Universität Münster

Social Media and Network Analysis: Exploration of Posts, Users, and Communities with Centrality Metrics, Sentiment Analysis, and Clustering

Descriptive Analysis – Basic Information

Data: Social media posts and network (friendship/following) relations
Time frame: October 31 0:00 – October 31 23:59, 2024
Activity: Highest at 3:00 and **lowest** at 18:00
Number of Posts: 70,260
Number of Users: 46,849
Avg. Degree of Users: 4.05
Language: English
Duplicates: High frequency of duplicates, especially by top users

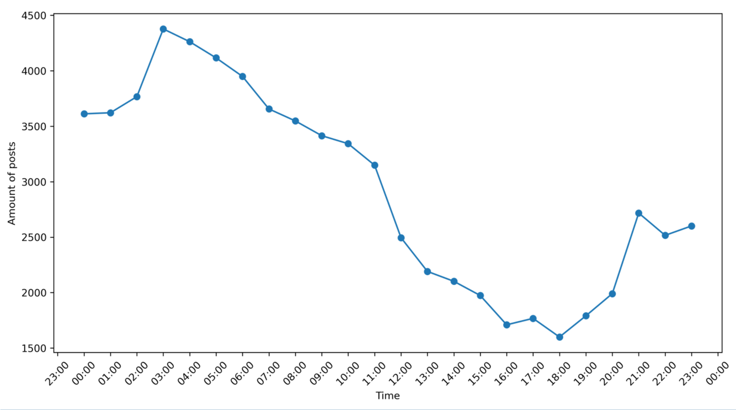

Figure 1: Number of posts per hour

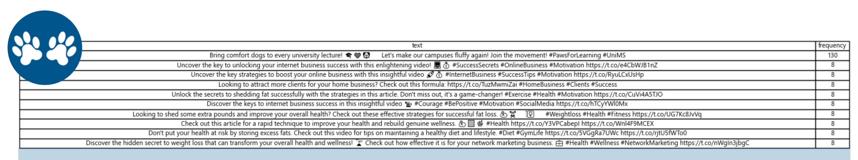

Figure 3: Most frequent duplicates

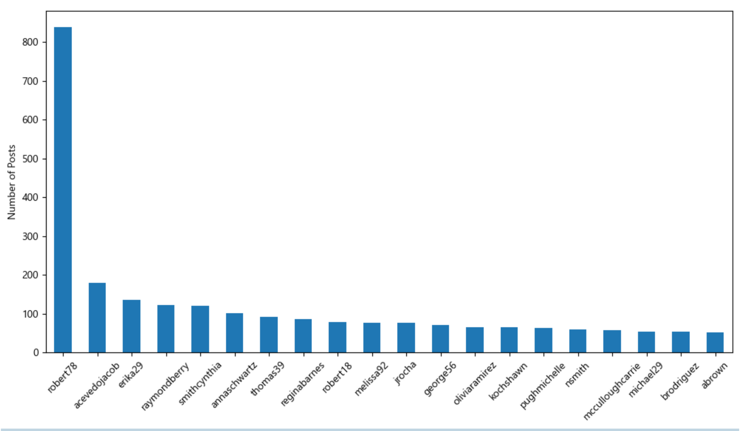

Figure 2: Top 10 Users with most posts across all hours

Network Analysis – Popularity of Users and Community Detection

1. Popularity Measured with Centrality Metrics

As a first step in exploring the social media network using **centrality metrics** to **assess user importance/popularity.**
Outlier: The user ***robert78*** dominates every metric due to 1700+ connections.

Degree: Measures **popularity based on direct connections** to other users
Eigenvector: Measures Popularity by **connection** to other **important users**
PageRank: Measures Popularity by the **incoming connection** from other **important users**
Closeness: Measures Popularity by **efficiently/fastly reaching** all other users
Betweenness: Popularity by the user acting as a **bridge** for other users

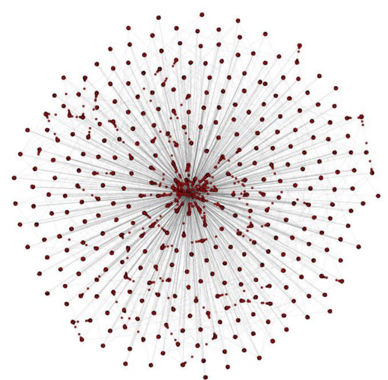

Figure 4: Plot with IGraph

User	Degree	User	Eigenvector	User	Betweenness	User	Page Rank	User	Closeness
robert78	0.03814	**robert78**	0.70588	**robert78**	1085964482	**robert78**	0.00736	**robert78**	0.30633
rharris	0.00113	matthew61	0.01893	john04	7686513	rharris	0.00024	ryan91	0.23646
davisjonathan	0.00105	davidcurry	0.01828	ryan91	6599323	reidelizabeth	0.00023	taylorjeremy	0.23623
reidelizabeth	0.00105	katherinejones	0.01821	charlesbuckley	6504327	davisjonathan	0.00022	karismith	0.23563
jenniferbenton	0.00098	khenry	0.01816	lejacqueline	6486049	edwardcabrera	0.00021	michael58	0.2356

Figure 5: Top 5 popular users based on popularity metrics

2. Community Detection with Louvain

The **Louvain approach** is a **community detection algorithm** that can be computed directly on the graph representation of the network. The algorithm consists of two steps:

1. Assign each node to be in its own cluster
2. Try to gain maximum modularity by relocating each node to the cluster of its neighbor

Results

Number of Communities: 245
Modularity Score: 0.96

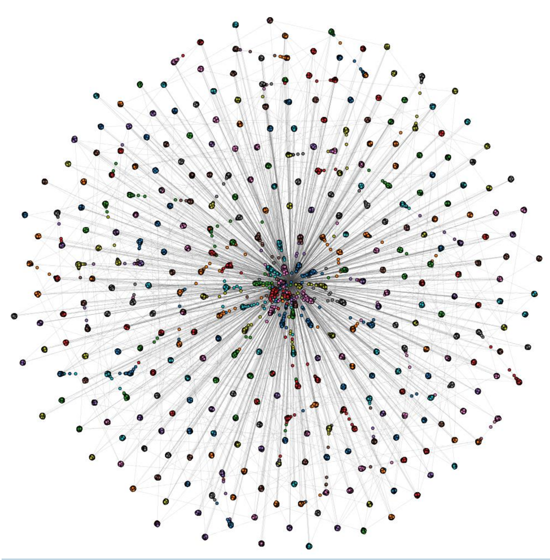

Figure 6: Community detection with Louvain

3. Community Detection with DBSCAN

Numeric vectors are necessary to use the **DBSCAN algorithm**. Consequently, the graph needs to be **converted into a vector representation:**

1. Convert Graph into Vector with Node2Vec (d=128)
2. Dimensionality Reduction with TruncatedSVD
3. K-Distance Plot
4. DBScan

Results

Number of Communities: 348
Silhouette Score: 0.46

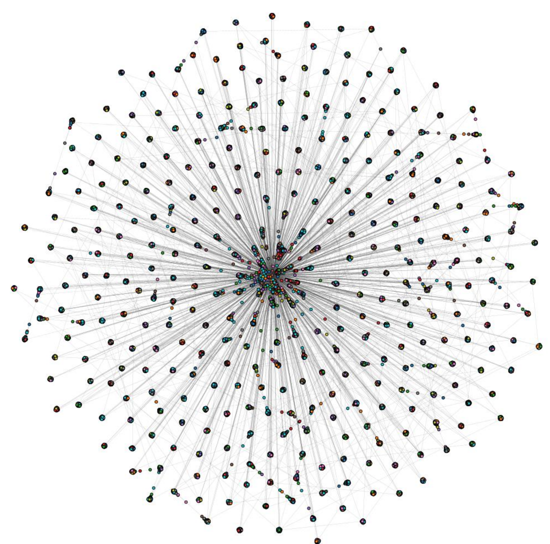

Figure 7: Community detection with DBSCAN

Outlook

- Handling of **outliers**, probably spam posts or bot users
- Examination of **homophily in the network** based on characteristics such as the topic clusters
- Speed of **information spreading** through the network

Sentiment Analysis – Text Mining with Pre-Trained Models

1. Overall Sentiment Distribution

An overview of **sentiment distribution** in the dataset shows that the majority of posts are positive, followed by neutral, and lastly negative.

But **which topics arise** on that day, and how are they **emotionally charged**?

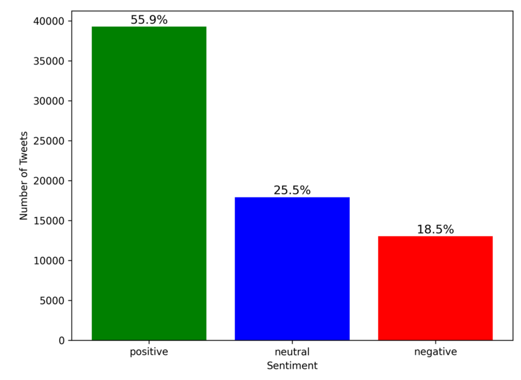

Figure 8: Overall sentiment distribution

2. Sentiment Distribution Across Topics

The ***news & social concern*** category has the **highest negative sentiment**. Social media often amplifies negativity in these areas, reinforcing 'bubbles'.

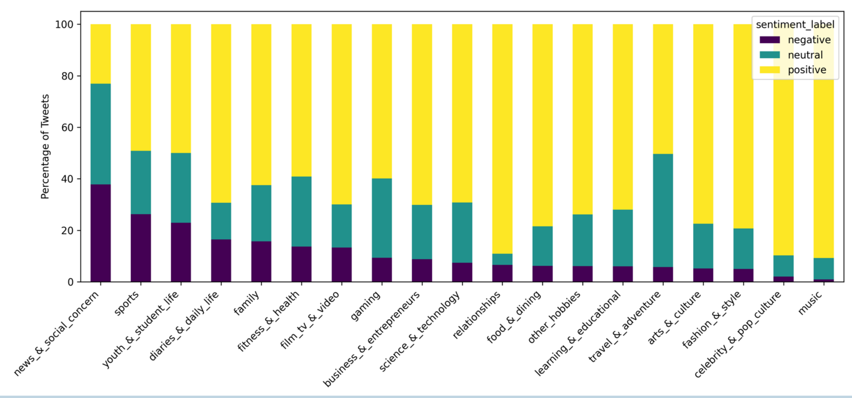

Figure 9: Sentiment distribution across all topics in %

3.

“ Do users who predominantly post **tweets with negative sentiment** tend to **form denser clusters** in the social network? ”

Pearson Correlation

- Correlation = 0.0601
- p-value = 0.2936

No, because the **distribution** between low and high Negative Sentiment Ratio data points is **equal** and correlation is **low** and **non-significant**.

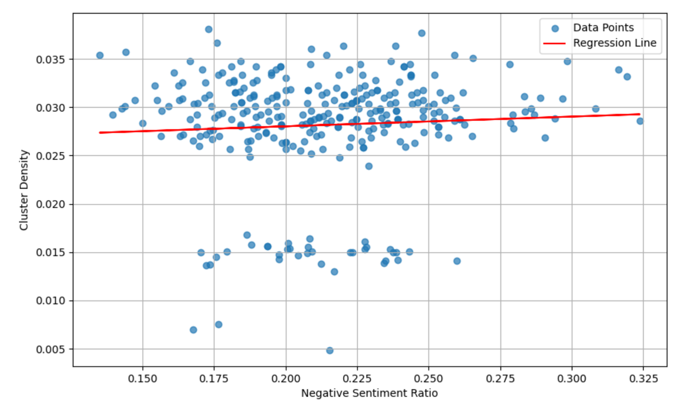

Figure 10: Negative Sentiment Ratio and Cluster Density

Thematic Clusters and User Behavior Analysis – Dimensionality Reduction and Clustering

1. Text Vectorization with TF-IDF

To uncover thematic clusters **through data dimensionality reduction**, the **top 500 terms** are retained (elbow method).

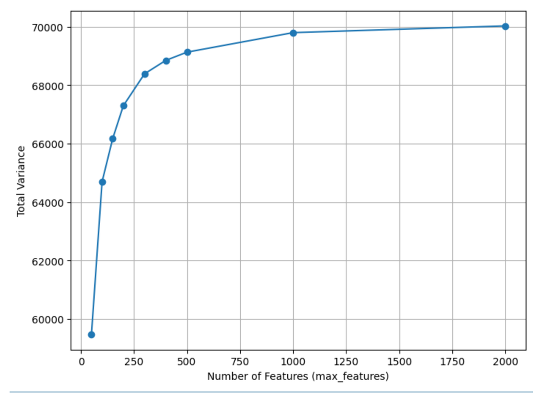

Figure 11: Total Variance and Number of Features

2. Dimensionality Reduction with TruncatedSVD

Dimensionality reduction **requires correlation**, with some feature pairs showing strong links, indicating thematic overlaps in the text.

The given text data is **highly distributed**: No clear elbow point or optimal reduction to n dimensions

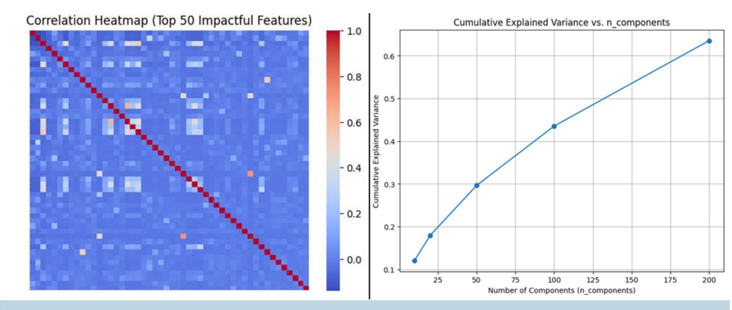

Figure 12: Correlation Heatmap and text data distribution

3. Clustering with KMeans

Evaluate clustering performance across dimensions and cluster sizes and use the **silhouette score** to assess the quality of clusters.

FOCUS

	Coarse Binary Cluster Distinction	Finer-grained Cluster Distinctions
n-components	12	8
n-clusters	2	9
Silhoutte Score	0.56	0.47

Figure 13: Best configured MiniBatchKMeans-Model

4. Cluster Visualization and Key Results

Preprocessing: Identified **10 key terms per cluster**, assigned posts to clusters, and analyzed **user activity by post count**.

Cluster-Model: The finer-grained cluster distinctions model was chosen because this represents distinct themes supporting targeted analyses.

Results

Most clusters are led **by low-activity users**, except **niche clusters** where **a few high-activity users** (e.g., influencers) drive the conversation.

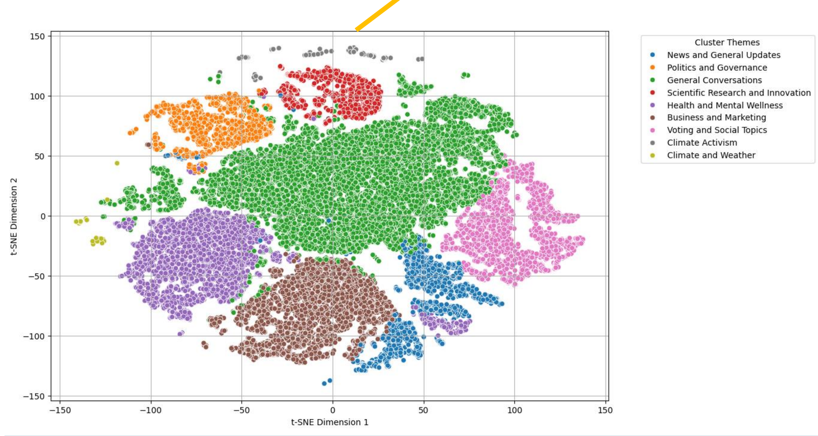

Figure 14: Finer-grained cluster distinctions and its cluster themes visualized as t-SNE Plot

living.knowledge

Group F: Dimitrij Eli, Kevin Martin Schulz, Marc David Sutjipto, Piet Vinnbruck

CSSSA Computational Social Science and Systems Analysis