Master’s research plan seminar

Structural Topic Modeling and Sentiment Analysis of X (Twitter) Data to Quantitatively Appraise Global Public Opinion

NGUYEN Viet Hoa
m5281033

Supervisor: Prof. LE Doan Hoang
University of Aizu

To Advance Knowledge for Humanity
THE UNIVERSITY OF AIZU
· 1993 ·

2024 Jul 23th

Content

1 Introduction

2 Existing Technique and Its Changllenges

3 Ideas for Improvement and Proposed Solution

4 Model Framework

5 Case Study and Initial Result

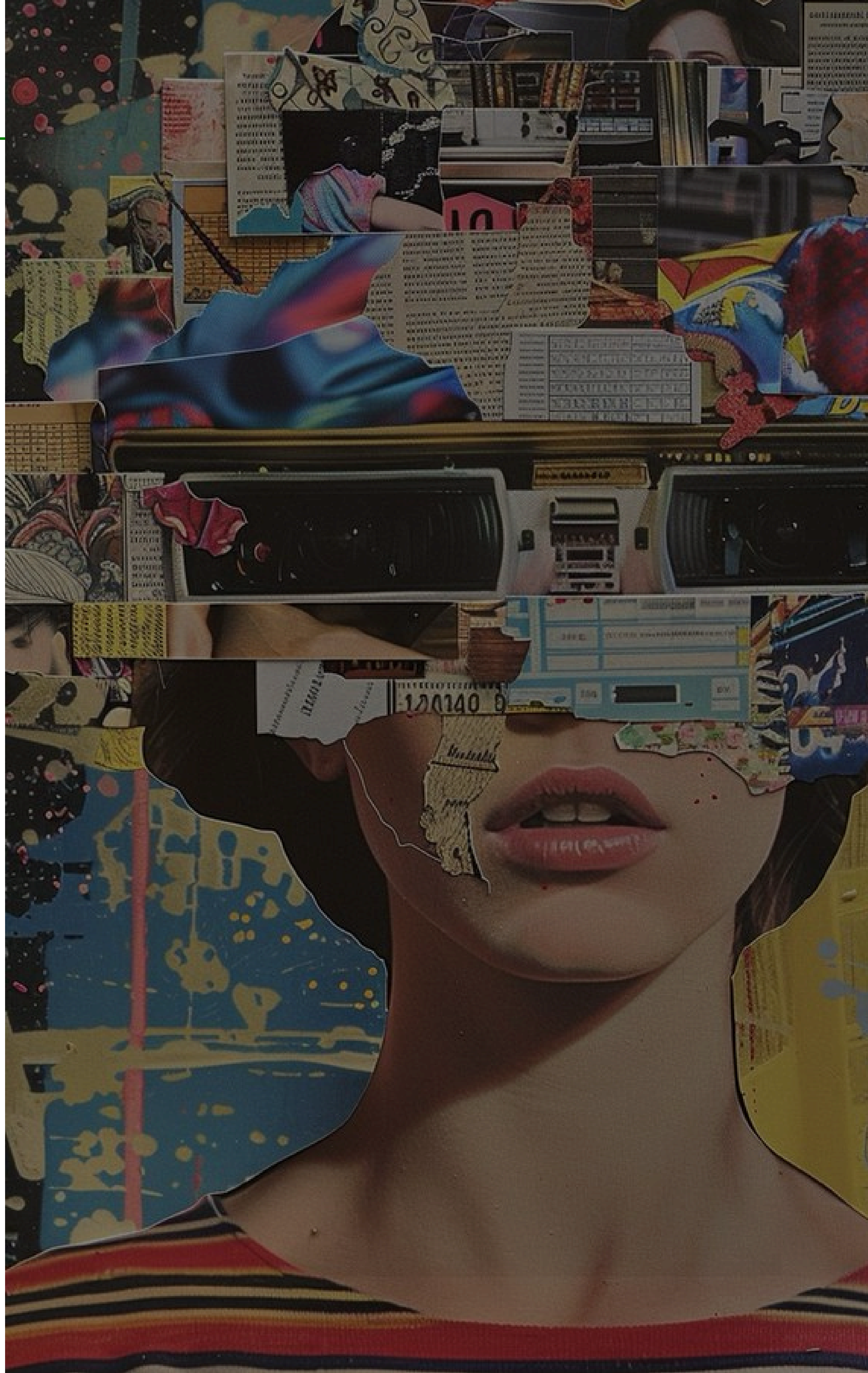
Introduction

The Glosbal Use of Twitter:

- X (Twitter) has become a bustling global forum where individuals, communities, and even nations share their thoughts, opinions, and concerns.
- This real-time stream of consciousness presents an unprecedented opportunity to understand of global public opinion, or in other words, "WHAT DO PEOPLE THINK?"

Reseach to Quantify Public Opinion

- Researchers and organizations have recognized the value of Twitter data and have employed various techniques to analyze and quantify public sentiment.



Research Approaches for Quantify Public Opinion and Sentiment Analysis

1

Rule-based Systems

Model rely on handcrafted rules that utilize lexicons, dictionary and other linguistic features

2

Machine Learning-based Systems

Utilize algorithms that learn to classify sentiment from labeled training data

3

Hybrid Systems

Combine the strengths of both rule-based and machine learning approaches

Existing Techniques for Analyzing Public Opinion on Twitter and Its Challenges

1. Lexicon-Based Sentiment Analysis [1]:

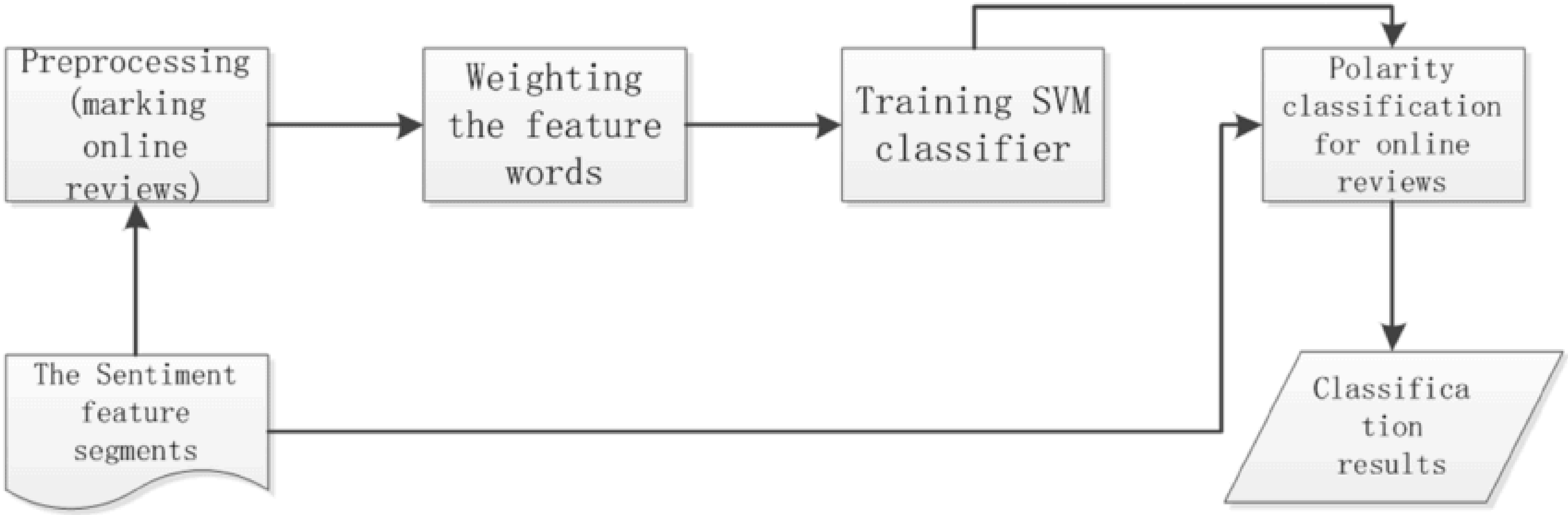
- **Mechanism:** Relies on pre-defined dictionaries of words and their associated sentiment scores (e.g., positive,negative). Tweets are analyzed by matching words against these dictionaries and aggregating their scores.
- **Advantages:** Simple, fast, and requires no training data.
- **Limitations:** Struggles with context, sarcasm, negation, and evolving language.

[1] S. K. Tripathi et al., "Sentiment Analysis of Twitter Data," IJET, 2020.

Existing Techniques for Analyzing Public Opinion on Twitter and Its Challenges

2. Supervised Machine Learning [2]:

- **Mechanism:** Algorithms (e.g., Naive Bayes, Support Vector Machines) learn to classify tweet sentiment from labeled training data.
- **Advantages:** Can capture complex patterns and adapt to new language if retrained.
- **Limitations:** Requires substantial labeled data, which is costly and time-consuming to obtain. May overfit to training data and not generalize well to unseen examples.


Preprocessing (marking online reviews)
Weighting the feature words
Training SVM classifier
Polarity classification for online reviews
The Sentiment feature segments
Classifica tion results

[2] Xia, H., Yang, Y., Pan, X., & An, W. "Sentiment analysis for online reviews using conditional random fields and support vector machines," 2020.

Existing Techniques for Analyzing Public Opinion on Twitter and Its Challenges

3. Topic Modeling [3]:

- **Mechanism:** Statistical models like Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF) identify latent topics within a corpus of tweets.
- **Advantages:** Uncovers underlying themes and discussions, can handle large-scale data.
- **Limitations:** Often assumes topic independence, doesn't directly incorporate sentiment analysis, and topic interpretation can be subjective.

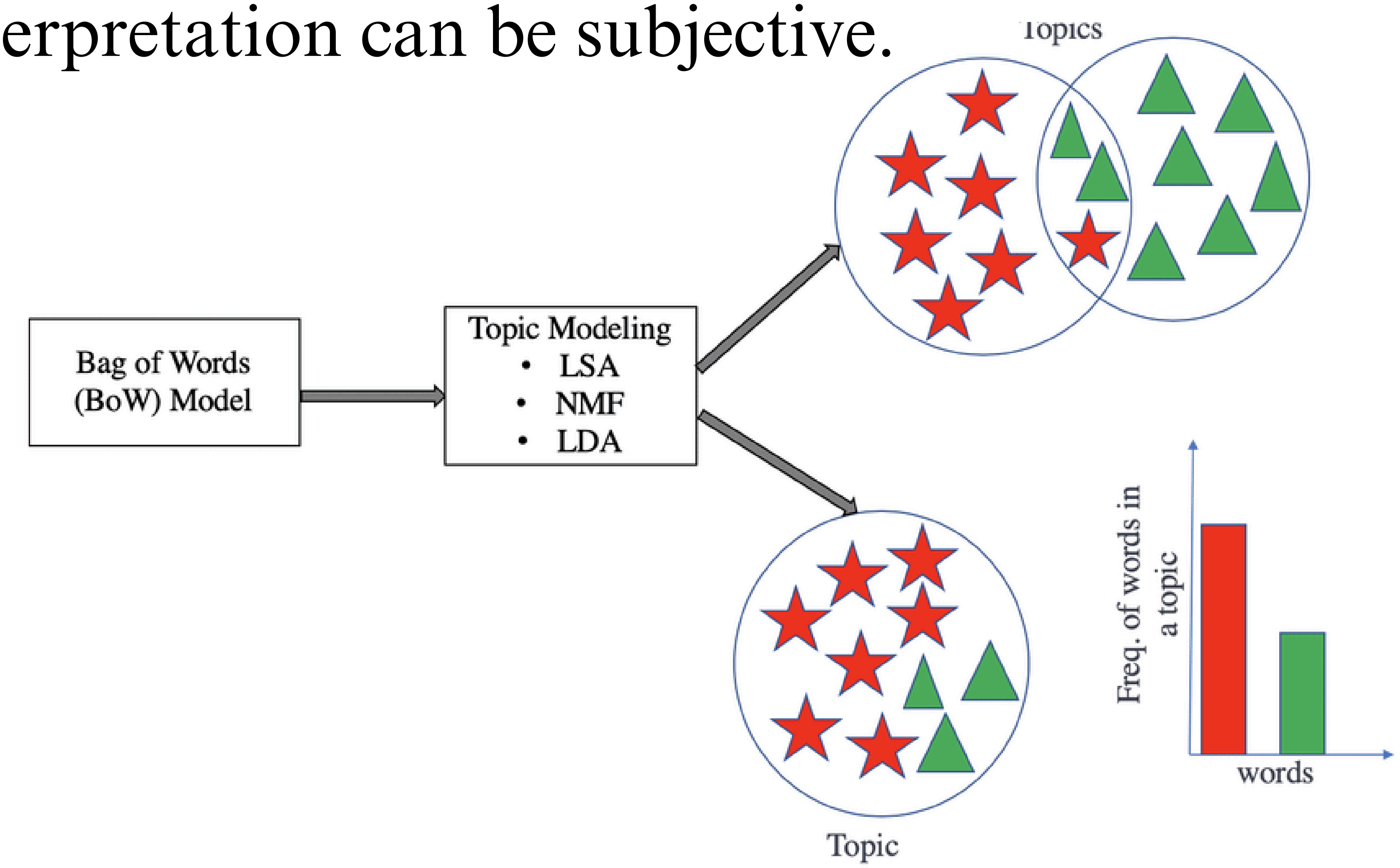


Image adapted from: "Yoga-Veganism: Correlation Mining of Twitter Health Data," 2019.

[3] J. Boyd-Graber, Y. Hu, and D. Mimno, "Applications of Topic Models," Foundations and Trends in Information Retrieval, vol. 11, no. 2-3, pp. 143-296, 2017.

Existing Techniques for Analyzing Public Opinion on Twitter and Its Challenges

4. Hybrid Approach [5]:

- **Mechanism:** Combine multiple techniques (e.g., lexicon-based sentiment analysis with topic modeling, topic modeling and ABSA) to leverage their strengths.
- **Advantages:** Provides a more deeper analysis of opinions than general sentiment analysis.
- **Limitations:**
    - Focus on specific domains (e.g., product reviews, hotel feedback) or smaller datasets.
    - Often focus on either technique separately. Missing the opportunity to gain deeper insights from their combined application.
    - While those has been applied to public opinion research, it is often used in conjunction with surveys or other data sources, not the meta data.
    - Face challenges related to data quality (e.g., noise, sarcasm), aspect identification, and sentiment classification accuracy.

[5] Kumar, A., Jaiswal, A., & Kumar, A. (2019). Sentiment Analysis of Twitter Data: A Hybrid Approach. Journal of Advances in Information Technology, 10(2), 1-16.

Ideas for Improvement

- Design and implement algorithms specifically tailored to address and handle the massive volume of social media data
- Work with specific data with noise, slang, and figurative language.
- Real-time monitoring and analysis to track public opinion trends
- Go beyond binary sentiment classification and explore nuanced sentiments towards multiple aspects of a topic

Propose Solution

=> integrates the two model of **Structural Topic Modeling (STM) and Aspect-Based Sentiment Analysis (ABSA)** for analyzing public opinion on X (Twitter)

Basic Ideas of Structural Topic Modeling (STM)

- **Definition: A method to discover hidden topics within large sets of text [5].**
  - **Example**: Imagine reading thousands of datas and automatically finding recurring themes like "economy," "politics," or "entertainment."
- **Its advantages:**
  - Allows discover **latent topics** within **large volumes** of **unstructured text data**
  - Design specially to **incorporate with metadata.**
  - Allows for **correlations between topics**, for example, how do topics correlate with specific events or variables?

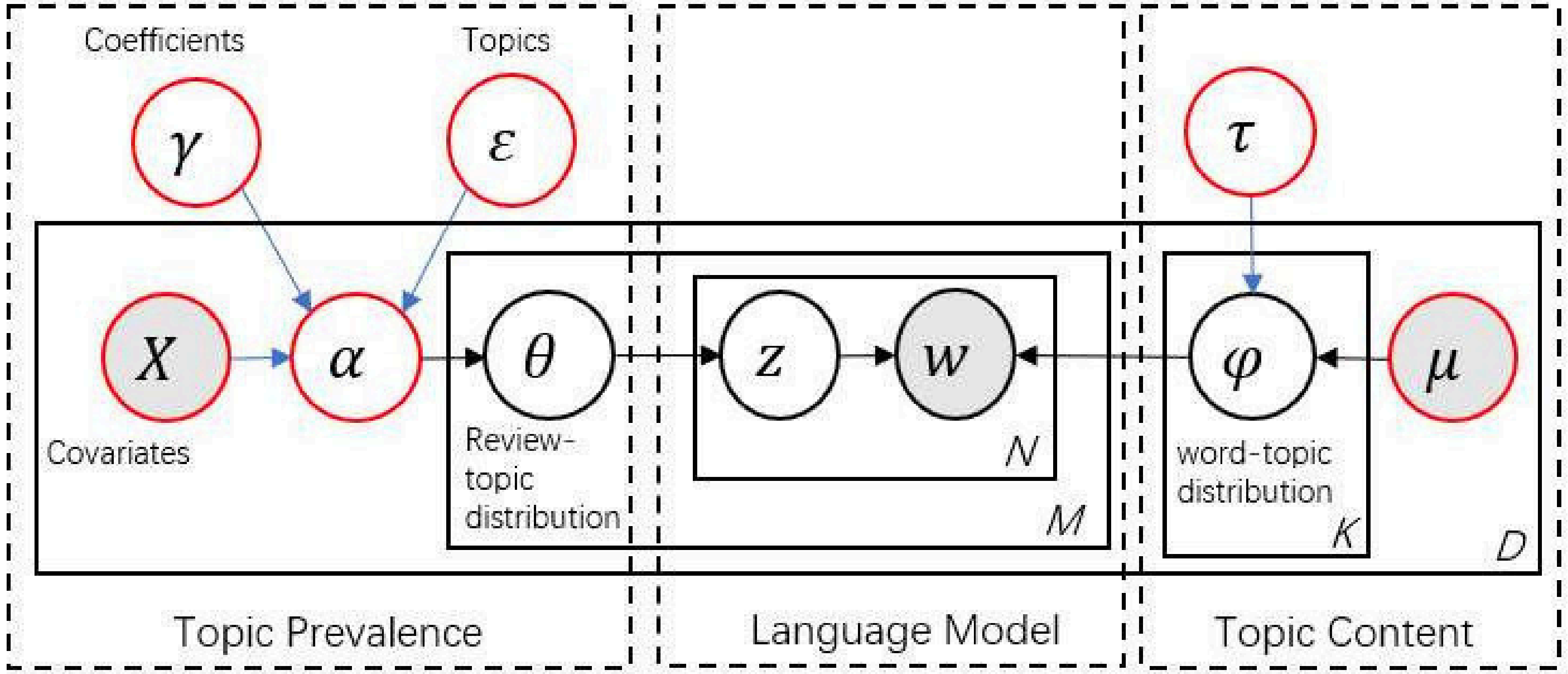


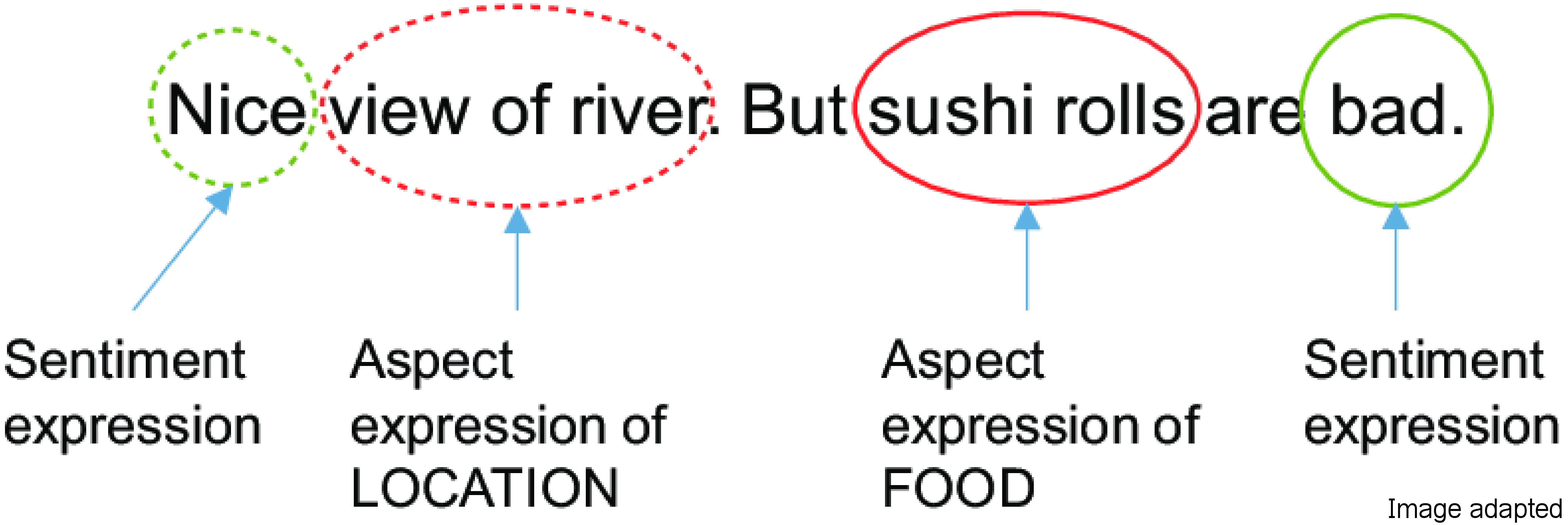
Image adapted from: He, L., Han, D., Zhou, X., & Qu, Z. (2020). "The Voice of Drug Consumers: Online Textual Review Analysis Using Structural Topic Model."

[5] Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). "stm: An R Package for Structural Topic Models." Journal of Statistical Software, 91(2), 1-40. DOI: 10.18637/jss.v091.i02

Basic Ideas of Aspect-Based Sentiment Analysis (ABSA)

There are 4 level of Sentiments Analysis

- Level 1: Basic Sentiment Analysis**:** Determines the overall polarity of a text (positive, negative, or neutral).
- Level 2: Categorization of Sentiment: Identifies specific emotions like joy, anger, or sadness in the text.
- Level 3: Sentiment by Topic: Analyzes sentiment towards specific topics or entities mentioned in the text.
- **Level 4: Aspect-Based Sentiment Analysis: Goes further by identifying specific aspects or features of those topics/entities and determining the sentiment expressed towards each aspect.**



Nice view of river. But sushi rolls are bad.

Sentiment expression

Aspect expression of LOCATION

Aspect expression of FOOD

Sentiment expression

Image adapted from: "An example of aspect-based sentiment analysis (ABSA)," ResearchGate.

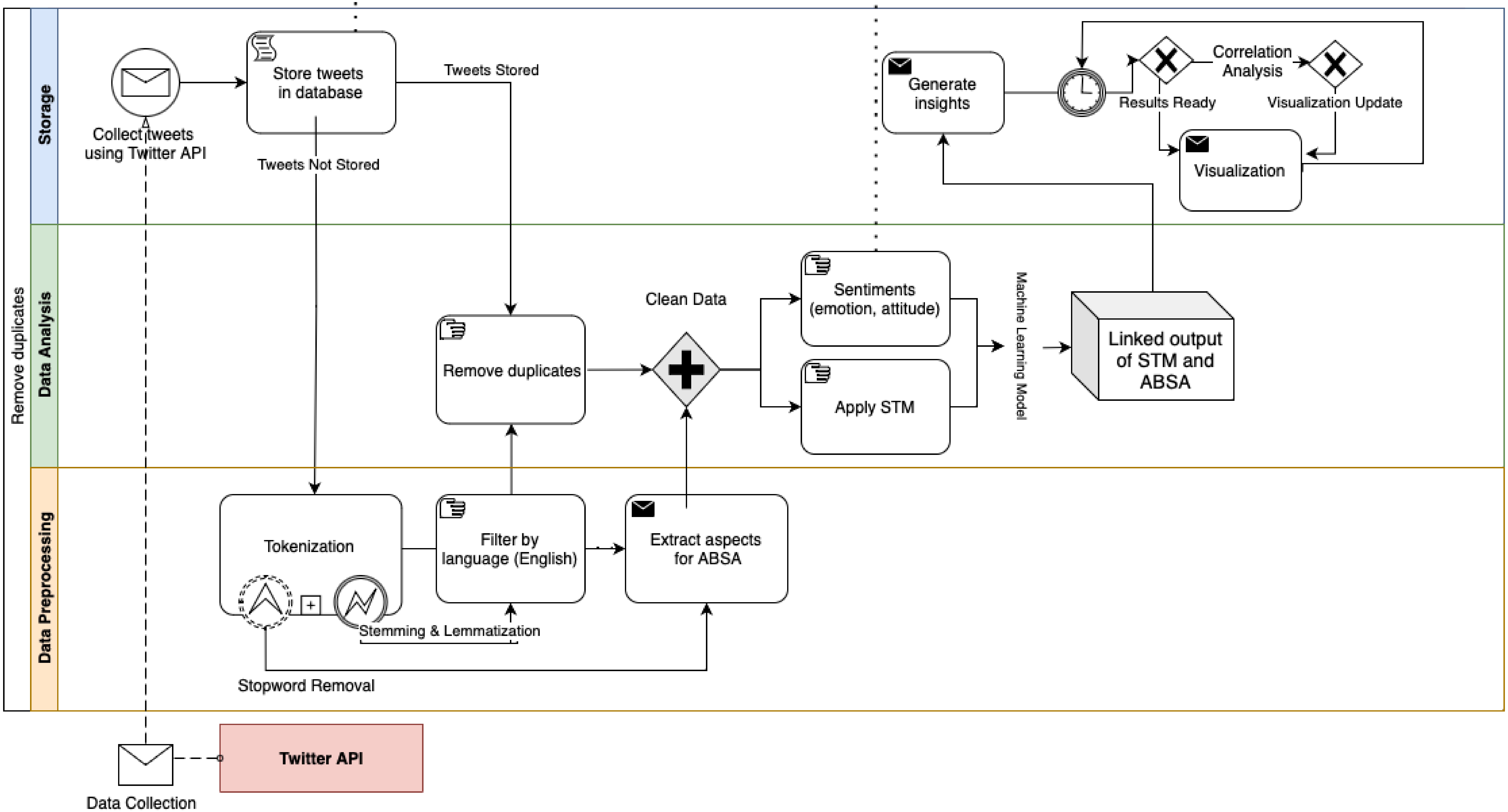
Nazir, A., Rao, Y., Wu, L., & Sun, L. "Issues and Challenges of Aspect-Based Sentiment Analysis: A Comprehensive Survey." IEEE Transactions on Affective Computing, 2022.

Key Improvements and Expected Research Goals

What is my novelty by combining STM and ABSA compared to previous work?

- First to apply this hybrid method on Twitter data.
- Developing tailored algorithms and techniques to handle the unique characteristics of Twitter data, capable of handling the decent volume of Twitter data.
- Providing a more comprehensive and nuanced analysis. STM will uncover the main themes and discussions, while ABSA will delve into the specific aspects.
- Analyzing sentiment towards multiple aspects simultaneously. Going beyond the basic positive/negative sentiment classification often limited in previous work.
- Real-time analysis, e.g., how people's thoughts change over time.

Model Framework



Remove duplicates

Storage

Data Analysis

Data Preprocessing

Collect tweets using Twitter API

Store tweets in database

Tweets Stored

Tweets Not Stored

Generate insights

Correlation Analysis

Results Ready

Visualization Update

Visualization

Remove duplicates

Clean Data

Sentiments (emotion, attitude)

Apply STM

Machine Learning Model

Linked output of STM and ABSA

Tokenization

Filter by language (English)

Extract aspects for ABSA

Stemming & Lemmatization

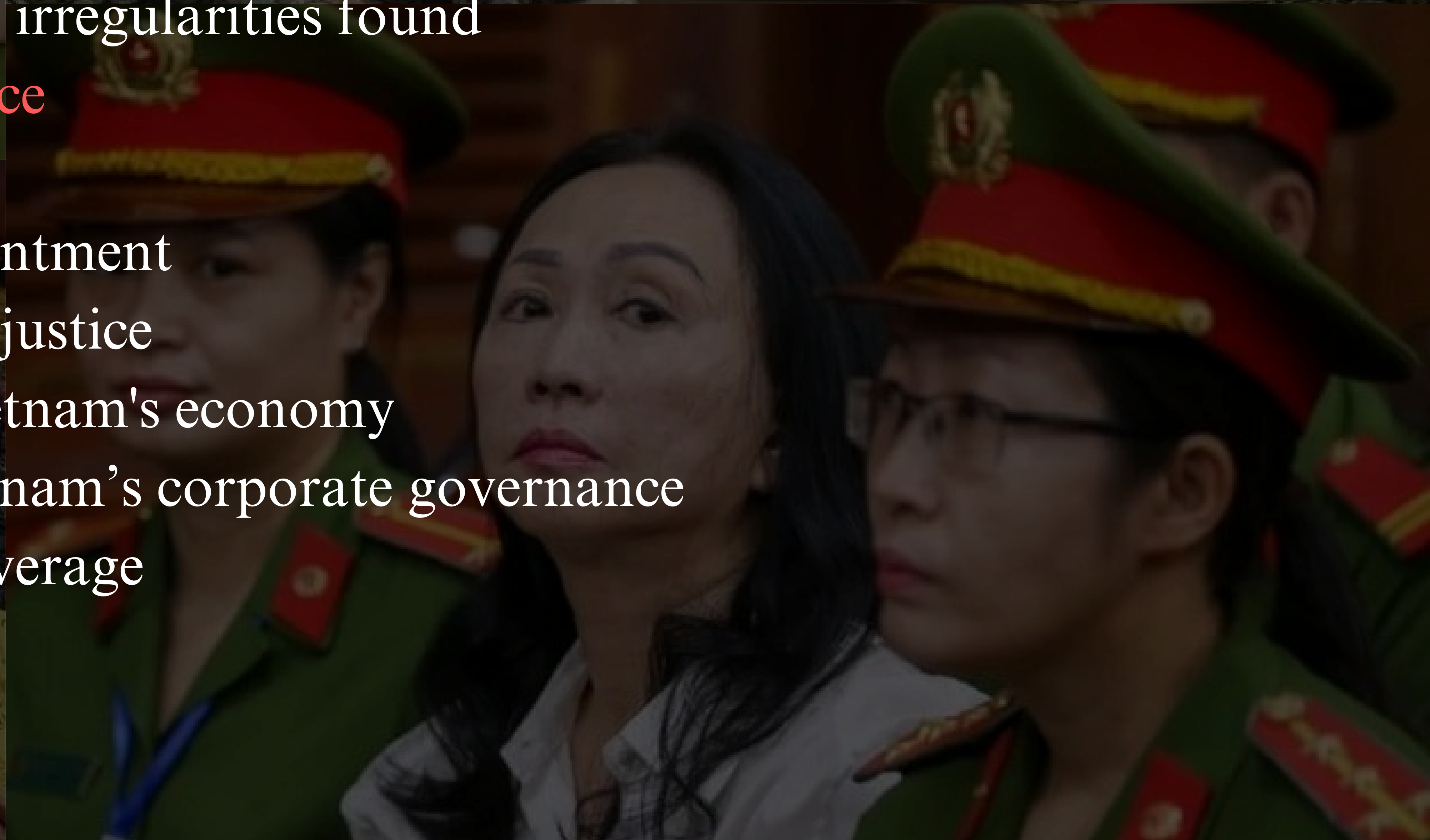
Stopword Removal

Twitter API

Data Collection

Case Study and Initial Result

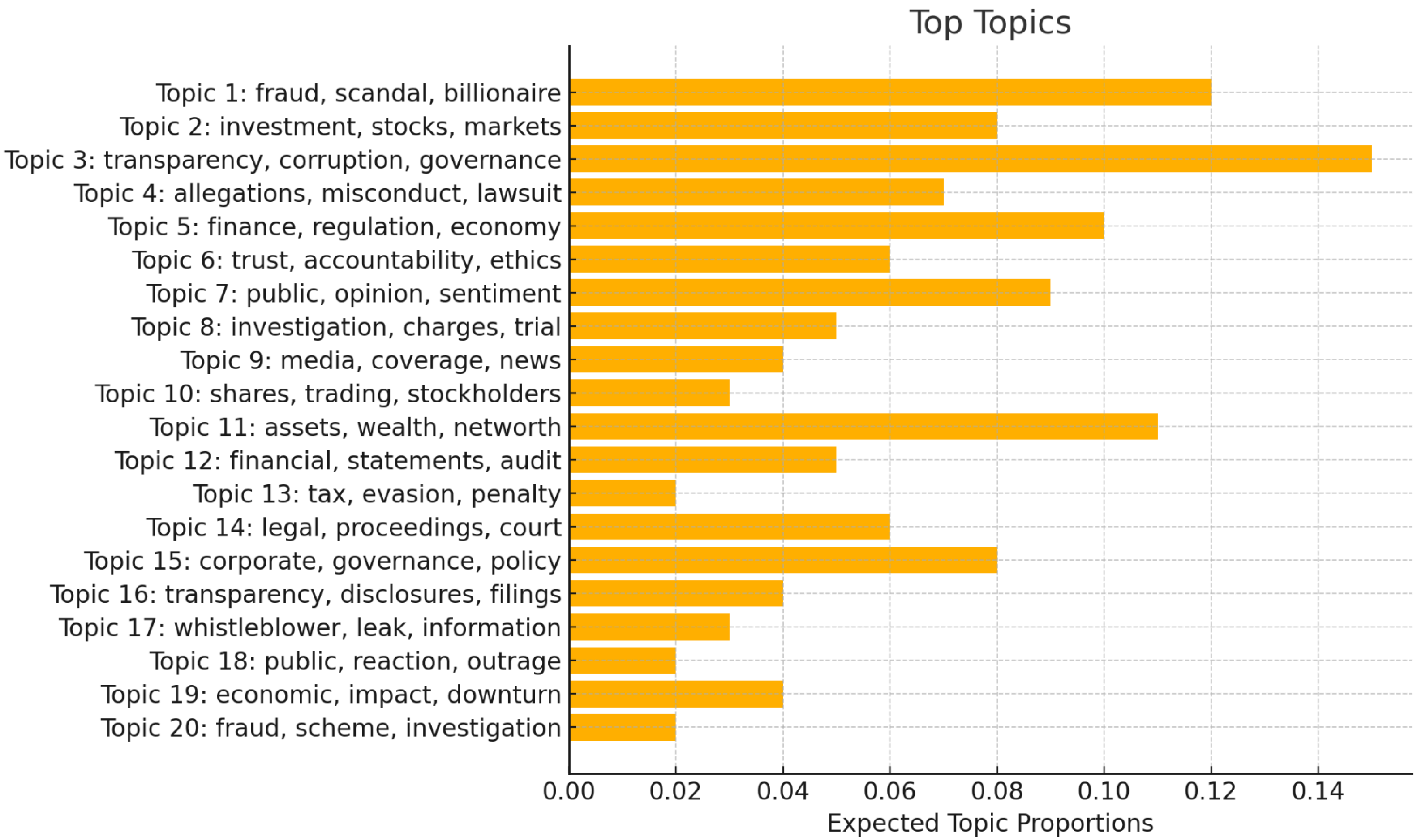
Case Selection: #TruongMyLan Financial Crime



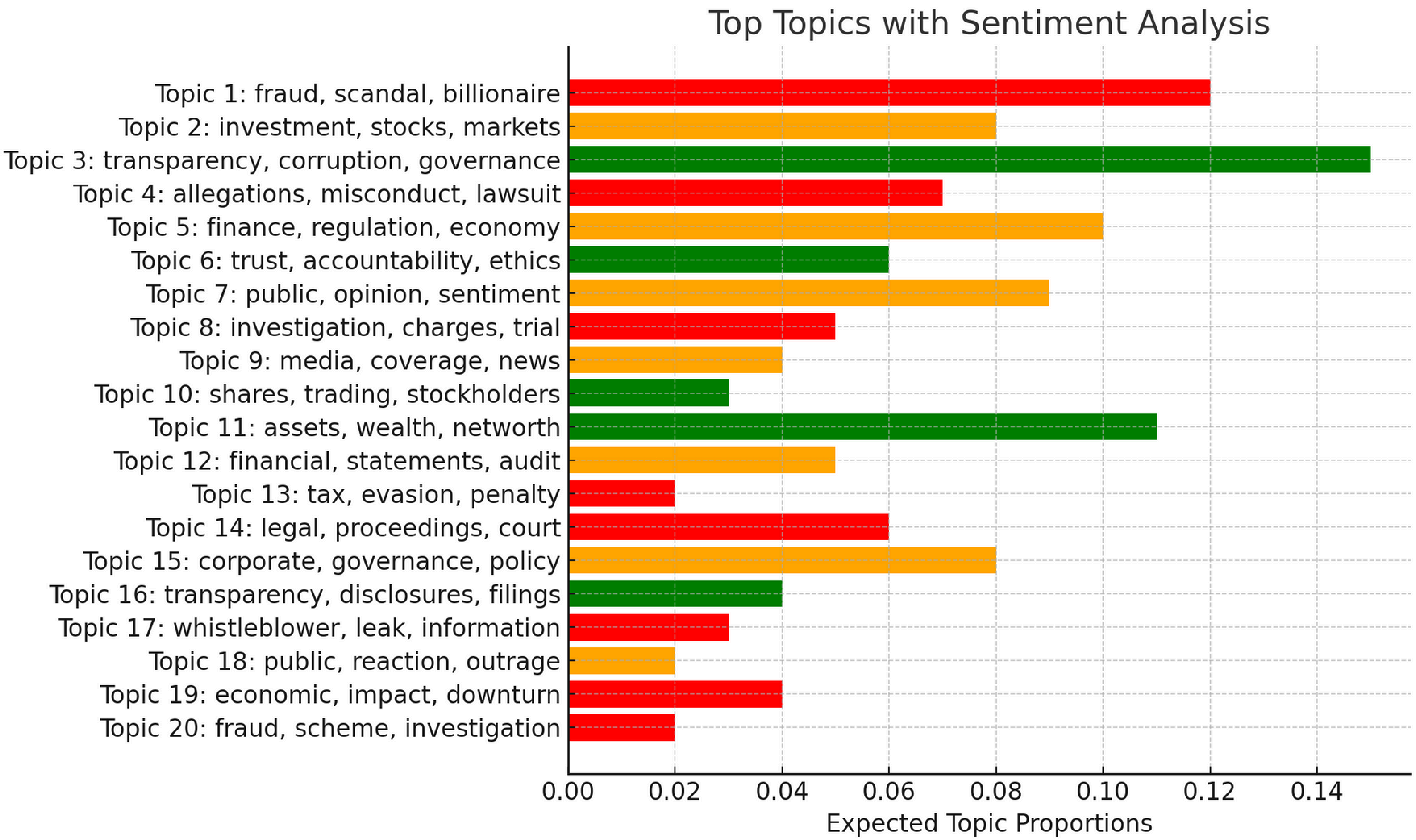
Trương Mỹ Lan Case Overview

- Date: April 11, 2024
- Person: Trương Mỹ Lan, Vietnamese billionaire
- Allegations: Embezzlement and fraud
- Investigation: Financial irregularities found
- Outcome: Death sentence
- Public Reaction:
  - Shock and disappointment
  - Anger and calls for justice
  - Concerns about Vietnam's economy
  - Discussions on Vietnam’s corporate governance
  - Extensive media coverage

Expected Result of STM



Expected Result of ABSA

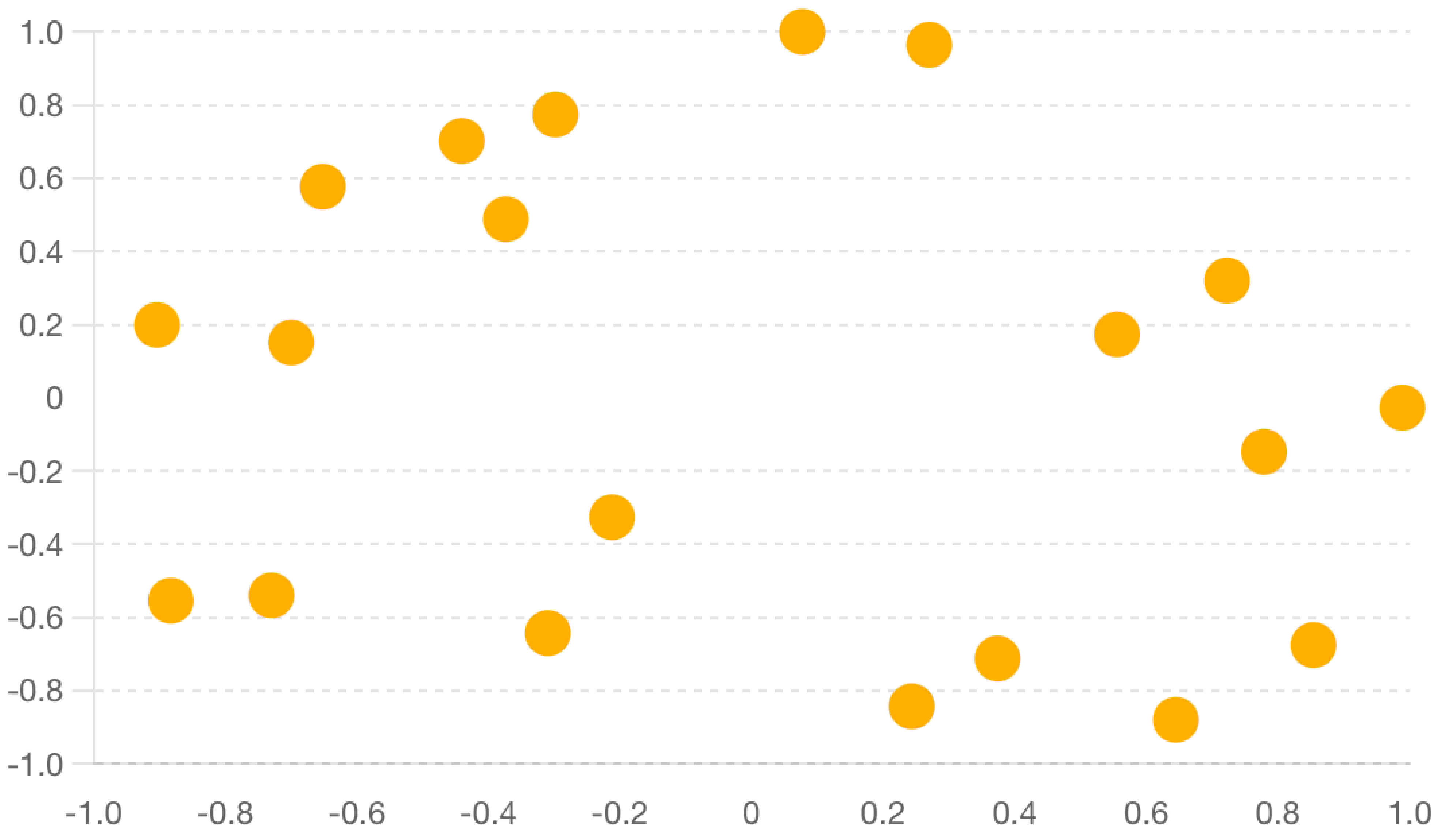


- Green: Positive Sentiment
- Red: Negative Sentiment
- Orange: Mixed Sentiment

Expected Result of Correlation Analysis

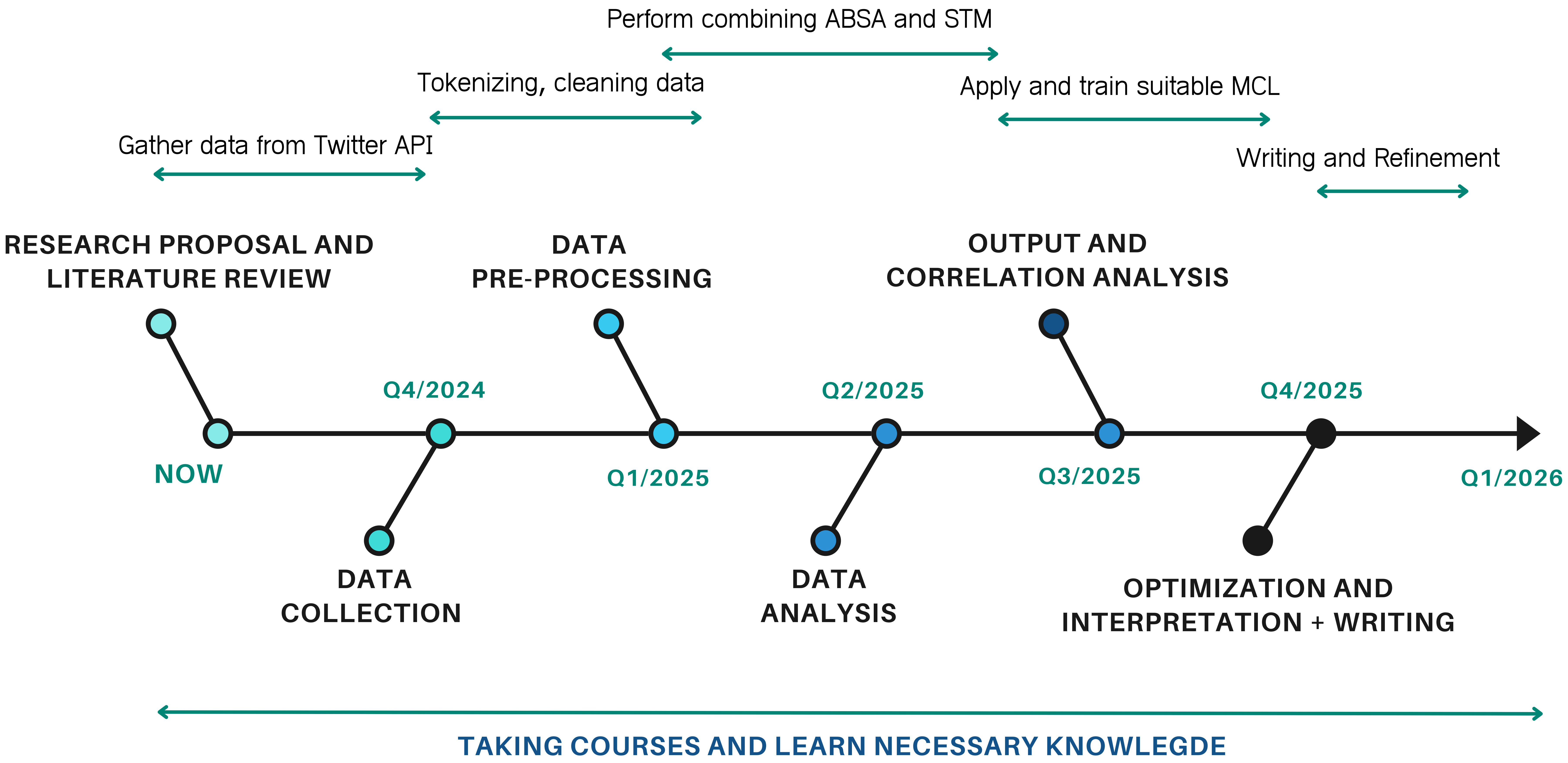
- Topics that are closely related tend to cluster together

- Interpretation:
  - Identifying Key Relationships
  - Understanding Public Opinion
  - Insights



- T1 (fraud, scandal, billionaire) == T3 (transparency, corruption, governance) & T6 (trust, accountability, ethics)
  => suggesting discussions about transparency and ethics in the context of financial scandals.
- T2 (investment, stocks, markets) == T5 (finance, regulation, economy) & T18 (public, reaction, outrage)
  => indicating connections between market investments, financial regulations, and public reactions

Research Plan Timeline



Perform combining ABSA and STM

Tokenizing, cleaning data

Apply and train suitable MCL

Gather data from Twitter API

Writing and Refinement

RESEARCH PROPOSAL AND LITERATURE REVIEW

DATA PRE-PROCESSING

OUTPUT AND CORRELATION ANALYSIS

Q4/2024

Q2/2025

Q4/2025

NOW

Q1/2025

Q3/2025

Q1/2026

DATA COLLECTION

DATA ANALYSIS

OPTIMIZATION AND INTERPRETATION + WRITING

TAKING COURSES AND LEARN NECESSARY KNOWLEGDE

Thank you.