Stance Detection of Russia-Ukraine War Across Nations on Mainstream News

Anqi Wu
dept. of Statistics
Columbia University
aw3088@columbia.edu

Zhanghao Ni
dept. of Statistics
Columbia University
zn2209@columbia.edu

***Abstract*—This project addressed the challenge of detecting and quantifying international stances during the Russia-Ukraine war with mainstream news articles. Using a dataset of 96,015 articles from outlets like British Broadcasting Corporation (BBC), Cable News Network (CNN), Russia Today (RT), etc., we developed a stance detection framework that classifies countries' positions toward Russia as positive, negative, or neutral. Our approach of corporate data augmentation techniques, including synonym replacement, back-translation, and country-specific text modification, along with fine-tuned RoBERTa models, overcome challenges such as limited labeled data and class imbalance. Key results demonstrated significant improvements in model accuracy and recall, particularly for minority classes, achieved through data augmentation and GPT-assisted labeling. This framework provides insights into evolving geopolitical narratives and demonstrates potential for broader applications in sentiment and stance analysis, particularly for news articles where neutral stances are prevalent and challenging to detect.**

***Index Terms*—Stance Detection, Russia-Ukraine War, Data Augmentation, RoBERTa**

I. Introduction

Where do you acquire world news nowadays? News consumption has changed drastically in recent decades due to the rapid evolution of information dissemination methods. From traditional newspaper and radio podcasts to the emergence of television, and more recently, digital platforms, the way people consume news has undergone a gigantic shift. Newspapers were once the dominated pathway of information that provided detailed global events and regional perspectives, serving as a bridge to connect people around the world. Today, online platforms, social media, and the digital form of mainstream news outlets have become the primary sources of world news, allowing each individual to access global incidents on their own demand.

On 24 February 2022, Russia launched a full-scale invasion of Ukraine, signifying the beginning of one of the most significant international crises of the 21st century. The Russia-Ukraine war not only exacerbated the prolonged geopolitical tensions, but also became the focus of a worldwide media scrum. The complex dynamics of Russia-Ukraine conflicts ranged from military operations, humanitarian crises, to global economic impacts, which elicited a wide discussion. Multiple mainstream news outlets began to closely monitor the ever-evolving situation, scrutinizing official statements, diplomatic efforts, and on-the-ground developments.

The media analysis allowed the audience to understand how conflicts are framed and perceived globally. For example, Media Cloud's study of news narratives regarding the Israel-Hamas conflict(2023) illustrated how media outlets shape public interpretation by highlighting specific themes and narratives over time. Similarly, in the context of the Russia-Ukraine conflict, the nations have expressed their positions through official statements, either condemning or supporting the actions of Russia. However, these declarations often diverged from their economic dealings, strategic movements, and even their political decisions, showing more subtle attitudes. For example, France initially expressed strong support for Ukraine and publicly accused Russia violent actions. However, its continued purchases of Russian natural gas presented a more ambivalent stance toward Russia. Such contradictions highlighted the importance of examining the position of countries not only from the official statements of the nation but also from their diplomatic actions and regional opinion articles, which provide a more accurate representation of their positions.

Our project focused on monitoring and quantifying each country's position over time by analyzing the convergence of diplomacy in mainstream news media, particularly regarding Russia. By quantifying the actual trends and patterns in the media narratives, the project aimed to illustrate the consistency or inconsistency between public statements and media coverage that specifically focused on the Russia-Ukraine war.

II. Related works

Stance detection, a pivotal task in opinion mining, involves identifying an individual's position as favorable, neutral, or opposing toward a given target. Unlike traditional sentiment analysis, which assesses the overall tone of text, stance detection allows identifying attitudes toward a topic often fails to capture the nuanced perspectives in stance detection (Aldayel et al., 2021). The rise of support vector machines (SVMs) and decision trees improved the overall performance of stance detection by employing methods such as bag-of-words and term frequency inverse document frequency(TF-IDF)(Mohammad et al., 2016).

Kareem Darwish et al. (2020) further advanced this field by proposing a robust unsupervised framework that exploits dimensionality reduction and clustering techniques, enabling effective stance detection without requiring manual labeling.

Similarly, Glandt et al. (2020) presented the COVID-19-Stance dataset, which uses tweets related to public health mandates during the pandemic to benchmark stance detection models, exploring self-training and domain adaptation for enhanced performance.

Building upon these innovations, transformer-based architectures like BERT and RoBERTa offered more flexibility and contextual embeddings, which are more suitable for the complicated nature of articles. Liu et al.(2019) emphasized the ability to interpret multilingual social networks with fine-tuned BERT models. Similarly, RoBERTa models showed better performance in capturing subtle context differences. Umit Can et al. (2021) expanded on these advancements and introduced optimization-driven approaches focused on optimization, like-such as hybrid metaheuristics, to tackle stance detection with enhanced accuracy.

Based on the foundation, the current project bridges these methodologies by focusing on stance dynamics in international news articles, specifically focusing on the Russia-Ukraine conflict. Drawing on the advancements of previous researchers, we extended the current project employing data enhancement such as synonym replacement and back-translation to address data scarcity and mitigate data class imbalance. We also utilized fine-tuned RoBerta models to enhance detection accuracy,especially for better contextualization in long-form news articles. Beyond classification, we extend previous research by introducing temporal tracking to analyze the evolution of stances over time to provide deeper insights into the relationship between public statements and media narratives during significant international crises.

III. Data

A. Data Collection

The data for this project consist of 96,015 textual news articles from multiple mainstream media outlets, focusing on coverage related to diplomacy, particularly concerning Russia and Ukraine, between February 2022 and October 2024. The data set was a combination of structured and unstructured text data sourced from five key news outlets.

TABLE I
Article Sources and Collection Methods

Source	*Number*	*Method*
New York Times	29,154	API
The Guardian	34,294	API
BBC News	9,800	Scraped
Russia Today	15,166	Scraped
Cable News Network	7,601	Scraped

To compile the data set, the Guardian and NYT APIs were utilized for direct article retrieval. For Sources without APIs, such as CNN, BBC, and RT, web scraping using Beautiful Soup and Requests was employed. A SQLite3 database was implemented for efficient data storage and fast querying. In addition, URLs were collected through Media.org to enhance the overall coverage across various media outlets. Despite the extensive work on extracting articles from multiple sources, due to privacy restrictions and API limitations, 29154 articles from NYT were only partially presented in this project.

B. Data Preprocessing

The dataset included article metadata, such as title, dates of publication, authors, sources, and full text of the news release. The preprocessing involved removing duplicates, common stop words, punctuation, unnecessary links, and symbols. Furthermore, filtering news articles solely focused on the Russia-Ukraine conflict was also used in the original data set to ensure complete relevance.

C. Special Treatment – Labeling

Since there was no existing data set for training, we performed data labeling ourselves, during the labeling process we constructed rules to label the country-to-country stance on countries shown in the news as positive-negative as well as none, where positive was labeled 1, negative was labeled 2, and none was labeled 0. The owner of the position was the origin country, and the country towards which the position was oriented is the target country.

Considering labor, time, and financial constraints, we chose 918 news articles mentioning France and Germany as our training set to be labeled, and chose Russia as the target country. We believe that France and Germany are relatively representative in terms of changes in their stance towards the Russia-Ukraine conflict in all of the countries. There are 13 countries of origin mentioned in the news that we will be position testing. They were the United States of America, China, Japan, Germany, India, United Kingdom, France, Canada, Italy, South Korea, Saudi Arabia, Spain, and Turkey. These countries were selected due to their significant influence and active role in the geopolitical landscape surrounding the Russia-Ukraine conflict. The target country for the position we selected was Russia, because it is a party to the Russo-Ukrainian war.

During the labeling process, we first constructed a set of basic rules. These rules were incorporated into the GPT prompts, which were used to label the news articles and provide a reasoning for each label. The labels and their corresponding reasoning were manually reviewed to ensure consistency. This review was based on the assumption that GPT's reasoning was accurate. If an incorrect label was identified, we update the labeling rules based on the specific context of the news article. If the label was correct, we proceeded with the labeling process.

IV. Methods

A. Input Modification

Stance analysis is fundamentally a classification problem. Our initial approach to the classification task was to experiment with various classification algorithms to identify the one with the best performance. Since we needed to distinguish between the origin country and the target country, we modified

the input text to help the model recognize the roles of these entities and their directional relationship. The modification was done as follows:

```
Target: {origin_country}'s attitudes
towards {target_country} Text: {news}
```

B. Algorithm

The stance classification task was first approached using traditional machine learning algorithms such as Support Vector Machines (SVM) and Logistic Regression. These algorithms were selected for their simplicity and effectiveness in text classification tasks. However, due to the imbalanced nature of the dataset (with "neither" labels dominating), these models struggled to achieve satisfactory performance. The classification results exhibited low accuracy and recall, particularly for the minority classes ("positive" and "negative" attitudes), as the models were biased toward predicting the majority class.

TABLE II
PERFORMANCE OF SVM AND LOGISTIC REGRESSION

Model	Label	*Precision*	*Recall*	*F1-Score*
SVM	No Attitude	0.64	0.93	0.76
	Positive	0.50	0.08	0.13
	Negative	0.62	0.24	0.35
	Overall Accuracy	0.64		
	Macro Avg	0.59	0.42	0.41
	Weighted Avg	0.62	0.64	0.58
Logistic Regression	No Attitude	0.65	0.92	0.76
	Positive	0.00	0.00	0.00
	Negative	0.61	0.27	0.38
	Overall Accuracy	0.64		
	Macro Avg	0.42	0.40	0.38
	Weighted Avg	0.59	0.64	0.58

Given the limitations of traditional models, we explored transformerbased architectures, specifically BERT and RoBERTa. By incorporating data augmentation, biases were significantly reduced, and both accuracy and recall improved.

After extensive experimentation, RoBERTa was chosen as the final model for its best performance in precision, recall, and F1 scores in all classes, mainly minority labels. The combination of advaced transformer-based models and data augmentation techniques proved essential to achieve reliable and consistent results.

C. Data Augmentation

The size and composition of a training dataset significantly impact the performance of the model, and small datasets often result in suboptimal training results. In this study, the initial dataset contained approximately 50 positive labels and 200 negative labels, with the majority of samples labeled "no." Given the inherent class imbalance and the limited size of the data set, data augmentation techniques were used to enhance the data set and improve the model performance.

a) Oversampling: To address the class imbalance that existed in the data set, oversampling was employed as a resampling strategy to ensure a more balanced distribution of classes in the training data. With class imbalance, a disproportionate representation of classes can bias the model words predicting the majority class, thus overlooking the performance of the minority class(He & Garcia, 2009). In this project, in particular, oversampling was utilized by duplicating samples from the positive and negative attitude class. This approach employed random sampling with replacement using the resample function from sklearn.utils. With a balanced class for our training dataset, the model can effectively learn and contextualize the content from all classes. Random oversampling is a simple but effective approach in handling data imbalance(Chawla et al., 2002), which improved learning efficiency for training dataset and provided more reliable evaluation for the entire dataset.

TABLE III
BERT MODEL PERFORMANCE WITH OVERSAMPLING

Label	*Precision*	*Recall*	*F1-Score*	*Support*
No Attitude	0.62	0.76	0.68	51
Positive	0.86	0.93	0.89	58
Negative	0.75	0.58	0.65	73
Overall Accuracy	0.74			
Macro Avg	0.74	0.76	0.74	182
Weighted Avg	0.75	0.74	0.74	182

b) Synonym Replacement and Back-Translation: To further enhance the diversity of the dataset, synonym replacement and back-translation were employed as data augmentation strategies.

Synonym replacement introduces lexical diversity by substituting selected words in the original text with their synonyms. Using NLTK's WordNet, the algorithm identifies synonyms for each word in the input text and randomly replaces a subset of words with appropriate synonyms, ensuring the replacements are non-identical to the original words. For instance, the sentence "The sanctions imposed on Russia by the European Union aim to pressure the Kremlin" could be transformed into "The penalties imposed on Russia by the European Union aim to coerce the Kremlin" by replacing "sanctions" with "penalties" and "pressure" with "coerce." This method preserves the semantic integrity of the text while introducing lexical variation, enabling the model to generalize across diverse word forms.

Back-translation creates sentence-level diversity by translating a sentence into another language and subsequently back into the original language. This process often results in slight variations in sentence structure or word choice, while maintaining the original meaning. For example, the sentence "The United States has pledged additional military aid to Ukraine in its fight against Russian aggression" might be back-translated as "The United States has committed further

military assistance to Ukraine to combat Russian hostility." Employing Google Translator's capabilities, this approach introduces syntactic and semantic diversity, which enhances the model's robustness and ability to generalize to unseen data.

The application of synonym replacement and back-translation resulted in significant improvements in precision and recall for both BERT and RoBERTa models. Notably, these techniques enhanced the prediction accuracy for the negative class, outperforming oversampling alone.

TABLE IV
PERFORMANCE OF BERT AND ROBERTA WITH SYNONYM REPLACEMENT AND BACK-TRANSLATION

Model	Label	*Precision*	*Recall*	*F1-Score*
BERT	No Attitude	0.65	0.80	0.72
	Positive	0.81	0.96	0.88
	Negative	0.91	0.77	0.83
	Overall Accuracy	0.81		
	Macro Avg	0.79	0.84	0.81
	Weighted Avg	0.82	0.81	0.81
RoBERTa	No Attitude	0.65	0.85	0.73
	Positive	0.82	1.00	0.90
	Negative	0.92	0.73	0.81
	Overall Accuracy	0.80		
	Macro Avg	0.79	0.86	0.82
	Weighted Avg	0.83	0.80	0.80

c) Country Name Replacement: Given the study's objective of analyzing the positions of various countries toward Russia, but the limited availability of stance labels for only France and Germany, a novel augmentation strategy was devised. Specifically, terms representing France or Germany, such as the names of leaders, capitals, or other country-specific entities, were systematically replaced with equivalent terms from other countries. This method effectively expanded the dataset by simulating stance data for additional countries, introducing geographical diversity while maintaining contextual relevance. The augmentation significantly enriched the dataset, enabling the model to learn from a broader spectrum of geopolitical scenarios.

TABLE V
ROBERTA MODEL PERFORMANCE WITH OVERSAMPLING

Label	*Precision*	*Recall*	*F1-Score*	*Support*
No Attitude	0.99	0.99	0.99	1594
Positive	0.96	0.99	0.97	841
Negative	0.99	0.98	0.99	2463
Overall Accuracy	0.99			
Macro Avg	0.98	0.99	0.98	4898
Weighted Avg	0.99	0.99	0.99	4898

V. SYSTEM OVERVIEW

A. Application Design and Technology Stack

The application is a web-based interactive dashboard designed to analyze and visualize international positions during the Russia-Ukraine conflict. It integrates multiple technologies and tools to provide users with an in-depth understanding of the perspectives of the country and the media. The front-end is built using HTML, CSS, and JavaScript, with D3.js handling interactive data visualizations. The backend utlized Python for data preprocessing, augmentation, and analysis pipelines, while Google Cloud Platform is used for data storage and access. Machine learning models, specifically RoBERTa, are employed for stance detection, trained on augmented datasets using Hugging Face Transformers. Additional tools like Pandas and NLTK facilitate data handling and linguistic processing. For data orchestration and scalability, Airflow manages workflows, and Spark supports data processing. The entire application is supported by robust system resources, including GPUs (A100) and CPUs for efficient computations.

B. Challenge and Proposed Solutions

The bottleneck lies mainly in the limited size of training data, which can lead to biased model training. This limitation stems from the fact that the labeling process is currently done manually, which is labor intensive and time consuming. Additionally, the labeled data are restricted to only two origin countries, France and Germany, with Russia as the sole target country. This narrow scope further impacts the diversity and generalizability of the dataset.

To address the challenge of limited training data, the use of GPT APIs for automated labeling can significantly enhance efficiency and accuracy. By iteratively refining labeling rules and expanding the range of origin countries, the dataset can become more comprehensive, improving the model's robustness, and enabling precise stance detection across diverse international contexts. Additionally, adopting a self-training or bootstrapping approach can further address this limitation. In this method, the model is initially trained on the small-labeled dataset and then predicts labels for unlabeled data. The most confident predictions are added to the training set for retraining, and this process is repeated iteratively. This approach gradually expands training data, reduces the reliance on manual labeling, and maximizes the use of available data for improved model performance.

C. Interactive Features and User Experience

The application is user-friendly and interactive. Users can explore various tabs, including 'About Project', which provides an overview of the project's goals and methodologies.

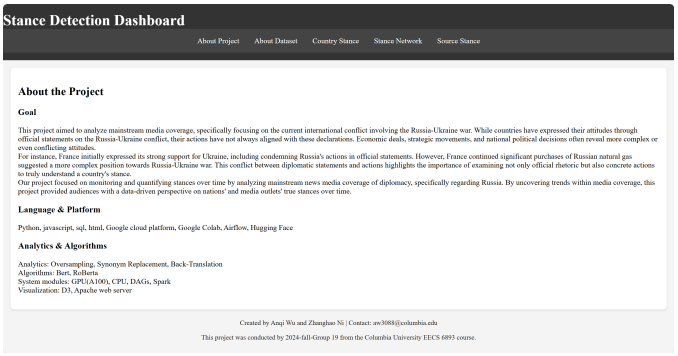

Fig. 1. 'About Project' section

The "About Dataset" section allows users to explore trends in Ukrainian-Russia-related news coverage over time. A dropdown menu enables users to filter by news sources, such as CNN or BBC, and the line chart dynamically updates to reflect the selected sources, displaying the number of articles over time.

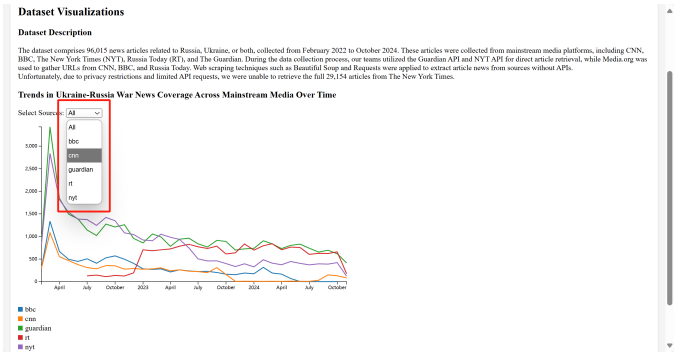

Fig. 2. 'About Dataset' section

The "Country Stance" section provides tools to analyze stance trends of various countries toward Russia. Users can select countries by checking boxes and adjust the weights of neutral, positive, and negative positions using numeric input fields labeled "a", "b," and "c." The formula of stance score is shown above. The chart updates in real-time to display the stance trends based on the selected countries and weights.

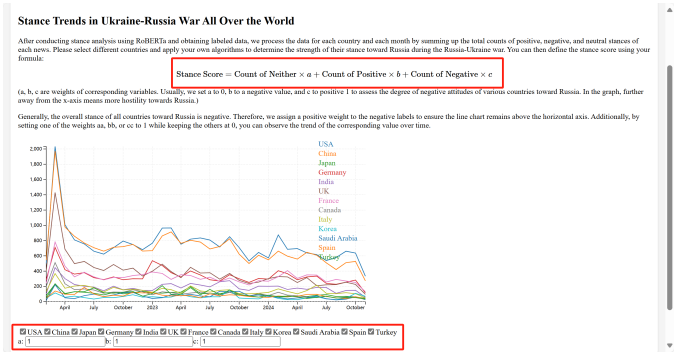

Fig. 3. 'Country Stance' section

In the "Stance Network" section, users can visualize the relationships between countries and their stances toward Russia through an interactive network graph. By adjusting the sliders for parameters such as positive and negative stance weights, users can modify the way stance scores are calculated. The network graph updates dynamically, with nodes representing countries and links reflecting their stance scores toward Russia. Hovering over the nodes provides detailed stance score information.

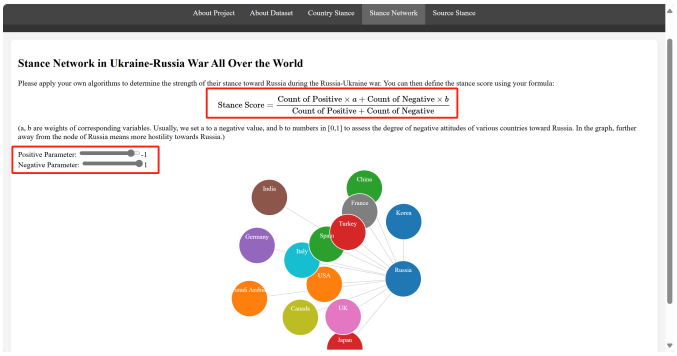

Fig. 4. 'Stance Network' section

The "Source Stance" section enables users to analyze how the stances of media outlets toward Russia have changed over time. By selecting a source from the dropdown menu, users can view a stacked area chart showing the distribution of neutral, positive, and negative stances for that source. Hovering over the chart reveals detailed percentages for each stance at different time points.

Fig. 5. 'Source Stance' section

VI. Experiments

To demonstrate the effectiveness of our approach, we conducted several experiments aimed at evaluating performance, comparing it to baseline methods, and analyzing the impact of various components. First, we compared our method, which incorporates advanced data augmentation techniques and fine-tuned BERT and RoBERTa models, against a baseline approach using oversampling. Although oversampling addressed data imbalance, it lacked the diversity needed for robust training. Our data augmentation techniques, specifically Synonym

Replacement and Back-Translation, introduced lexical and semantic diversity, resulting in significant improvements in precision, recall, and F1 score. This was especially evident for negative position labels, where our method outperformed the baseline by a wide margin.

To better understand the role of data augmentation, we analyzed model performance with and without augmented data. Without augmentation, the limited dataset size led to under performance, particularly for minority labels like "negative." Incorporating data augmentation significantly improved model accuracy and recall across all labels, as the expanded and diversified dataset allowed the model to generalize better. This experiment clearly highlighted the importance of augmentation in addressing data scarcity and improving robustness.

We also conducted hyperparameter tuning to optimize the model performance. Key hyperparameters such as learning rate, batch size, and number of training epochs were tested. The best results were achieved with a learning rate of 2e-5, a batch size of 16, and 5 epochs. These parameters provided a balance between training stability and efficiency, as overly high learning rates caused instability, and smaller batch sizes prolonged training without yielding noticeable improvements. This tuning process demonstrated the sensitivity of model performance to hyperparameter choices and the value of systematic experimentation.

Finally, we evaluated the system's end-to-end performance by visualizing stance trends across countries and media sources. Using interactive line charts and network graphs, we illustrated how stances evolved over time and how different countries and media outlets positioned themselves relative to Russia. The visualizations validated the effectiveness of the system in capturing and presenting stance dynamics, offering clear and actionable insights. In general, these experiments demonstrated that our approach successfully addresses the challenges of stance detection and provides a robust framework for understanding complex international relations.

VII. Conclusions

This project achieved significant progress in understanding and quantifying international stances during the Russia-Ukraine conflict. By leveraging data augmentation techniques like Synonym Replacement and Back-Translation, coupled with advanced models such as BERT and RoBERTa, we effectively addressed challenges posed by a limited dataset, improving performance across all metrics.

Key challenges included the narrow scope and size of the labeled dataset, which introduced training bias. To overcome these, we employed GPT-based automated labeling, iterative rule refinement, and expanded the range of analyzed countries, resulting in a more robust and comprehensive system.

Future extensions could involve adding more countries, incorporating multilingual capabilities, and integrating advanced models such as GPT-4 for improved scalability and accuracy. Beyond geopolitical analysis, this framework can be adapted for corporate sentiment analysis, public opinion tracking, and other applications, demonstrating its broad utility and potential for impact.

References

[1] S. M. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "SemEval-2016 Task 6: Detecting Stance in Tweets," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, CA, USA, 2016, pp. 31–41. DOI: 10.18653/v1/S16-1003.

[2] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019. DOI: 10.48550/arXiv.1907.11692.

[3] K. Darwish, P. Stefanov, M. Aupetit, and P. Nakov, "Unsupervised User Stance Detection on Twitter," *arXiv preprint arXiv:1904.02000*, 2020. DOI: 10.48550/arXiv.1904.02000.

[4] A. ALDayel and W. Magdy, "Stance detection on social media: State of the art and trends," *Information Processing & Management*, vol. 58, no. 4, p. 102597, 2021. DOI: 10.1016/j.ipm.2020.102597.

[5] K. Glandt, S. Khanal, Y. Li, D. Caragea, and C. Caragea, "Stance Detection in COVID-19 Tweets," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online, 2021, pp. 1596–1611. DOI: 10.18653/v1/2021.acl-long.127.

[6] U. Can and B. Alatas, "A novel approach for efficient stance detection in online social networks with metaheuristic optimization," *Technology in Society*, vol. 64, 2021. DOI: 10.1016/j.techsoc.2020.101428.

[7] Media Cloud, "How the News Talked About the Israel-Hamas Conflict in Its First Month," Media Cloud, 2023. Available: https://www.mediacloud.org/research/how-the-news-talked-about-the-israel-hamas-conflict-in-its-first-month. [Accessed: 21-Dec-2024].

[8] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sept. 2009. DOI: 10.1109/TKDE.2008.239.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, June 2002. DOI: 10.1613/jair.953.