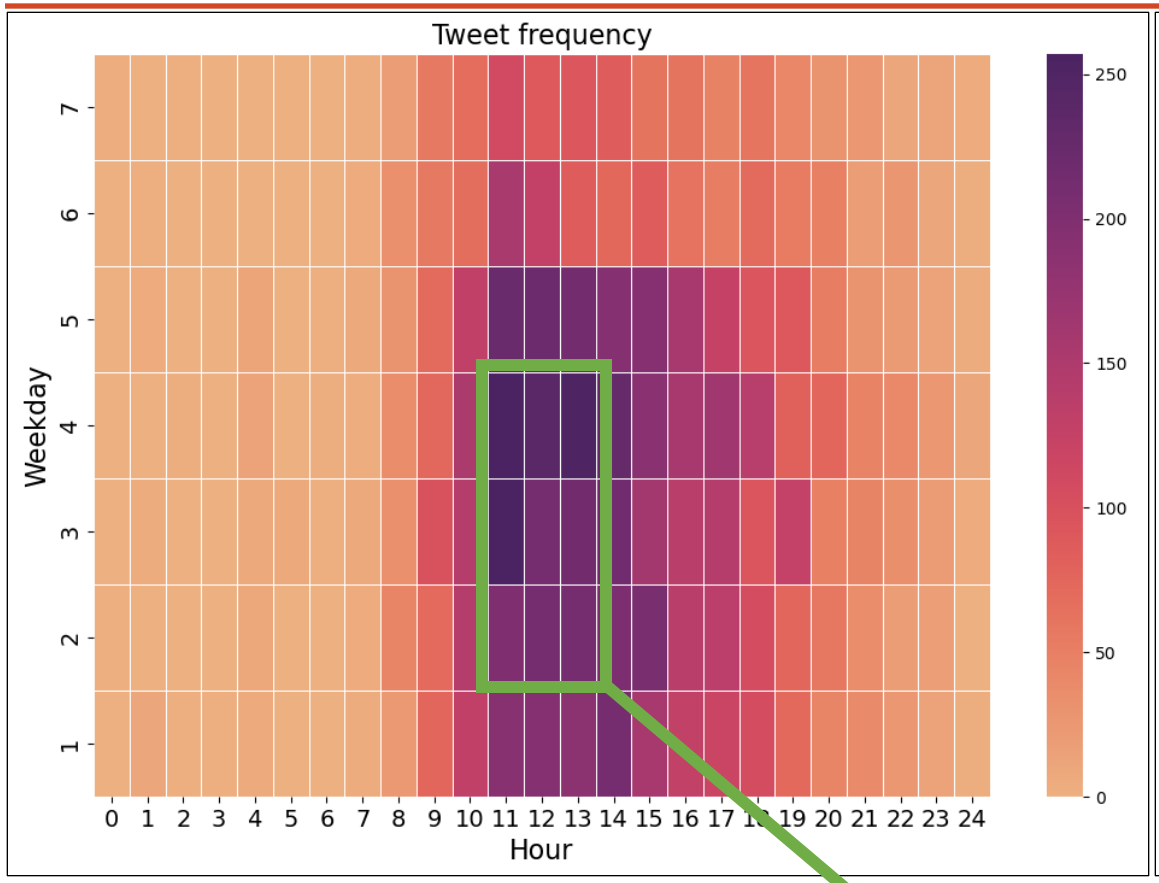
Data Imputation:
Median of Views per HEI

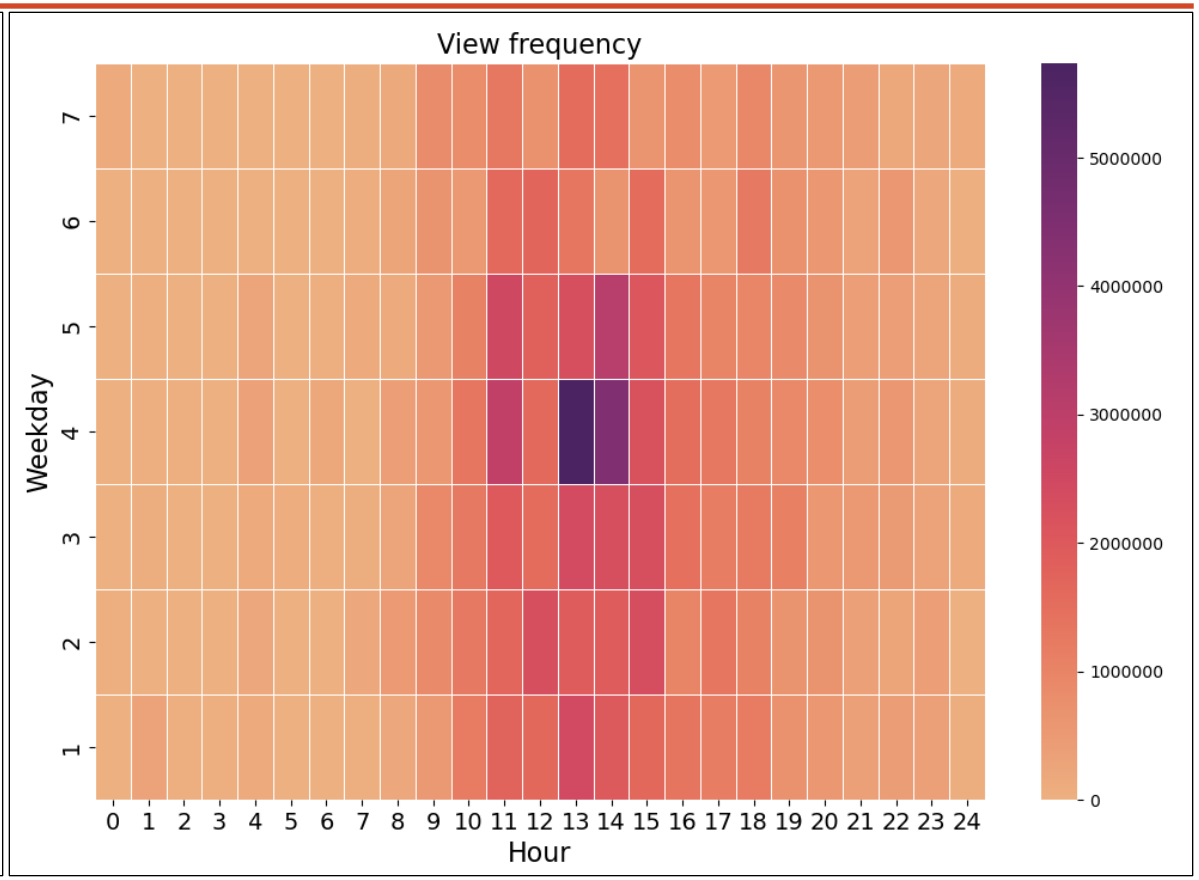
Binary Classification:
Present or Absent

Transformations:

- Complutense HEI dropped as only one observation was present.
- MITs "We did not subscribe to Twitter Blue." Tweet dropped as the consequent skewness was too high.
- Tweet length calculated including punctuation and spaces.
- Specifying type of media.
- "created_at" variable divided into weekday and hour of the day.

Part 3B: Univariate Analysis

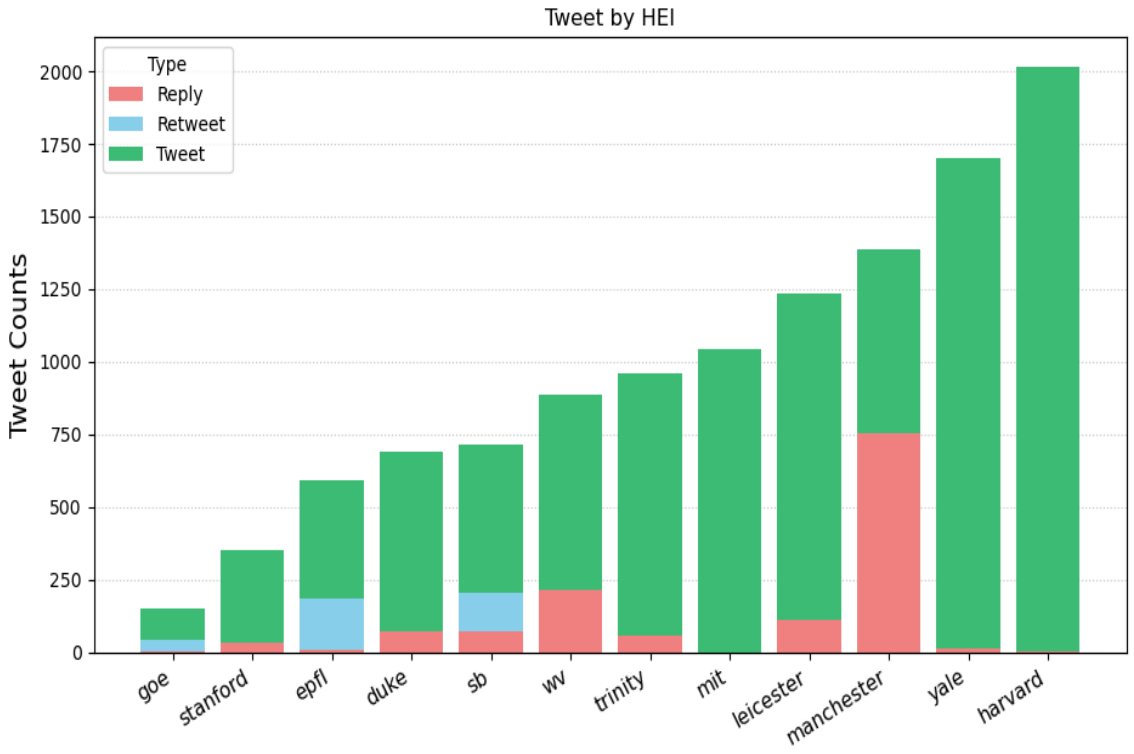


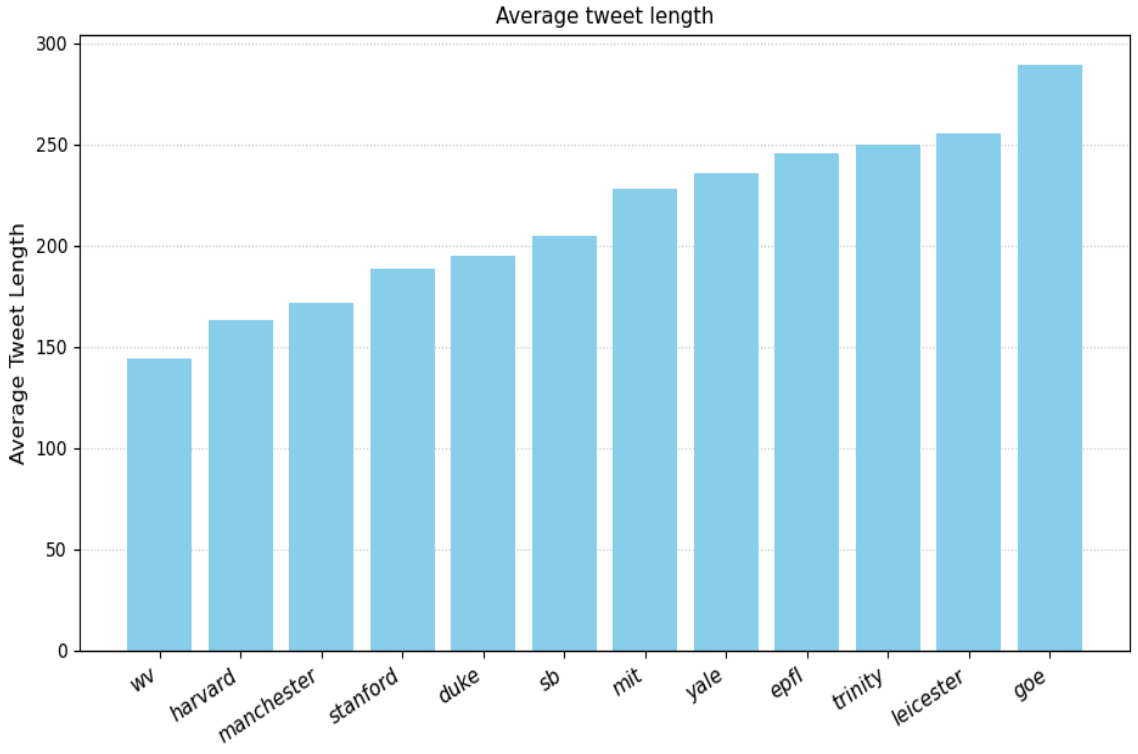


Mode

- Most tweets published by HEIs during weekdays between 11 and 16 UTC.
- Most viewed tweets are the ones published on Thursday's early afternoon.

Part 3B: Univariate Analysis

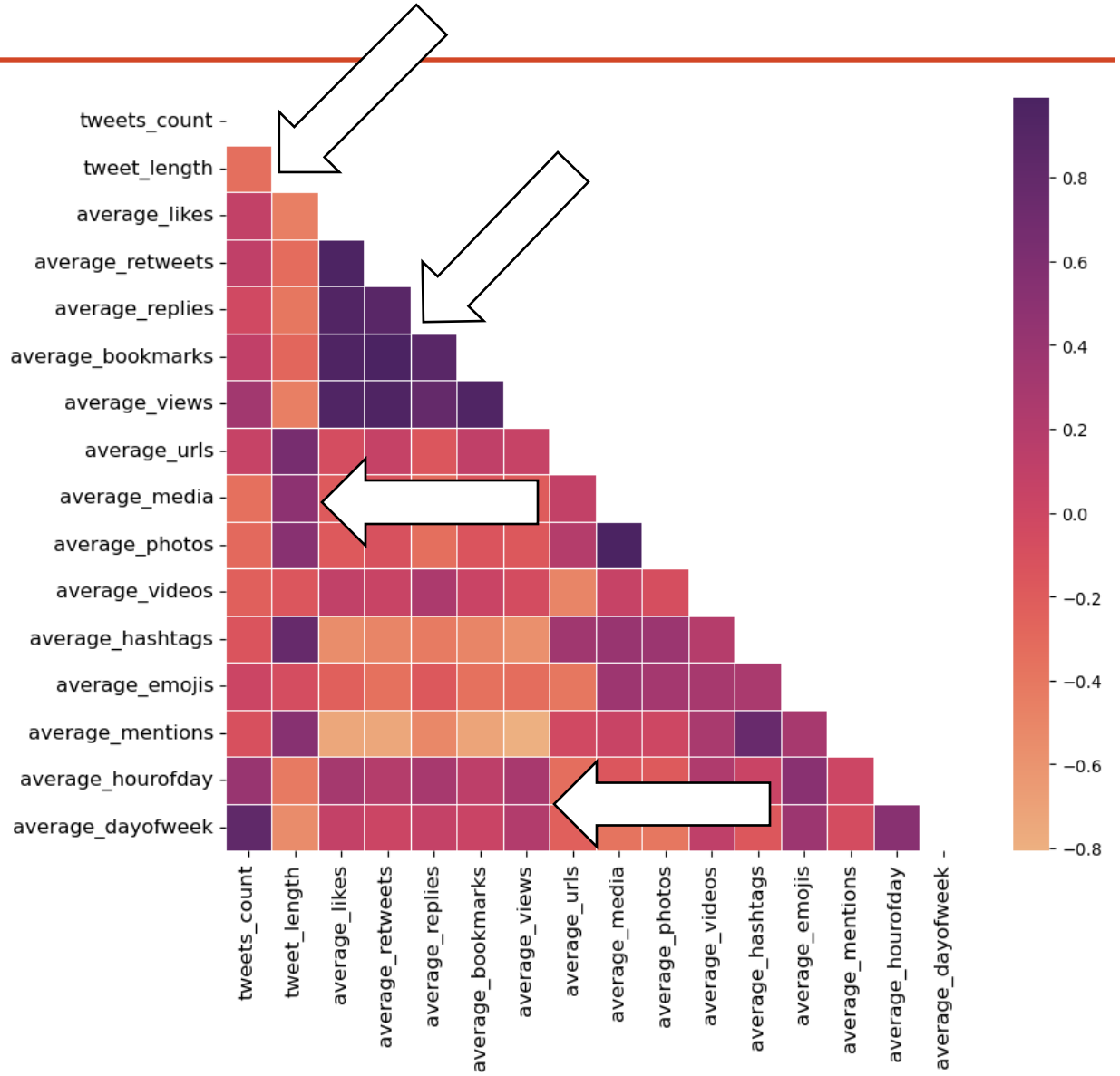




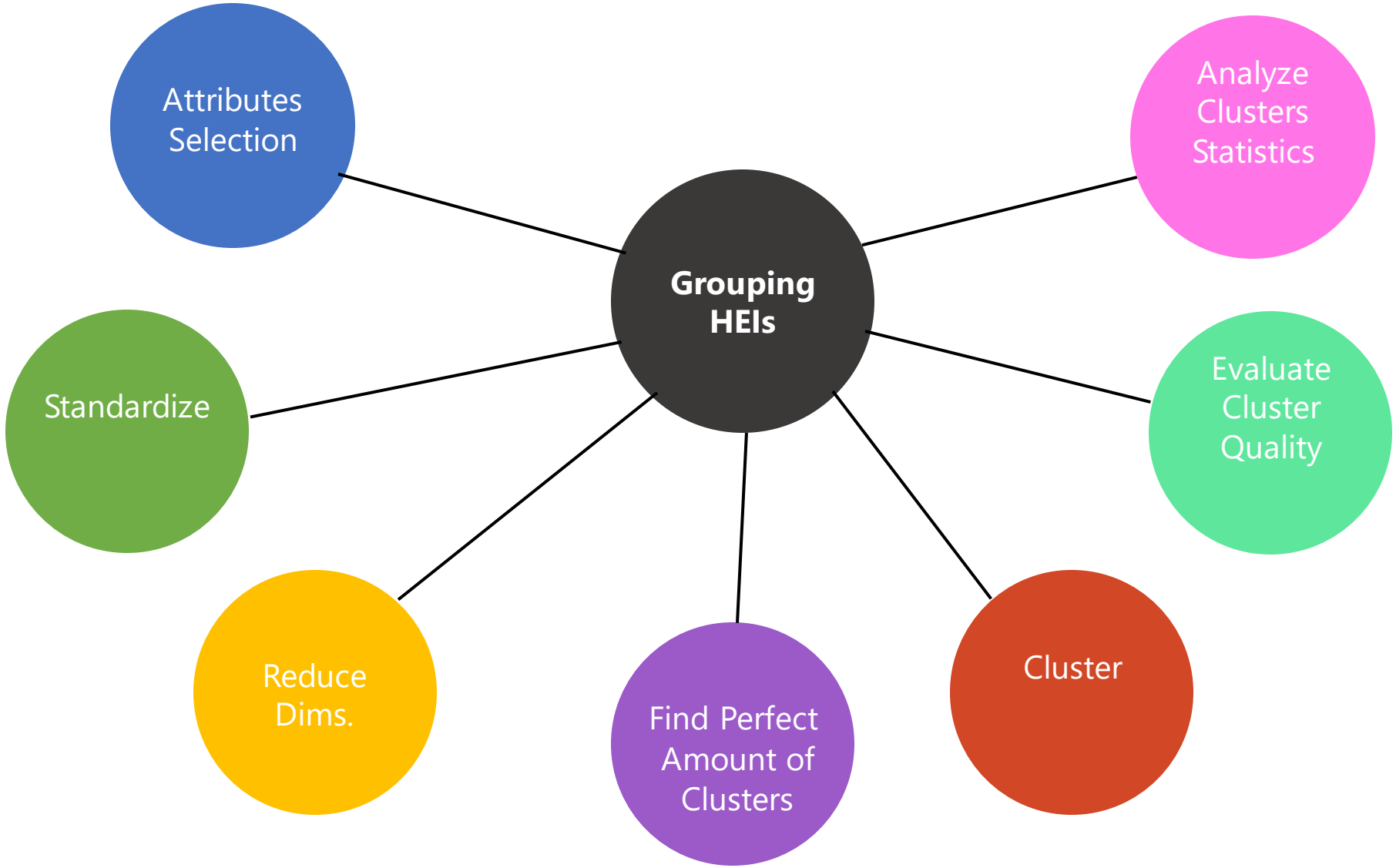
Part 3B: Bivariate Analysis

Correlations:

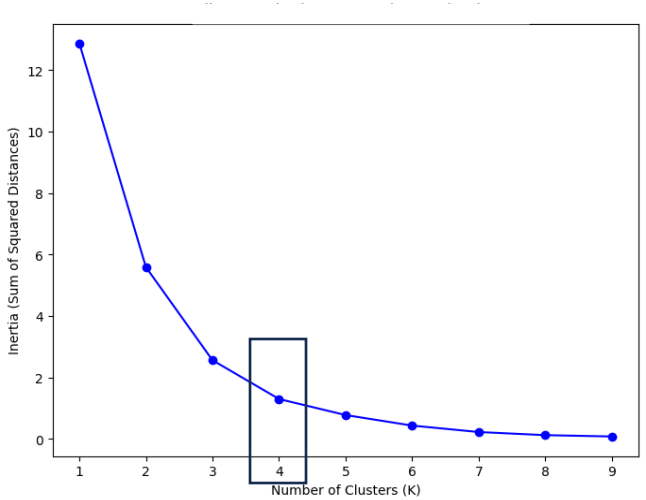
- "tweet_count" has moderate positive correlation to "average_views" and a moderate negative correlation with "tweet_length". So HEIs that tweet more often tend to get more views and write shorter tweets.
- "Engagement metrics" are strongly interrelated.
- Longer tweets with more content elements (URLs, media, photos) tend to receive less engagement.
- Hour of the day has a stronger correlation with engagement then day of the week.



Part 3C: Clustering Analysis



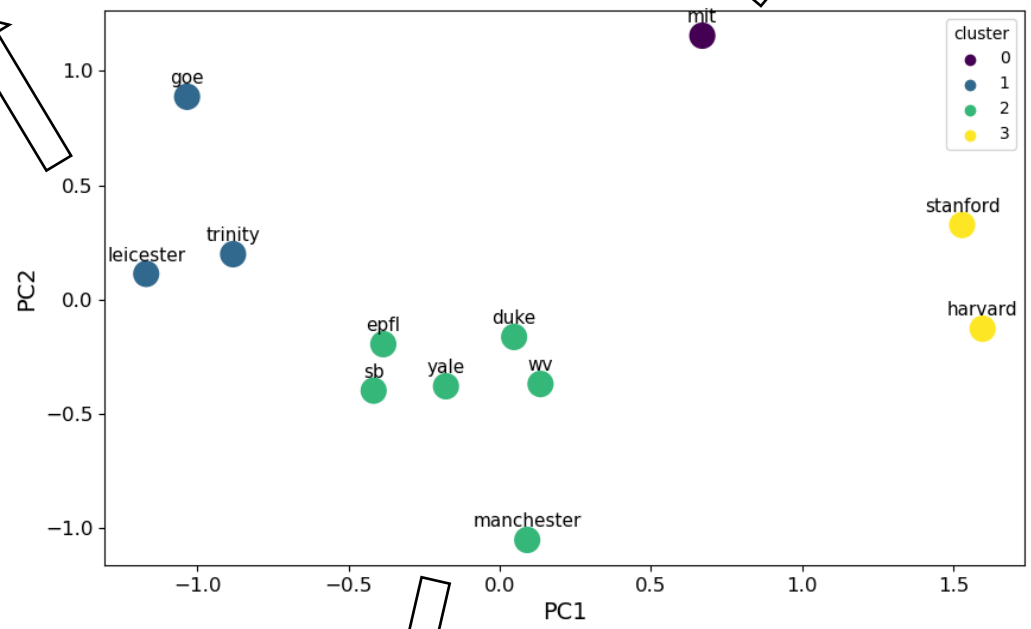
Part 3C: Clustering Analysis



Silhouette Coefficient: 0.542

Length, Hashtags, Emojis, Mentions, Day of Week, Amount, Views, Interactions.

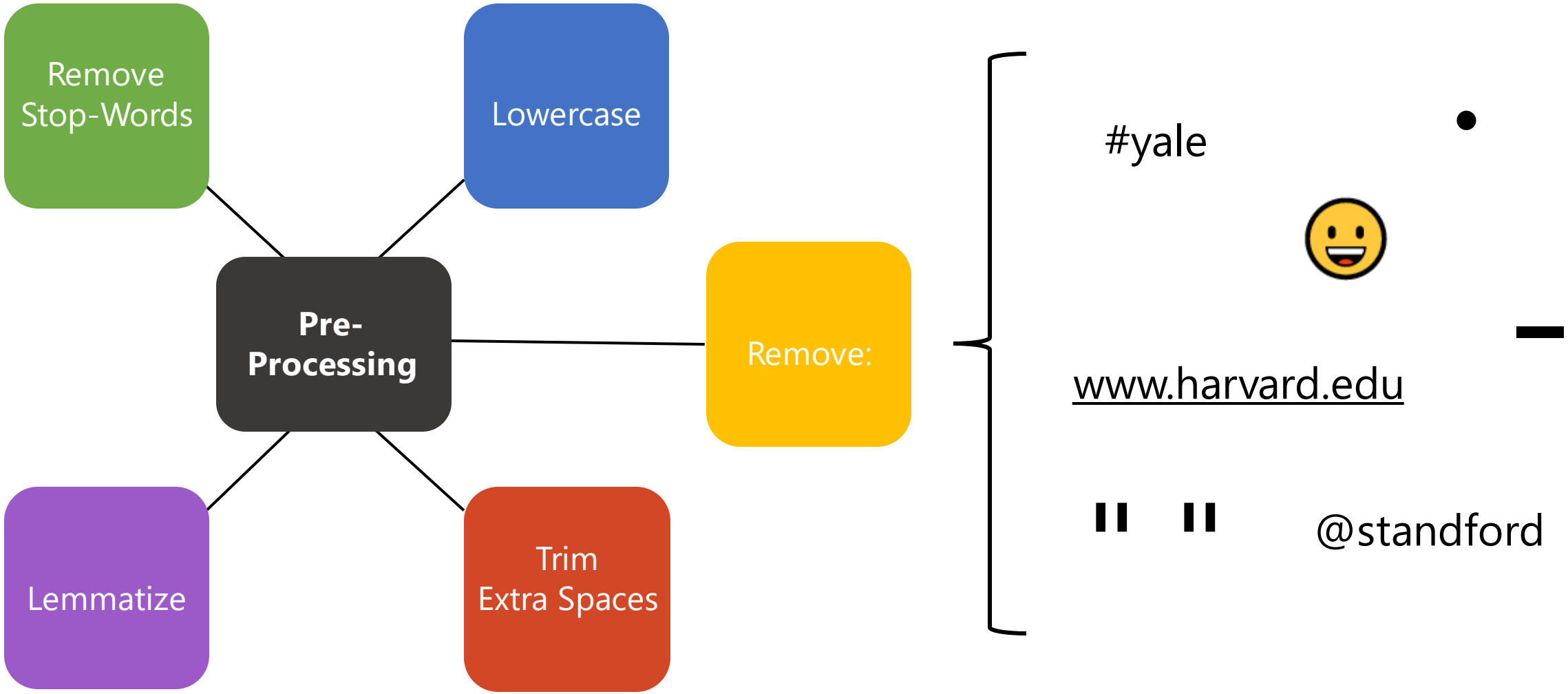
Media, Photos, URLs, Hour of Day, Videos, Hashtags, Emojis, Mentions.



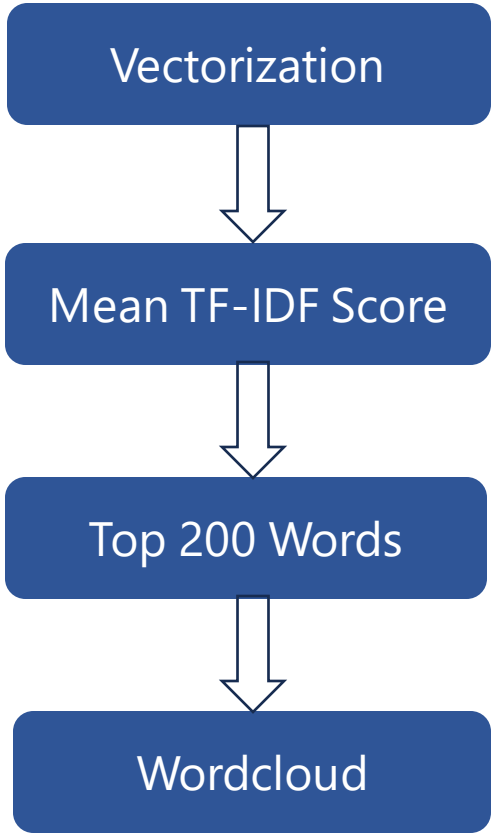
Amount, Views, Interactions, Mentions, URLs, Videos, Length, Media, Photos, Day of Week, Hour of Day.

URLs.

Part 4A: Pre-Processing



Part 4B: Top TF-IDF Words



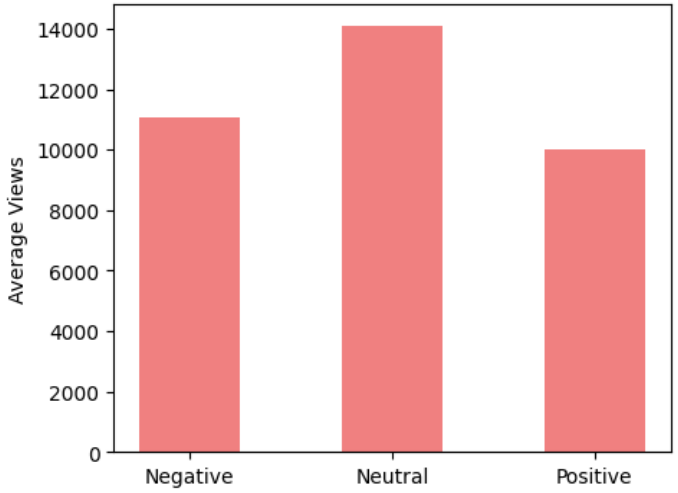
Part 4C: Sentiment Analysis

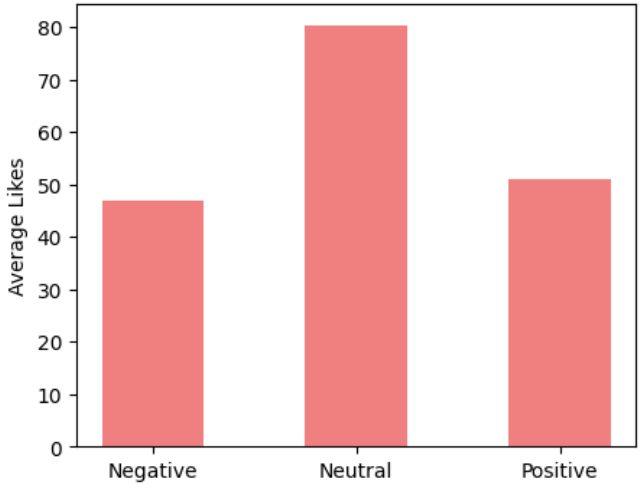
Sentiment Value

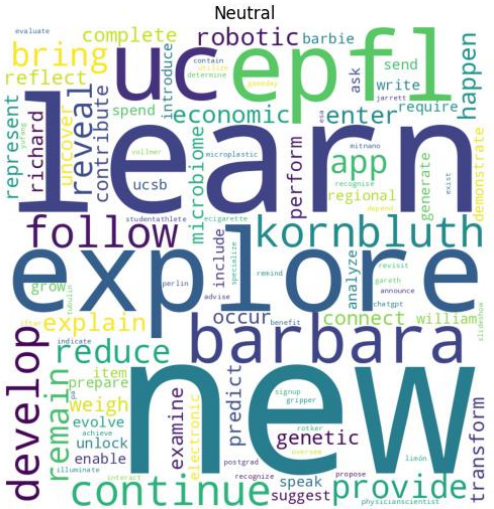
[-1, 1]

Sentiment

(Negative, Neutral, Positive)



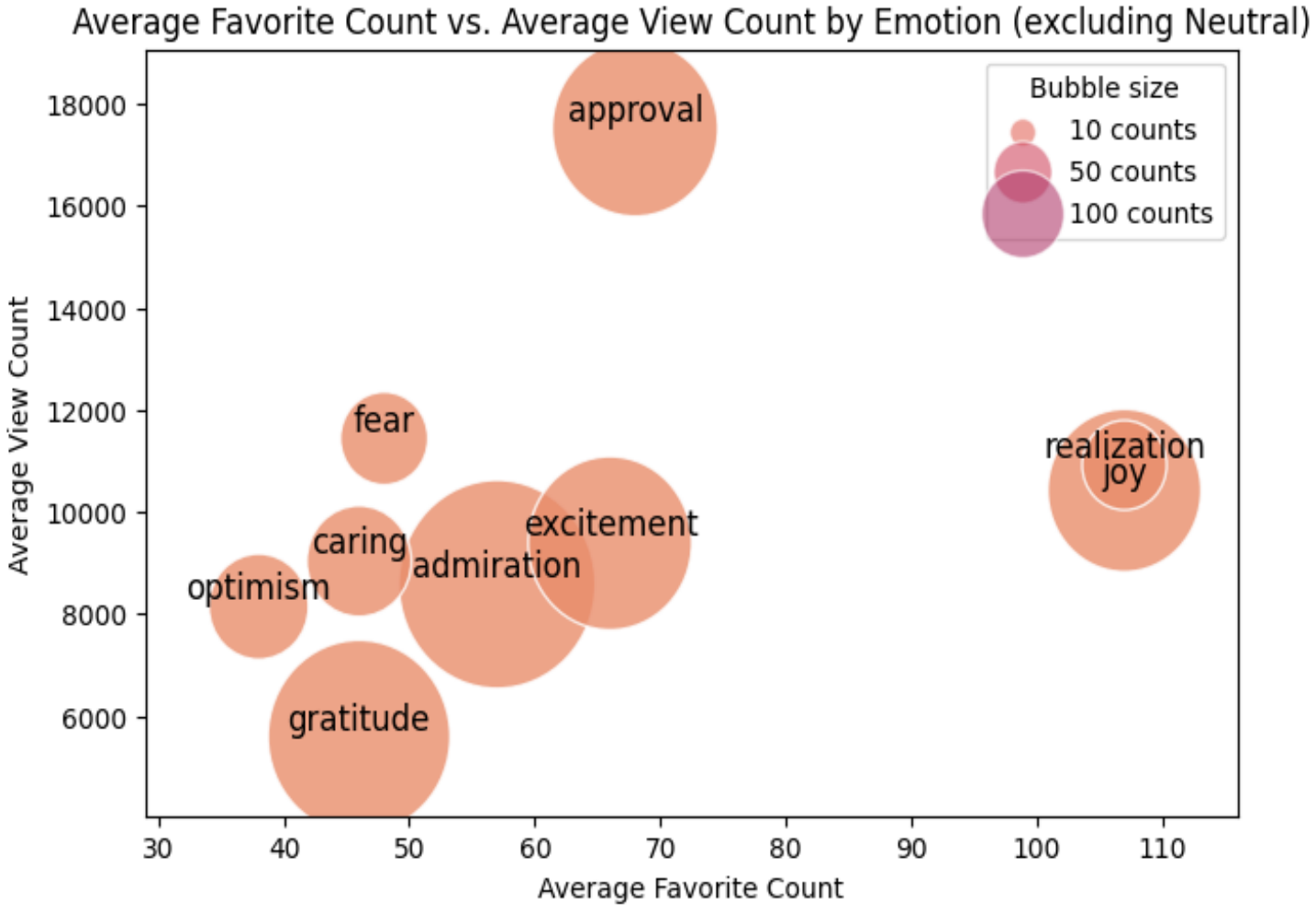




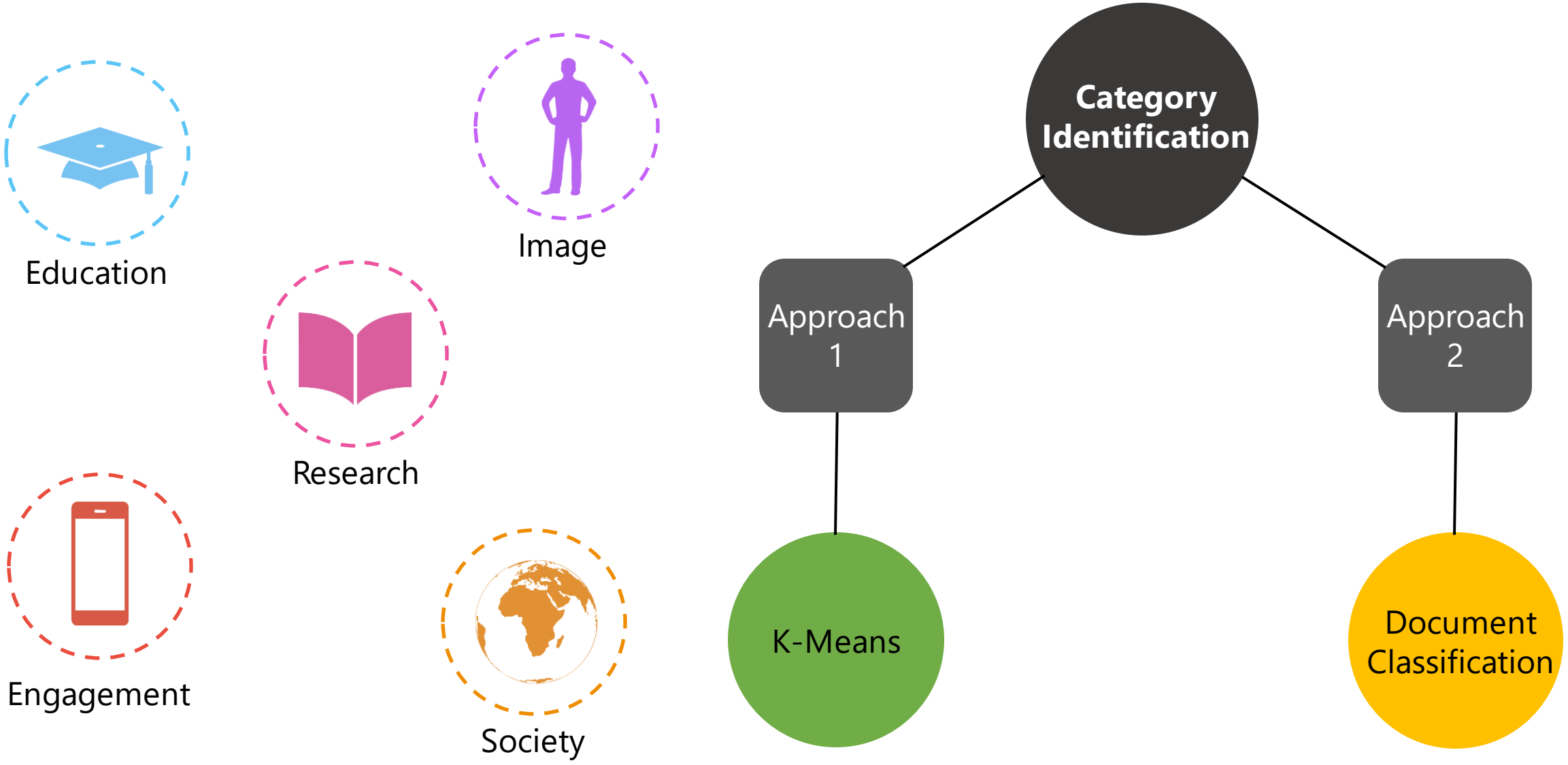
Part 4D: Emotion Recognition

Views and Likes by Emotion:

- The size of the bubble represents the volume of tweets for each emotion. 'Neutral' excluded.
- 'Gratitude' and 'Admiration', despite frequent have worse engagement metrics
- 'Approval' has the most average views.
- 'Realization', and 'Joy' have the highest likes-to-views ratios.



Part 4E: Category Identification



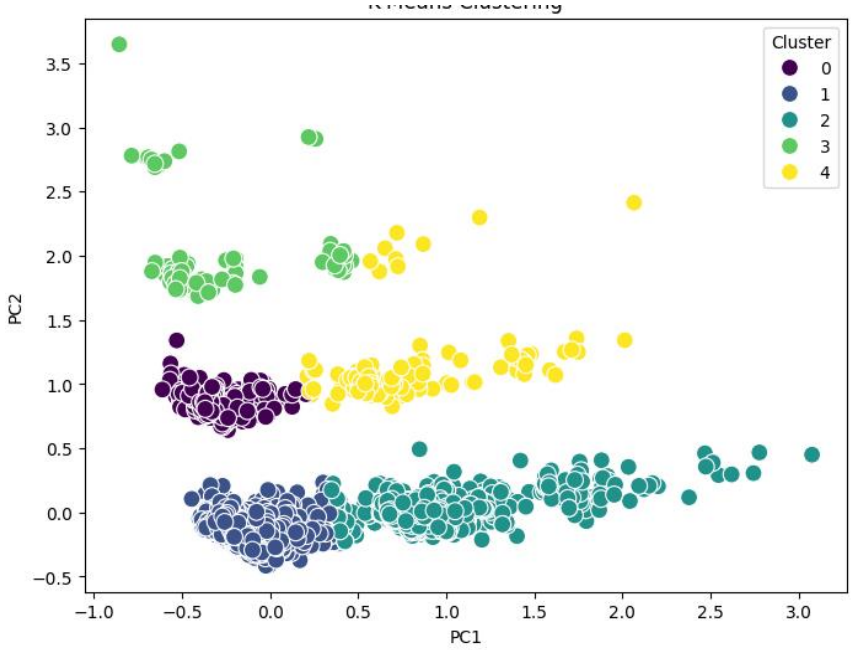
Part 4E: Category Identification (Approach 1)

Build Dataframe

Reduce Dims.

Clustering

Manual Assignment

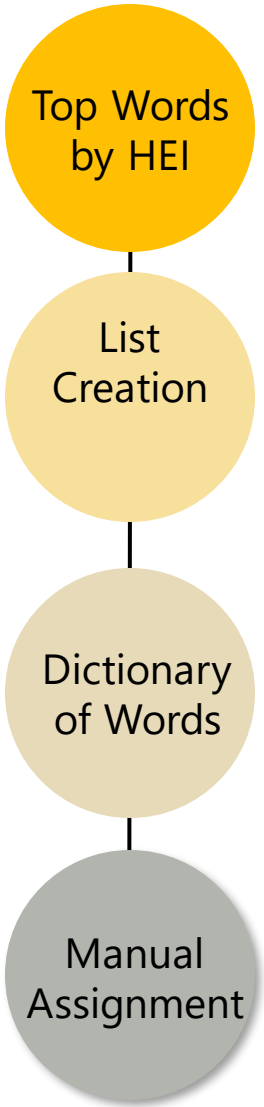


Silhouette Coefficient: 0.819

- **Sentiments and Emotions.**
- **Most Used Words.**
- **Random Tweets.**

Manual Assignment

Part 4E: Category Identification (Approach 2)



- **Top 20 Words by HEI**: Using TF-IDF.

- **Join All the Words in a List**: For simplicity.

- **Creation of a Dictionary of words**: Manually.

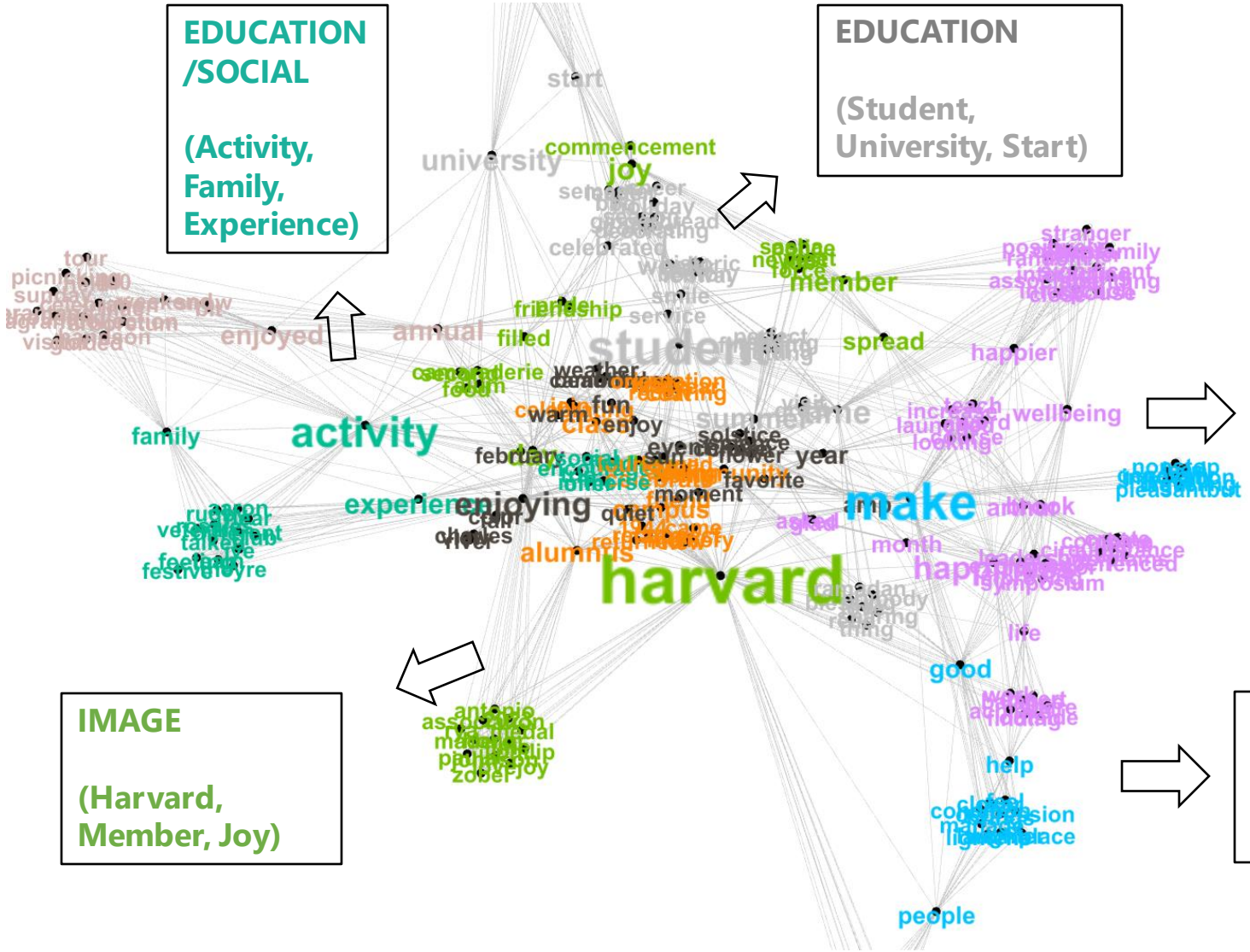
- **Assign the Tweets to Each Category**: Based on the most common words from a category, for each tweet.

- **Vectorization**: To build a data frame.

- **Top 100 Words by HEI.**
- **Load a pre-trained word vectors.**
- **Calculate similarity.**

Expansion of the Dictionary (Rule-based)

Part 4E: Category Identification (Approach 2 with Network)



EDUCATION /SOCIAL

(Activity, Family, Experience)

EDUCATION

(Student, University, Start)

SOCIAL/IMAGE

(Happier, Wellbeing, Life)

IMAGE

(Harvard, Member, Joy)

SOCIAL

(Make, Good, Help, People)

Network (Harvard):

- Co-occurrence of words within tweets.
- Communities: Modularity classes.

Part 5: Conclusion

- The project revealed posting trends and social media strategies of HEIs by exploring:
    - Engagement metrics (likes, retweets, replies, bookmarks, and views), the post's characteristics (count and length);
    - Types of content (URLs, media, photos, videos, hashtags, mentions, and emojis);
    - Time of posting (weekday and hour of the day);
    - Content of the texts (sentiment, emotions, top words and similarity of words).
- Category identification with different approaches to ensure a robust labeling of tweets.