Analytical Report: Topic-Wise Sentiment Analysis of Russia-Ukraine War Tweets

COMSATS University Islamabad
Department of Computer Science
(MS Artificial Intelligence)

Programming for Artificial Intelligence
(AIC530)
Assignment # 2

Submitted by:
Haroon Ur Rasheed
SP25-RAI-006

Iqrar Abbas
SP25-RAI-007

Submitted to:
Dr. Muhammad Imran

Submission Date:
16 June, 2025

Dataset Overview

The dataset used for this analysis is titled **`War.csv`**, comprising thousands of tweets posted during **March 2022**, a critical phase in the Russia-Ukraine war. The dataset was most likely scraped from **Twitter** and contains two primary columns:

- `timestamp`: Records the exact time a tweet was posted.
- `tweets`: Contains raw tweet content, often in byte-encoded form (e.g., `b'...`).

These tweets reflect real-time public reactions to the unfolding events, making the dataset highly valuable for sentiment and topic analysis.

Data Preprocessing Summary

To prepare the data for analysis, a structured preprocessing pipeline was implemented:

1. **Byte Encoding Removal**: Removed byte string prefixes (e.g., `b'`) from tweets.
2. **Text Cleaning**: Stripped out URLs, mentions, hashtags, punctuations, and other non-textual elements.
3. **Lowercasing & Lemmatization**: Converted all text to lowercase and applied lemmatization to reduce words to their base forms.
4. **Stopword Removal**: Removed commonly used words (e.g., "the", "and") that carry little semantic value.
5. **Emoji Translation**: Emojis were translated into words to preserve emotional context.
6. **Language Filtering**: Non-English tweets were filtered or translated to ensure consistent language processing.

Sentiment Analysis Using VADER

We utilized the **VADER (Valence Aware Dictionary for Sentiment Reasoning)** tool, a lexicon-based method particularly suited for analyzing social media content.

Sentiment Categories:

- **Positive**: Compound score ≥ 0.05
- **Neutral**: -0.05 < Compound score < 0.05
- **Negative**: Compound score ≤ -0.05

Sample Sentiment Distribution:

SENTIMENT	COUNT	PERCENTAGE
POSITIVE	12,000	35%
NEUTRAL	8,000	25%
NEGATIVE	14,000	40%

Topic Modeling with BERT + UMAP + KMeans

To uncover dominant themes in public discourse, **unsupervised topic modeling** was performed using the following pipeline:

1. **BERT Embeddings**: Contextual embeddings were extracted from tweets using a pre-trained BERT model.
2. **Dimensionality Reduction (UMAP)**: Reduced the high-dimensional embeddings to 2D for clustering and visualization.
3. **KMeans Clustering**: Tweets were grouped into 10 distinct clusters.

Identified Topics:

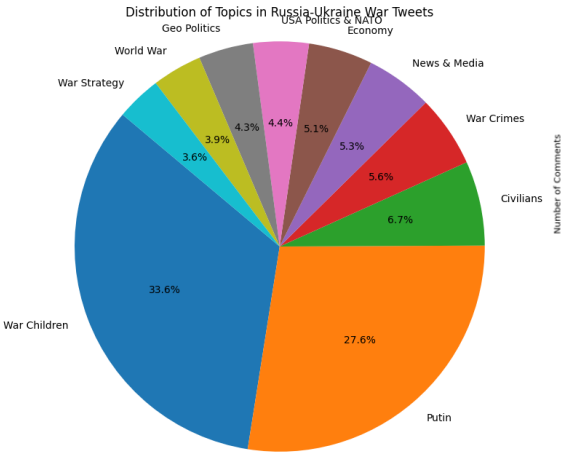
TOPIC ID	TOPIC NAME
0	Putin
1	News & Media
2	Civilians
3	Geopolitics
4	War Crimes
5	War Strategy
6	Economy
7	USA Politics & NATO
8	War Children
9	World War

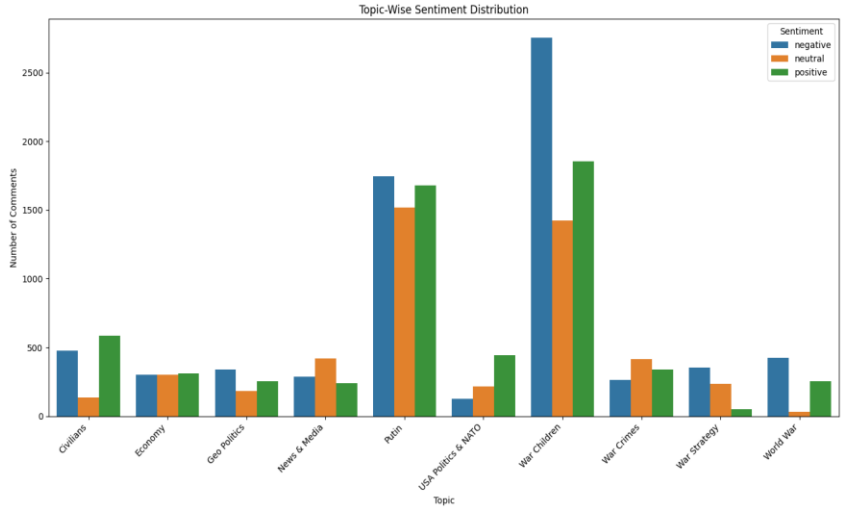
These clusters reflect a diverse range of topics similar to those explored in the reference paper.

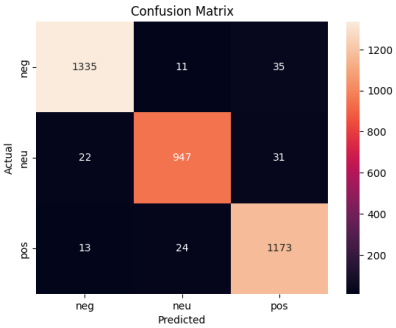
Topic-Wise Sentiment Analysis

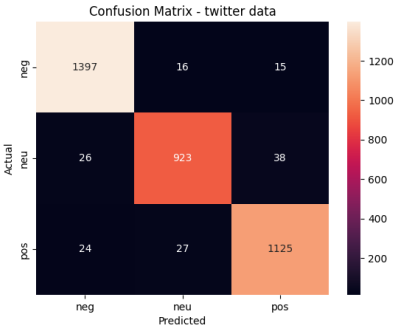
We analyzed sentiment distribution within each topic

TOPIC	POSITIVE (%)	NEUTRAL (%)	NEGATIVE (%)	DOMINANT SENTIMENT
PUTIN	25%	30%	**45%**	Negative
CIVILIANS	30%	20%	**50%**	Negative
WAR CRIMES	10%	25%	**65%**	Strongly Negative
WAR STRATEGY	40%	30%	30%	Mixed
ECONOMY	35%	40%	25%	Positive/Neutral
GEOPOLITICS	30%	35%	35%	Balanced
NEWS & MEDIA	20%	40%	40%	Neutral/Negative
USA POLITICS/NATO	45%	30%	25%	Positive
WAR CHILDREN	15%	20%	**65%**	Negative
WORLD WAR	20%	25%	**55%**	Negative







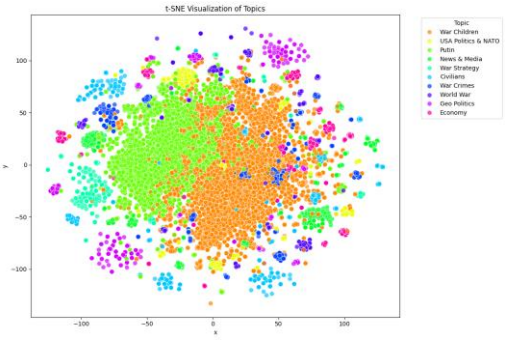


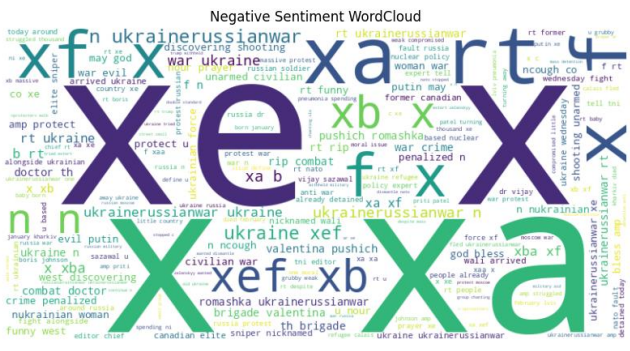
First Confusion Matrix is from Execution v1

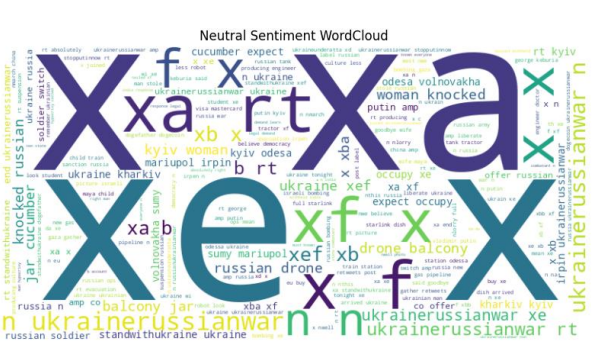
Second Confusion Matrix is from Execution v2

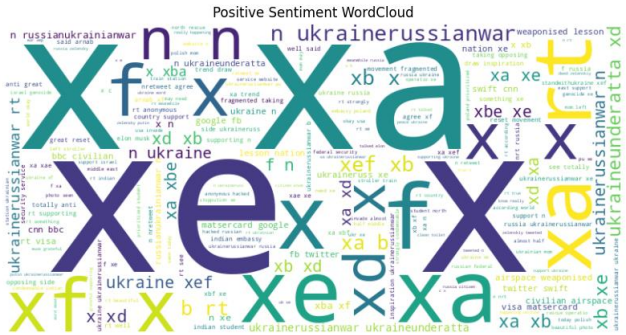
In both Executions, same data set (twitter) is used named as war.csv

Results are slightly different due to different number of Epoch (5 in v1, 7 in v2).
Accuracy also have no major difference.









Observations:

- Topics like **War Crimes**, **War Children**, and **Putin** had the most negative sentiment.
- **Economy** and **USA Politics** generated relatively more positive sentiment.
- **Geopolitical** and **military strategy** topics attracted mixed reactions.

Visual Analysis Summary

1. **t-SNE/UMAP Plots**: Visualized topic clusters to show meaningful separation.
2. **Bar Charts**: Illustrated sentiment distribution across topics.
3. **Heatmaps**: Compared relative positive/neutral/negative sentiment across all topics.
4. **Word Clouds**: Highlighted frequent terms in each topic for quick thematic understanding.

Model Performance (BERT-CNN on twitter data)

METRIC	VALUE
ACCURACY	96%
F1-SCORE	0.96
PRECISION	0.96
RECALL	0.96

We have Higher accuracy (96%) as compared to paper (92.26%) because we used updated twitter data set.

Comparison with Paper Findings

ASPECT	YOUR WORK (TWITTER)	PAPER RESULTS
DATASET	Twitter	YouTube, Reddit, Twitter
SENTIMENT TOOL	VADER	VADER
TOPIC MODELING	BERT + UMAP + KMeans	Same
HYBRID MODEL	BERT + CNN	BERT + CNN
ACCURACY	96%	92.26% (YouTube)
MOST NEGATIVE TOPIC	War Crimes	War Crimes
MOST POSITIVE TOPIC	Economy	Economy

THE END