https://github.com/ulalud/UDA_FinalProject_Ludwiczak

Urszula Ludwiczak

Sentiment and Conversational Trends among Reddit users regarding South Korea

Analysis of the subset of 'self-contained' posts

The objective of this project was to analyze Reddit post titles mentioning South Korea and their associated comments, retrieved using the Reddit API. The analysis focused specifically on comments related to the recently lifted martial law. The first key task involved identifying the ten most common words across all post titles following pre-processing. Next, single linkage hierarchical clustering was applied to these titles, resulting in the identification of eight distinct clusters. The second major task focused on conducting sentiment analysis of unlabeled comments containing the term 'martial' using the VADER sentiment analysis tool. The normalized, weighted composite sentiment scores were then visualized over time.

The program was developed on a machine with an Apple M3 chip, with a system configured with 16GB RAM, 256GB SSD , 8-core CPU with 4 performance cores and 4 efficiency cores, 2 GB of memory per each CPU, and core 8-core GPU. The program was developed using Python 3.11, with Jupiter Notebook as the primary editor. The program runs in 0.92 seconds excluding import of libraries and fetching the data from Reddit API.

Data Selection

To conduct Natural Language Processing (NLP), post topics were retrieved from Reddit open access API, using a Python library Python Edit API wrapper (PRAW). Fetching the original post URL is not possible, if an external redirecting link is provided, hence the number of available posts narrowed down to 49 ‘self-contained’ post. The author's Reddit credentials have been removed from the Jupyter Notebook, but to reproduce the results, all the fetched data has been saved in additional files. Each title contains ‘South Korea’ as wel as ‘South Korean’. According to Wikipedia (*Wikipedia Article “Koreans,”* n.d.) ‘Koreans are an East Asian ethnic group and nation native to Korea.’, thus to avoid combining results from both states, the keyword ‘Koreans’ was not included

in the search. The articles were posted between 2014-07-13 and 2024-12-04. The data comprise such variables as post title, text, core, number of comments, URL address, and time. The posts were saved as CSV under ' reddit_posts.csv'.

To conduct sentiment analysis on comments under posts referring to the martial low in South Korea, which was introduced and lifted shortly thereafter, 140 comments were fetched and saved as both a DataFrame and a CSV file under 'reddit_martial_law_comments'. The data frame consists of two variables- comment body and timestamp.

Most common topics and post clustering

Data Pre-Processing

As part of the pre-processing, punctuation and numbers were removed from the titles. Additionally, all letters were converted to lower-case. 'North Korea' and 'South Korea' and their conjugated forms were converted to expressions with an underscore 'north_korea' and 'south_korea' respectively. All mentions of 'korea', 'korean', and 'south_korea' itself were removed from the titles in order to find more meaningful common subjects in the later stage of this analysis. Such titles were subsequently tokenized, using 'split' Python built-in function, and stemmed with Porter Stemmer.

Finding ten most common words in the document

For simplicity, as word weight is not required for finding common keywords or clustering, CountVectorizer was used on the tidy titles to further tokenize the data and transform them into document-term-matrix. During this process, 10 features were retained, and English stop-wors were removed. Subsequently, the total frequency of each word across all documents was calculated and the top ten most common phrases were identified as:

- north_korea, Frequency: 9;
- law, Frequency: 7;
- martial, Frequency: 7;
- group, Frequency: 5;
- clone, Frequency: 4;
- declar, Frequency: 4;
- kim, Frequency: 4;
- war, Frequency: 4;
- autist, Frequency: 4;
- match, Frequency: 4.

Ward hierarchical clustering of posts' topics

The analysis employed the Agglomerative Hierarchical Clustering Algorithm to group documents into clusters based on their pairwise cosine similarity. The data preprocessing pipeline included tokenization using the TF-IDF Vectorizer, a method that assigns weights to terms based on their frequency in a document and their inverse frequency across the corpus. This weighting scheme ensures that the clustering process focuses on the most relevant terms, thereby improving the quality of the clusters. To evaluate the clustering performance, the silhouette score was used, which indicated the highest performance for tokenized data without applying stemming. The number of features was set to 10, and English stop words were removed to reduce noise and enhance interpretability. Cosine distance was used to measure similarity, because it captures the orientation of document vectors rather than their magnitude (such as document length), making it ideal for text data where relative term importance is crucial.

At first, Single Linkage Clustering was tested. The results were unsatisfactory- a few too many clusters were identified. The method struggled with the sparse representation of the data. Moreover, single linkage is highly sensitive to outliers- documents or text samples that significantly differ from the majority of the dataset. Having regard to the above, Ward's Clustering was selected due to its computational efficiency and ability to minimize the variance within clusters, making it suitable for high-dimensional data like document embeddings. Ward linkage clustering calculates Euclidean distances in the process, however it can take raw feature data as input. The new distance between the newly formed cluster u and each v is calculated according to formula (1).

(1)

$$d(u,v) = \sqrt{\frac{|u|+|s|}{T} d(v,s)^2 + \frac{|v|+|t|}{T} d(v,t)^2 - \frac{|v|}{T} d(s,t)^2}$$

, where d is an Euclidean distance, *u* is the newly joined cluster consisting of clusters *s* and *t*, *v* is an unused cluster in the forest, T = |v| + |s| + |t|, and |*| is the cardinality of its argument. This is also known as the incremental algorithm (SciPy).

Therefore, the similarity between clusters was defined as the increase in the sum of squared distances from each point to the centroid of its cluster when the clusters are merged, as illustrated in Figure 1.

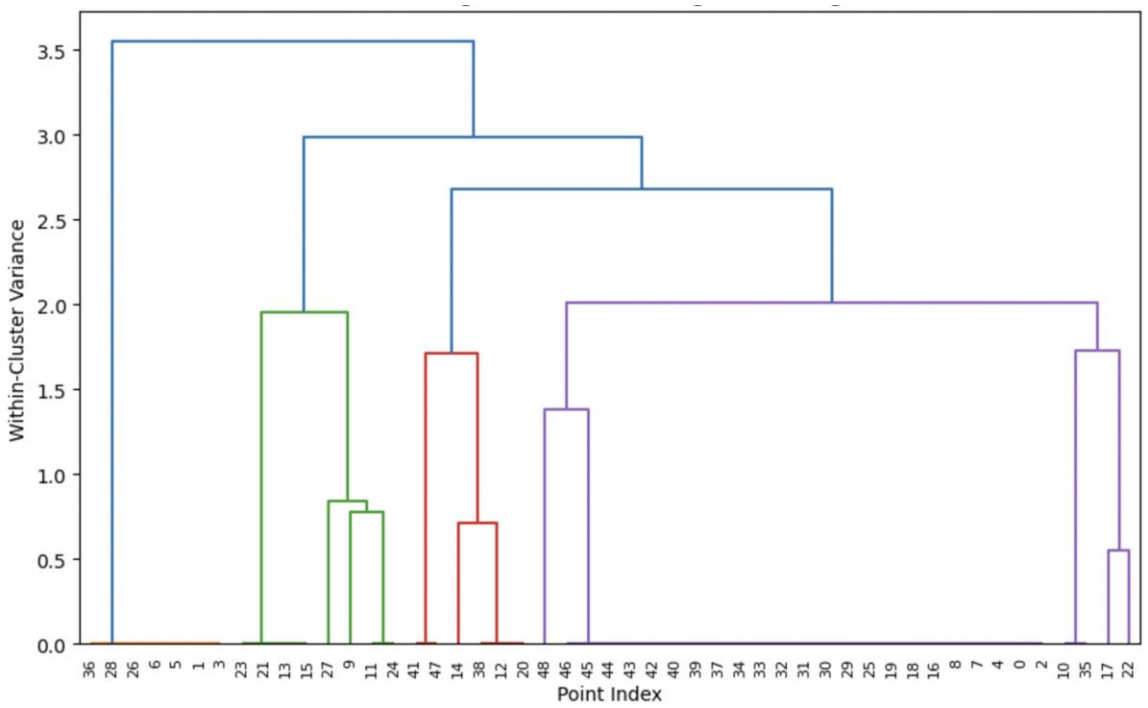

Figure 1: Dendrogram for Ward Clustering

At a within-cluster variance threshold of 1.5, the algorithm identified 8 clusters with a silhouette score of 0.86, indicating very good separation between clusters. It was the highest silhouette score with a moderate number of clusters, hence t =1.5 was identified as an optimal cutoff. Sample documents from two first clusters (C1 and C2) are provided to illustrate the cluster content and coherence. A visual inspection can be conducted to determine the most common topics among the clusters:

Titles with label 1:

C1 = ['SOUTH KOREAN MARTIAL LAW MEGATHREAD!', 'South Korean stocks down 6% today on South Korea declaring martial law, buy the dip', 'MT: South Korea Martial Law ', 'Emergency Martial Law declared in South Korea', 'South Korean Parliament overturns Martial Law decree', 'South Korea invokes Martial Law - MDs on strike specifically told to return to work or suffer sonsequences', 'We can get prepared Trump will try to follow this: "South Korea\'s president declares martial law"']

Titles with label 2:

C2 = ['North Korean Defector Who is Sending Information to North Korea', ' Japan publicly claims at UN, "There\'s no basis for comfort women claims"... South Korean representative remains silent, while the North Korean representative engages in heated debate', 'Why don’t North Korean border guards just immediately cross the border into South Korea?', 'My grandfather was born in North Korea but escaped to the South, ask him anything!']

Sentiment Analysis of comments

The final part of the analysis was focused on comments under posts referring to the newly introduced (and already lifted) martial law in South Korea. 140 comments were fetched from the Reddit, together with the time (GTM) of their submission and they were saved as both a DataFrame and CSV file under 'reddit_martial_law_comments'. The time range of the analyzed comments is from 03 Dec 2024 17:05:15 to 06 Dec 2024 03:10:0, which suggests that the topic has been discussed only recently. It is expected, as the political discussions and about martial law in South Korea, followed by introduction of the new law took place on the 3rd of December 2024 ("Fear, Fury and Triumph: Six Hours That Shook South Korea," 2024).

Data Pre-Processing

Similarly to the post titles, the comments were prepared for the analysis by removing punctuation signs and converting all the letters into lower-case.

At first, TextBlob classifier was used to conduct the sentiment analysis of the comments. It returns several parameters, such as intensity, subjectivity and polarity of a sentence, out of which only the latter one was used for the analysis. Polarity lies between [-1,1], where -1 defines a negative sentiment and 1 defines a positive sentiment. Negation words reverse the polarity. 66 comments were classified as positive, and 34 as negative.

Subsequently, VADER (Hutto & Gilbert, 2014) sentiment analysis tool was leveraged. It is a lexicon-based approach, where a sentiment is defined by its semantic orientation and the intensity of each word in the sentence. This required to download a pre-defined dictionary classifying negative and positive words. The compound score was computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). This is the most useful metric to measure sentiment for a given comment. As a result, 48 comments were classified as positive, and 59 as negative.

Due to the absence of a labeled dataset, no validation of the sentiment analysis can be conducted. The decision was made to proceed with VADER for sentiment analysis, as it is highly optimized for social media text. The results obtained using VADER are presented over time in Figure 2. To enhance readability, the data was smoothed using a moving average window of size 10. The sentiment experiences prolonged negative values on the 3rd of December between approx. 22:00 and 23:00 GTM, which was when the decision about the introduction of the new law was made. The graph indicates that sentiment normalizes towards the end of the analyzed period, potentially reflecting the National Assembly's decision to lift martial law at 1:00 GMT on December 4th ("Fear, Fury and Triumph: Six Hours That Shook South Korea," 2024).

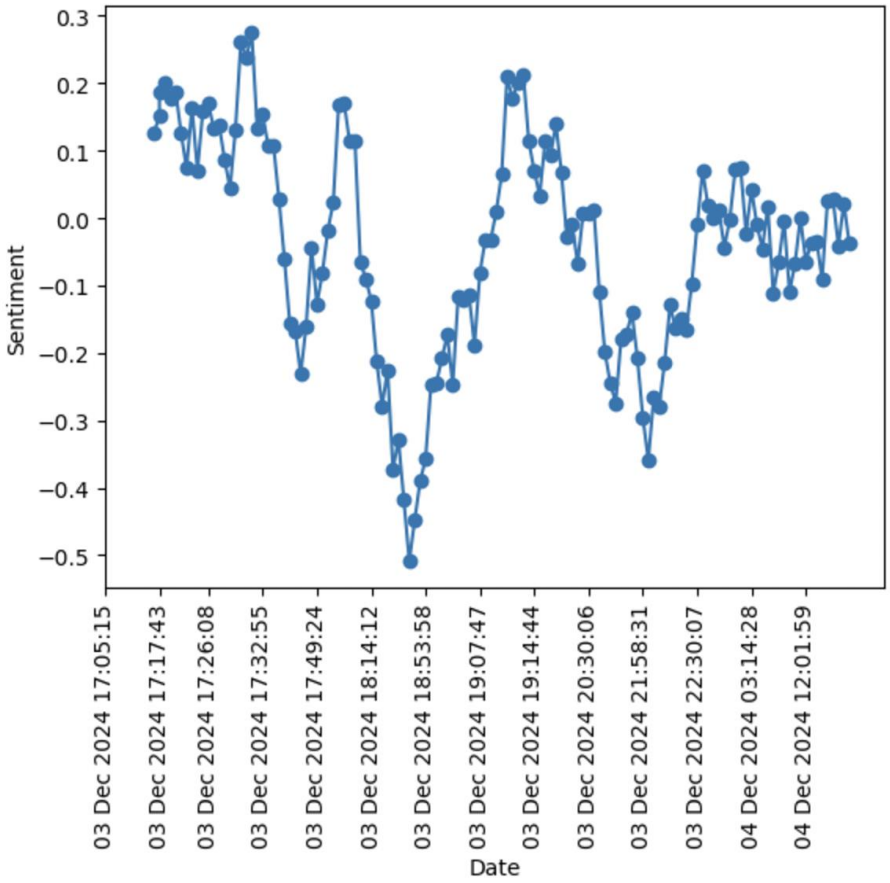

Figure 2: Sentiment over time

Conclusions

The analysis of post topics provided valuable insights into the general tone and sentiment of the dataset, identifying the ten most common themes. Among these, the most frequent topics were related to 'North Korea' (and its conjugated forms), 'law,' and 'martial.' Posts were further grouped into eight distinct clusters using Ward linkage hierarchical clustering. A very high silhouette score of 0.86 was obtained, revealing thematic patterns in the data. Hence, minimization of variance in the clusters occurred to be the most efficient method for clustering of this corpus.

The sentiment analysis of comments under the posts about martial law provided valuable insights. Using the VADER sentiment analysis tool, 18 fewer positive and 25 more negative comments were identified compared to TextBlob, leading to an overall more negative sentiment. VADER was deemed more appropriate tool due to its training on social media datasets. As expected, the majority of comments were neutral (33/140) or negative (59/140), reflecting the controversial nature of martial law.

Further exploratory analysis is recommended to validate and enhance these findings. Specifically, reviewing a small subset of the comments can help establish a rudimentary benchmark, enabling a more accurate assessment of whether TextBlob's and VADER's outputs align with human expectations. Incorporating labeled data or could refine the sentiment analysis and validate the resuls.

References

Fear, fury and triumph: Six hours that shook South Korea. (2024, December 3). *BBC News*.

Hutto, C. J., & Gilbert, E. E. (2014). *Vader: A parsimonious rule-based model for sentiment analysis of social media text*. Ann Arbor, MI,.

SciPy. (n.d.). *scipy.cluster.hierarchy.linkage*. SciPy. Retrieved December 13, 2024, from https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html

Wikipedia article "Koreans." (n.d.). Retrieved December 12, 2024, from https://en.wikipedia.org/wiki/Koreans

I have worked independently.