Ukrainian Catholic University

Faculty of Applied Sciences

Artificial Intelligence Course

Perception Under Fire: U.S. Responses to Ukraine War Disinformation

Final Project Paper

Authors:

Yuliana Hrynda

Iryna Denysova

Roman Pavlosiuk

2025

Abstract

This project analyzes shifts in public sentiment on the Russia-Ukraine war by examining YouTube comments, classifying them as pro-Russian, pro-Ukrainian, or neutral. Using sentiment analysis and stance classification, we track opinion changes over time, offering insights into public discourse on the conflict.

Keywords: sentiment analysis, disinformation, Ukraine war, social media, transformers

1 Introduction

The U.S. plays a major role in supporting Ukraine, so American public opinion is important for continued aid. Since the war began, information about Ukraine has often been affected by propaganda and disinformation, which can change how people understand the conflict. Social media spreads this information quickly and strongly influences opinions. YouTube is especially important because many Americans watch news there and share their thoughts in comments. These comments show real public reactions and make YouTube a useful source for studying opinion changes over time. The goal of this project is to study sentiment in YouTube comments from U.S. news channels to see how American views on the Ukraine war change over time. We focus on whether the sentiment is mostly positive or negative and how it shifts during different periods of the war.

2 Related Work

The paper "Knowledge Graph Convolutional Networks for Sentiment Classification" [1] analyzes over 400,000 tweets about the Russia–Ukraine war using models like GCNs, LSTMs, and RoBERTa, achieving high accuracy in detecting themes like war, people, and peace. It also finds sentiment analysis more accurate on English-language data due to the complexity of Russian and limited labeled datasets.

The article "AI Sentiment Analysis and Russia's War in Ukraine" [2] discusses AI tools like FilterLabs for analyzing Russian sentiment, highlighting concerns about non-transparent methods. It also reviews models like RuSentNE-2023 and RuSentiTweet, noting that accuracy depends on dataset size, translation quality, and model fine-tuning.

3 Data Collection

To study American public sentiment toward the Russia–Ukraine war, we collected data from YouTube comments posted under news videos published by major U.S. media outlets. In the current stage of the project, we focus on CNN News as the primary data source, as it is one

of the most influential U.S. news channels on YouTube. Using the YouTube Data API v3, we collected 1,511 video IDs related to coverage of the Russia–Ukraine war. For each video, we extracted all available public comments along with their publication timestamps. The analyzed data covers the period from years 2023 to 2025. At this stage, comments from earlier years (including 2022) are not fully available for CNN due to platform and API limitations. The collected dataset serves as the basis for sentiment and temporal analysis of American public reactions to war-related news.

4 Comments Preprocessing

Before running sentiment analysis, we first cleaned and prepared the YouTube comments. We removed URLs, user mentions, emojis and some special symbols, and converted all text to lowercase. We also dropped comments that were extremely short or contained only links, because they did not give useful information about sentiment.

After this basic cleaning, we standardized the text encoding to UTF-8 and removed extra spaces. We did not use lemmatization or stemming, because modern transformer models are trained on raw text and can handle different word forms using subword tokenization. This also lets us keep important phrases, names and slogans in their original form, which is important for political and war-related content. For the same reason, we did not remove stopwords, since they are part of normal sentences and are already included in the data that the models were trained on.

Finally, we used the tokenizer from the transformer models to prepare input for training and inference. The tokenizer splits each comment into tokens, maps them to vocabulary indices, and pads or truncates them to a fixed maximum length so that we can process comments in batches on the GPU. In this way, we keep the text close to the original while making it suitable as input for our models.

5 Methodology

5.1 Creating a Labeled Dataset with Qwen2.5-1.5B-Instruct

In addition to using a ready sentiment model, we also built our own classifier that is specific to the Russia–Ukraine war comments. To do this, we first created a labeled training dataset using the Qwen2.5-1.5B-Instruct language model. Qwen2.5-1.5B-Instruct is a decoder-only transformer with about 1.5 billion parameters, designed for instruction following and efficient inference on consumer hardware. We used this model as a teacher model to automatically assign one of three sentiment labels (negative, neutral or positive) to each comment.

The main reason for choosing the 1.5B version was our hardware limit. Our GPU has only

8 GB of VRAM, which is usually not enough to run larger 7B models comfortably for long inference runs on hundreds of thousands of comments. With Qwen2.5-1.5B-Instruct, we were able to run the model locally, using low-precision weights, and still keep the whole pipeline stable on our machine. This allowed us to generate a large pseudo-labeled dataset without manual annotation.

We designed a simple prompt that described the task in plain English and asked Qwen2.5-1.5B-Instruct to decide whether a comment is negative, neutral or positive toward Ukraine or the situation in general. For each comment, we saved the model's label together with the original text and timestamp. This dataset later became the training data for our own sentiment classifier.

5.2 Training a DistilBERT Sentiment Classifier

After creating the pseudo-labeled dataset with Qwen2.5-1.5B-Instruct, we trained a smaller classifier based on DistilBERT. DistilBERT is a compact version of BERT that keeps most of the accuracy but is faster and uses less memory, which is important for our 8 GB GPU.[**?**] We used the `distilbert-base-uncased` model as the backbone and added a simple classification head that predicts three classes: negative, neutral and positive.

We fine-tuned DistilBERT on the pseudo-labeled comments using standard cross-entropy loss. During training, we split the data into training and validation sets and monitored the loss and accuracy on the validation part to check that the model was learning and not overfitting too much to the noise in the teacher labels. After training, we used this DistilBERT model as our main sentiment classifier for the full YouTube dataset.

During inference, each cleaned comment is tokenized with the DistilBERT tokenizer and passed through the fine-tuned model, which outputs probabilities for the three sentiment classes. We choose the class with the highest probability as the predicted label. For some analyses, we also map the labels to numeric values (-1 for negative, 0 for neutral and $+1$ for positive) and compute average scores over time, similar to the continuous score used with the RoBERTa model.

5.3 Sentiment Analysis Model

To quantify the emotional tone of user comments related to the war in Ukraine, we employ a transformer-based sentiment classifier from Hugging Face:

`cardiffnlp/twitter-roberta-base-sentiment-latest`.

This model is built on the RoBERTa architecture, an optimized variant of BERT that improves upon the original design through longer pretraining, dynamic masking, and the removal of the next-sentence prediction objective. RoBERTa uses multi-head self-attention layers to capture

contextual relationships within text, enabling it to model complex linguistic phenomena such as negation, sarcasm, and emphasis.

Training domain. The model was trained on large-scale Twitter data, which consists of short, informal, and emotionally expressive text. This training domain closely resembles YouTube comments in structure and style, making the model well-suited for our dataset. At the same time, the model is not specifically trained on geopolitical or conflict-related discourse; therefore, it captures *general sentiment polarity* rather than explicit political stance.

Output classes. For each input comment, the model produces a probability distribution over three sentiment classes:

- Negative
- Neutral
- Positive

These probabilities are obtained by applying a softmax layer to the final hidden representation of the transformer (corresponding to the CLS token), ensuring that the outputs sum to one.

Continuous sentiment score. Rather than relying solely on discrete sentiment labels, we convert the model's probabilistic output into a continuous sentiment score in the range $[-1, 1]$, which allows us to capture sentiment intensity. Let p_{pos}, p_{neg}, and p_{neu} denote the predicted probabilities for the positive, negative, and neutral classes, respectively. The sentiment score s for a comment is computed as:

$$s = p_{\text{pos}} - p_{\text{neg}}.$$

Under this formulation, values close to $+1$ indicate strongly positive sentiment, values close to -1 indicate strongly negative sentiment, and values near 0 correspond to neutral or uncertain sentiment.

5.4 Temporal Aggregation

To analyze sentiment dynamics over time, individual comment-level sentiment scores are aggregated by date using the mean. This produces a daily sentiment signal that reflects the average emotional tone of user discussions on a given day. Since daily sentiment can be highly volatile due to short-term reactions, varying comment volumes, and isolated events, we apply a centered 7-day rolling average to smooth the time series. This smoothing reduces noise while preserving meaningful medium-term trends.

The resulting smoothed sentiment trajectory serves as the basis for subsequent temporal analysis, including the detection of local extrema and the identification of periods with abrupt changes in public perception.

To facilitate interpretation, we visualize the sentiment evolution separately for recent years, focusing on 2023, 2024, and 2025. Presenting yearly plots allows clearer inspection of local dynamics and avoids visual compression effects that occur in long multi-year timelines.

Together, these yearly visualizations provide a clear temporal context for identifying sentiment extrema and relating them to specific high-impact media events discussed in subsequent sections.

6 Experiments

This section describes the experimental setup used to analyze sentiment dynamics over time, including data processing steps, aggregation intervals, and the statistical measures applied to identify and interpret sentiment trends.

6.1 Experimental Setup

The experiments are conducted on a large-scale dataset of YouTube comments related to videos discussing the war in Ukraine. Each comment is associated with a publication date of the source video, a stance label (pro-Ukrainian, pro-Russian, or neutral), and a continuous sentiment score in the range $[-1, 1]$ computed using a transformer-based sentiment model, as described in Section 5.1.

All experiments are performed offline using Python, with `pandas` for data manipulation, `NumPy` and `SciPy` for numerical analysis, and `matplotlib` and `Plotly` for visualization. To ensure reproducibility, timestamps are normalized to calendar dates, and all sentiment values are computed prior to aggregation.

6.2 Temporal Aggregation Strategy

Sentiment analysis is performed at multiple temporal resolutions. The primary aggregation interval is **daily**, where individual comment-level sentiment scores are averaged for each calendar day:

$$S_d = \frac{1}{N_d} \sum_{i=1}^{N_d} s_i,$$

where s_i denotes the sentiment score of comment i, and N_d is the number of comments observed on day d.

Daily aggregation strikes a balance between temporal resolution and statistical stability: it is

fine-grained enough to capture reactions to specific media events, while still averaging over a sufficient number of comments to reduce random noise.

6.3 Smoothing and Trend Extraction

Daily sentiment signals are inherently noisy due to fluctuations in comment volume and short-lived reactions. To highlight medium-term trends, we apply a centered **7-day rolling mean**:

$$\tilde{S}_d = \frac{1}{7} \sum_{k=-3}^{3} S_{d+k}.$$

This smoothing window was chosen empirically to reduce high-frequency variance while preserving meaningful shifts in sentiment that unfold over several days. The smoothed signal $\tilde{S}_d$ is used for all subsequent analyses, including visualization and extremum detection.

6.4 Statistical Measures

Several statistical measures are used to analyze sentiment trends:

- **Mean sentiment**: captures the overall polarity of discourse within a given time interval.
- **Rolling average**: used to extract stable trends from noisy daily data.
- **Local extrema**: local maxima and minima of the smoothed sentiment curve are detected using a neighborhood-based comparison. These extrema indicate abrupt positive or negative shifts in public perception.

Local extrema are identified algorithmically by comparing each point in the smoothed series to its neighboring values within a fixed temporal window. This approach allows for systematic detection of sentiment turning points without manual thresholding.

6.5 Yearly Analysis

To improve interpretability and avoid visual compression effects, sentiment trends are analyzed separately for each calendar year. Yearly segmentation enables more precise examination of local dynamics and facilitates alignment between sentiment shifts and contemporaneous media events. All reported extrema and visualizations are therefore contextualized within their respective yearly timelines.

Overall, this experimental design enables a structured and statistically grounded analysis of how public sentiment toward Ukraine evolves over time and how abrupt changes in sentiment relate to high-attention media content.

6.6 Running Our DistilBERT Pipeline on the Full Dataset

Besides experiments with the RoBERTa-based model, we also ran our own DistilBERT classifier on the full dataset of CNN comments. First, we applied the preprocessing steps described in Section 4 to all comments. Then we used the fine-tuned DistilBERT model from the previous subsection to predict sentiment labels for every comment in the dataset.

To make this step efficient on our hardware, we processed comments in batches and limited the maximum sequence length so that the model would fit into 8 GB of VRAM without out-of-memory errors. For each comment, we saved the predicted class (negative, neutral or positive), the model probabilities and the comment timestamp. After that, we aggregated these predictions by day in the same way as for the RoBERTa-based scores: we computed daily counts and proportions of each class and calculated average sentiment values.

This second pipeline gave us an additional view on the data based on a model that we trained ourselves using Qwen2.5-1.5B-Instruct labels. Comparing the RoBERTa-based scores with the DistilBERT predictions helped us to check the stability of the observed trends and to see how much the results depend on the choice of model and training data.

7 Event-Based Interpretation of Sentiment Spikes

For each local maximum and minimum, we identified the dominant video driving user discussion on that day and analyzed the associated topic.

7.1 2023

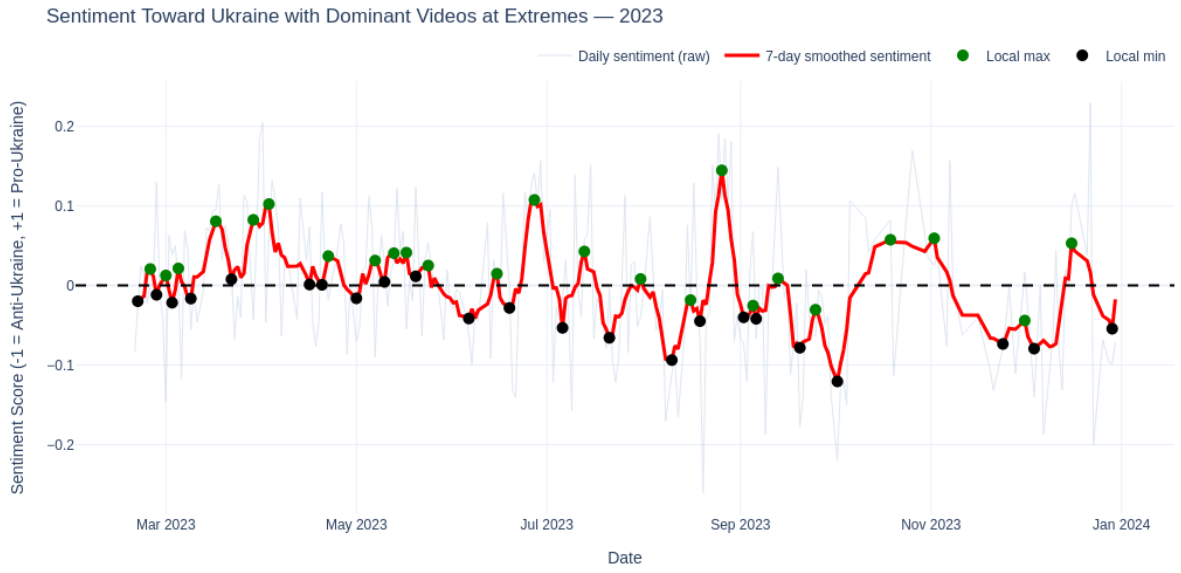

Figure 1: Daily and smoothed sentiment toward Ukraine in 2023. The red curve represents the 7-day rolling average, while the light blue curve shows raw daily sentiment.

Positive sentiment spikes (pro-Ukrainian, toward +1).

Several strong positive sentiment peaks are associated with news that portrays Russia as weakened or isolated, or that highlights international pressure on Russian leadership.

In March and early April 2023, positive spikes correspond to coverage of Russian battlefield losses in Bakhmut and major incidents inside Russia, such as deadly explosions. These topics generated pro-Ukrainian reactions by emphasizing Russian instability and military difficulties.

A notable positive spike in mid-March is linked to news about the International Criminal Court issuing an arrest warrant for Vladimir Putin. This event received strong pro-Ukrainian reactions, likely because it symbolized international accountability and legal pressure on Russia.

In late June and late August 2023, positive sentiment peaks align with coverage of internal Russian political instability, including analysis of Vladimir Putin's public statements and the reported death of Wagner Group leader Yevgeny Prigozhin. These events were widely interpreted as signs of internal conflict and weakening control within Russia, leading to increased pro-Ukrainian sentiment.

Overall, positive sentiment spikes are primarily driven by news emphasizing Russian losses, internal crises, or strong international actions against Russia.

Negative sentiment drops (pro-Russian or critical, toward −1).

Negative sentiment drops are most often associated with news that highlights uncertainty, controversy, or reduced support for Ukraine.

In late July and August 2023, sentiment drops occur during coverage of Ukraine's counteroffensive progress and the delivery of controversial military aid, such as cluster munitions. These topics appear to trigger more critical reactions, possibly due to war fatigue, ethical concerns, or skepticism about military effectiveness.

Strong negative dips in September, October, and December 2023 are closely linked to U.S. political debates about continuing aid to Ukraine. News about potential defunding, delays in military assistance, or internal disagreement within the U.S. government consistently coincides with sentiment moving toward negative values.

Another negative drop is associated with controversial reports suggesting Ukrainian involvement in military actions outside its borders, such as claims related to Sudan. These stories tend to provoke polarized and skeptical responses, reducing overall sentiment toward Ukraine.

Summary of topic-level sentiment patterns.

Taken together, the results show that sentiment on YouTube is highly sensitive to the framing of news topics. Stories that depict Russia as weakened, isolated, or facing internal instability tend to generate pro-Ukrainian sentiment spikes. In contrast, news emphasizing political division in the U.S., uncertainty about aid, ethical controversies, or ambiguous Ukrainian actions are more likely to produce negative or critical reactions.

7.2 2024

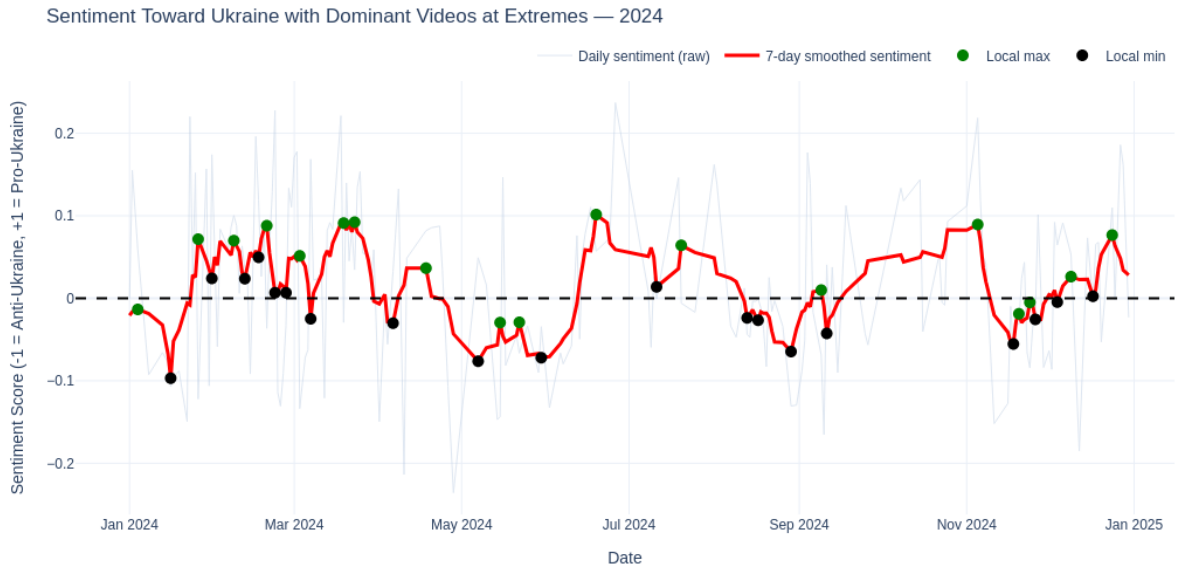

Figure 2: Daily and smoothed sentiment toward Ukraine in 2024. Separating the year highlights short-term sentiment fluctuations and emerging trends.

Compared to 2023, sentiment in 2024 shows more frequent short-term fluctuations, reflecting heightened political tension, debates over Western support, and several high-profile events involving Russian leadership.

Positive sentiment spikes (pro-Ukrainian, toward +1). The strongest positive sentiment peaks in 2024 are primarily associated with news emphasizing Russian internal repression, leadership instability, or increased international support for Ukraine.

In January and February 2024, several positive spikes coincide with coverage related to the death of Alexei Navalny. Videos such as statements blaming Vladimir Putin for Navalny's death and expert commentary describing Putin as "scared" generated strong pro-Ukrainian reactions. These events appear to reinforce narratives of authoritarian repression inside Russia, leading to increased sympathy for Ukraine and criticism of the Russian regime.

Another cluster of positive spikes appears in March and April 2024, driven by news about Putin's pre-election propaganda, warnings from former Russian officials, and international reactions to Russia's internal politics. Content highlighting cracks within the Russian political system or questioning the legitimacy of Putin's leadership consistently aligns with higher pro-Ukrainian sentiment.

During mid-2024, positive peaks are also linked to reports of Ukrainian military actions and symbolic successes, such as Ukraine advancing into Russian territory, gaining access to Ukrainian-held Russian towns, or targeting high-ranking Russian military figures. These stories often frame Ukraine as capable and resilient, prompting supportive reactions from viewers.

Toward the end of the year, renewed positive sentiment emerges around stories of continued Western engagement, including U.S. military support decisions and statements by Ukrainian leadership responding to expanded strike permissions inside Russia.

Negative sentiment drops (pro-Russian or critical, toward −1). Negative sentiment dips in 2024 are closely tied to uncertainty, escalation fears, and political disagreement within the United States.

Early negative drops in January and February are associated with battlefield setbacks and emotionally heavy coverage, such as Ukraine's final defense in Avdiivka and reports emphasizing Russian territorial advances. These narratives tend to reduce optimism and provoke more critical or pessimistic reactions.

Several pronounced negative dips appear in spring and summer 2024 during intense debate over U.S. aid to Ukraine. Videos discussing Ukraine aid as political leverage, Republican efforts to block funding, or alarmist reactions to escalation scenarios (for example, warnings following statements by Western leaders about possible troop involvement) often coincide with sentiment moving toward negative values.

Additional negative reactions are observed in late summer and early autumn around stories

describing large-scale escalation, drone attacks on Kyiv, or ambiguous diplomatic messaging from Russian officials. Such topics appear to increase anxiety and war fatigue rather than clear support for either side.

Summary of topic-level sentiment patterns.

Across 2024, sentiment spikes are strongly topic-dependent. News highlighting Russian repression, leadership weakness, internal dissent, or successful Ukrainian actions tends to produce pro-Ukrainian sentiment. In contrast, coverage focusing on Ukrainian losses, battlefield stalemates, U.S. political conflict over aid, or fears of escalation is more likely to generate negative or critical reactions.

7.3 2025

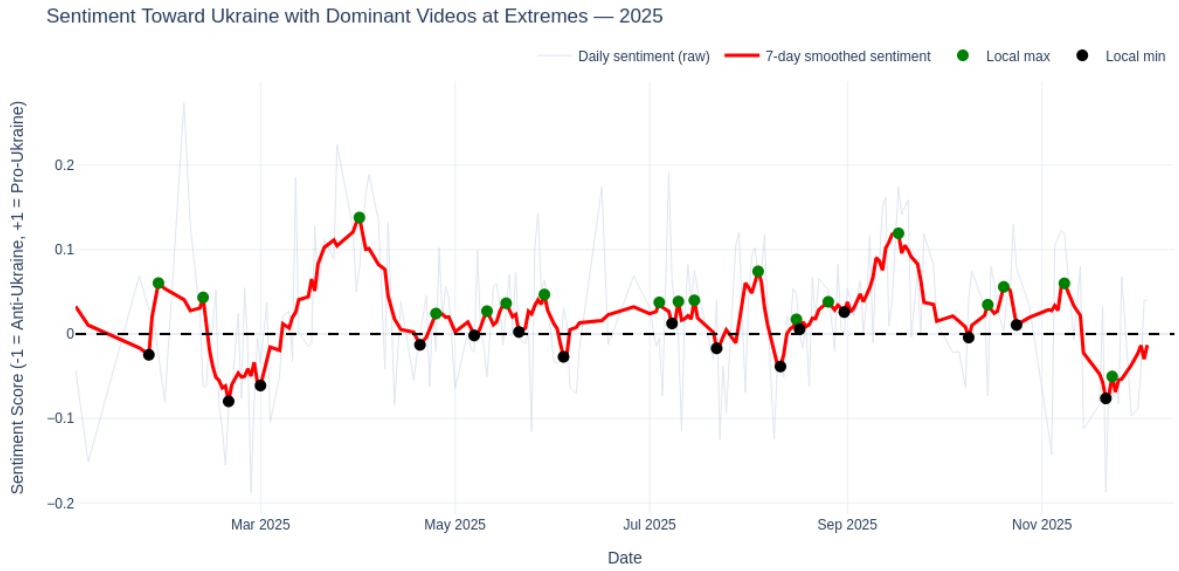

Figure 3: Daily and smoothed sentiment toward Ukraine in 2025. This visualization captures the most recent dynamics in public discourse.

Compared to previous years, sentiment in 2025 is strongly shaped by U.S. domestic political discourse, personal interactions between political leaders, and symbolic military actions rather than large battlefield turning points.

Positive sentiment spikes (pro-Ukrainian, toward +1).

Several of the strongest positive sentiment spikes in 2025 correspond to news framing Russia as morally compromised, politically isolated, or strategically vulnerable.

In late January and February 2025, positive spikes align with coverage of internal Russian repression and opposition, including renewed attention to the Navalny case and public confronta-

tions involving Russian leadership. These stories generated strong pro-Ukrainian reactions, as they reinforced narratives of authoritarian abuse and moral illegitimacy within the Russian regime.

Another set of positive spikes appears in spring 2025 and is associated with reports of Ukrainian military actions and symbolic strikes, such as drone attacks on Russian military infrastructure and incidents involving Russian oil refineries. These events are framed as demonstrations of Ukrainian capability and resilience, which consistently trigger supportive responses in YouTube comment sections.

Positive sentiment is also observed around stories highlighting diplomatic pressure on Russia or public criticism of pro-Russian rhetoric by U.S. politicians. For example, videos showing U.S. officials or former leaders openly condemning Putin or distancing themselves from Russian narratives are linked to noticeable sentiment increases.

Negative sentiment drops (pro-Russian or critical, toward -1).

Negative sentiment dips in 2025 are largely associated with political ambiguity, perceived concessions to Russia, and escalation anxiety.

Several sharp drops occur around news involving direct communication between U.S. political figures and Vladimir Putin, including phone calls, summit discussions, or statements suggesting compromise. Such events often provoke skepticism or frustration among viewers, leading to sentiment shifts toward negative values.

Additional negative reactions are linked to stories emphasizing large-scale Russian attacks on Ukraine or framing the conflict as increasingly costly and unresolved. These narratives tend to amplify war fatigue and uncertainty rather than clear alignment with either side.

Late-year negative dips also coincide with polarizing U.S. political debates, particularly when Ukraine is discussed as part of broader domestic power struggles. In these cases, sentiment becomes more critical and fragmented, reflecting declining emotional engagement and increased polarization.

Summary of topic-level sentiment patterns. Across 2025, pro-Ukrainian sentiment is most strongly associated with content exposing Russian repression, highlighting Ukrainian strikes inside Russia, or showcasing firm political opposition to Putin. In contrast, sentiment turns negative when news centers on political bargaining, ambiguous diplomacy, or repeated cycles of escalation without visible progress.

7.4 Cross-Year Conclusions

Across 2023–2025, sentiment toward Ukraine in U.S. YouTube comments is consistently driven by the framing of news topics rather than by steady long-term trends. Pro-Ukrainian sentiment

spikes most strongly in response to stories that portray Russia as weakened, internally unstable, or morally compromised, such as internal political crises, repression of opposition figures, international legal actions, and successful Ukrainian strikes. These topics create moments of moral clarity and reinforce support for Ukraine.

In contrast, negative sentiment is repeatedly associated with uncertainty and ambiguity. Coverage of U.S. political disagreement over aid, potential concessions to Russia, escalation fears, and prolonged or costly military developments tends to reduce public support and increase critical reactions. Over time, battlefield setbacks and unresolved conflict narratives also contribute to war fatigue, especially when progress is unclear.

While 2023 reactions are more closely tied to battlefield events and major international actions, 2024 and 2025 show a shift toward politically driven sentiment, with U.S. domestic debates and leader-level interactions playing a larger role. Overall, the results indicate that American public sentiment on YouTube is highly event-sensitive and framing-dependent, responding positively to narratives of accountability and resistance, and negatively to uncertainty, compromise, and prolonged conflict without clear resolution.

8 Future Work

There are several directions in which this project can be improved and extended. A natural next step is to experiment with larger transformer models, such as Qwen2.5-7B, by using quantization and offloading techniques. With 4-bit or 8-bit quantization, it becomes more realistic to run a 7B model on a consumer GPU with 8 GB of VRAM, at least for batched inference. This would allow us to compare how a stronger teacher model changes the quality of pseudo-labels and whether it leads to a better student classifier.

The dataset itself can also be expanded. We currently focus on CNN, but adding comments from other U.S. news channels like Fox News, MSNBC and independent media would make it possible to compare audiences and see how sentiment differs between information bubbles. It would also be useful to include data from earlier years and to align the sentiment time series with a detailed timeline of key war events and political decisions.

From a modeling point of view, future work could include fine-tuning DistilBERT (or another compact model) directly on a smaller set of human-labeled comments to correct systematic mistakes from the teacher model. We could also try multilingual or cross-lingual models to handle non-English comments. Finally, combining sentiment analysis with topic modeling or clustering would help to understand not only how people feel, but also what exactly they are discussing when sentiment becomes strongly positive or negative.

9 Conclusion

In this project, we studied how American opinions about the Russia–Ukraine war change over time by analyzing YouTube comments from CNN News. We used transformer-based sentiment models and time aggregation to measure whether comments were mostly positive, negative, or neutral. Our results show that public sentiment on YouTube is very changeable and often reacts strongly to news events. Using two different models gave similar trends, which makes our findings more reliable. Overall, this work shows that YouTube comments can be a useful source for understanding public opinion and the effect of media during the war.

References

[1] A sentiment analysis of the ukraine–russia war tweets using knowledge graph convolutional networks. *International Journal of Information Technology*, January 2025.

[2] Justin Young. Ai sentiment analysis and russia's war in ukraine, 2023.