Multi-Label Narrative Classification of Online News: Handling Large Label Spaces with Limited Data

Leon Kogler, Ines Theron, Raj Vasani
Technical University of Munich
Chair of Data Analytics and Statistics
School of Computation, Information and Technology

Abstract

In this study, we tackle the challenge of multi-label narrative classification as part of SemEval-2025 Task 10, focusing on news articles related to the Ukraine-Russia War and Climate Change. The task requires assigning multiple narrative and sub-narrative labels to articles, making it a complex multi-label classification problem. To address this, we explore various machine learning approaches, including logistic regression, SVMs, Random Forest, XGBoost, and transformer-based models such as BERT and RoBERTa. Our best-performing approach utilizes fine-tuned sentence transformer embeddings (all-MiniLM-L6-v2) combined with a Logistic classifier, achieving strong performance on narrative classification (Macro F1-score: 0.46) and sub-narrative classification (Macro F1-score: 0.30) on the test set. Despite challenges like label imbalance, data inconsistencies, and the need for large-scale training data, we demonstrate that high-quality data embeddings significantly improve performance, with high-quality labeled news articles outperforming paraphrased and translated ones. Our code is available here: Github

Keywords: Narrative Classification, Disinformation Detection, Multi-label Learning, Transformers

1 Introduction

In today's digital era, the web has become the primary source of information for people seeking to understand global challenges (World Economic Forum, 2022). However, this shift also complicates fact-checking and combating misinformation as information sources have become more diverse and decentralized. Identifying and classifying narratives is, therefore, crucial for detecting media bias and curbing disinformation.

This project, conducted for SemEval-2025 Task 10, tackles the problem of narrative classification in news articles flagged for potential misinformation. These articles are sourced from various web platforms and fall into two major categories: the Ukraine-Russia War (URW) and Climate Change (CC). The challenge lies in the multi-label nature of the classification, where each article can be assigned multiple narratives and sub-narratives. The dataset presents challenges such as class imbalance, data quality issues, and long articles, which demand robust modeling techniques. Our objective is to build a classifier capable of accurately predicting narrative labels.

The development of different text classification models in recent years has given us a multitude of different approaches to choose from. To achieve our goal, we explore a broad range of machine learning and deep learning techniques. We start with some classical models such as logistic regression, SVMs, and Random Forest (Mirończuk and Protasiewicz, 2018). We also used deep Neural networks like Feed Forward Neural Network (FNN), CNN, and LSTM. These helped us establish baseline performance and evaluate their strengths in multi-label classification. These methods also gave us a better understanding of this task's challenges and data. Building on these results, we incorporate transformer-based models such as BERT and RoBERTa, which leverage deep contextual embeddings to capture semantic nuances in multilingual articles. Additionally, we experiment with different word embedding techniques, such as sentence transformers and multilingual embeddings, to optimize model performance. Among all these approaches, the three best-performing methods were:

- Logistic Regression with Sentence Transformer Embeddings (all-MiniLM-L6-v2): One of the most basic classifier models paired with high-quality embeddings shows impressive results when dealing with high unbalanced and low quantity of training data.

- XGBoost with Sentence Transformer Embeddings (all-MiniLM-L6-v2): XGBoost stands for extreme Gradient Boosting and is based on decision trees. XGBoost outperforms the deep learning model, especially in a good missing value handling.

- Feed Forward Neural Network (FNN) with Sentence Transformer Embeddings (all-MiniLM-L6-v2): This approach uses the neural network learning algorithm, which receives sentence embeddings in the input layer and assigns the labels in the output layer.

This report presents our three best-performing models, detailing their development and the challenges encountered. It will cover data characteristics and challenges, word embedding techniques, classifier architectures, and the optimization and data augmentation methods applied.

Through this report, we aim to provide a comprehensive overview of our project, highlighting key insights and emphasizing the critical role of quality training data and the challenges of multi-label classification.

2 Dataset

This section introduces the dataset, describes its features, and discusses its limitations and challenges for model use.

2.1 Data Characteristics

In the context of this project, we obtained different text datasets. This section shows how this data is structured. Each article was provided with annotations, with every article receiving two different labels: narrative and subnarrative. Each label is either assigned to Climate Change *CC* or Ukraine Russian War *URW*. Each subnarrative is a subannotation derived from the narrative. To better understand, you can look at Table 1, which shows how the annotation is structured, one column being article_id, another possible narrative, and the last possible subnarratives.

2.2 Statistics

To better understand the data structure, we examine key statistics. We had a total of 399 articles. 2 gives an overview of the number of Narratives and Subnarratives for each category. We also had a total of 169 *"Other"* labeled narrative articles and 272 articles subnarrative labeled as *"Others"*.

article_id	**narrative**	**subnarrative**
EN_10001	*URW: Blaming others*	*URW: Ukraine is the aggressor*
EN_10002	*URW: Blaming others*	*URW: Blaming Others: Others*
EN_10003	*Others*	*Others*
EN_10004	*URW: Russia is the victim; URW: Blaming others*	*URW: Blaming Others: Others; URW: Russia is the victim: The West is russophobic*

Table 1: Example for annotation of articles

	CC	**URW**
Articles Count	103	128
Unique Narratives	11	11
Unique Sub-Narratives	42	42
Max Narratives assigned	10	11

Table 2: Summary of URW and CC Labels

2.3 Multi-label Nature

Another particularity is that each article can be assigned to multiple narratives and subnarratives as viewable in row 4 from Table 1. This type of classification is called a multilabel task. A typical classification task assigns the target variable to only one class. However, in a multilabel classification task, the target variable can be assigned to multiple classes simultaneously (Bogatinovski et al., 2022). Therefore, multiple classes should be detected for each article when predicting the classification of these articles.

2.4 Imbalance

The number of available articles for the task was relatively low compared to the number of unique labels the model was supposed to be trained on. For some subnarratives, there was only one occurrence in the entire dataset, such as *CC: Amplifying Climate Fears: Doomsday scenarios for humans*. Although narrative representation improved, it remained suboptimal, with some narratives appearing only five times. Any dataset with unequal class distribution can be considered imbalanced (Ramyachitra and Manikandan, 2014). An imbalanced dataset and the lack of representation of some classes over others can lead to a lack of information for an algorithm to learn from (Ramyachitra and Manikandan, 2014). For a clearer visualization of this imbalance, refer to Figures **??** and **??**.

2.5 Data Quality Issues

Another observation that stood out was when analyzing the annotations of the articles. The labels are highly specific and often difficult to differentiate. To visualize this issue, consider the following example: one narrative is *"Amplifying Climate Fears"* with the subnarrative *"Whatever we do, it is already too late"*, while another has the subnarrative *"Amplifying existing fears of global warming"*. The difference between these specific narratives and subnarratives is subtle and can be difficult to distinguish, even for humans. This increases the complexity of the classification task.

2.6 Long Articles

Article length is crucial, as it impacts classification differently depending on the task. Text length plays a critical role in some embedding methods, such as BERT or similar transformer-based embeddings. These models have a sequence length limitation (512 tokens for BERT) (Ding et al., 2020). When the sequence length exceeds this limit, articles must be truncated or split, which may result in losing important information. Our dataset has an average article length of approximately 477 words. Among the 399 articles, 159 articles (around 40%) exceed the BERT token limit, meaning a significant portion of the dataset is affected by this constraint.

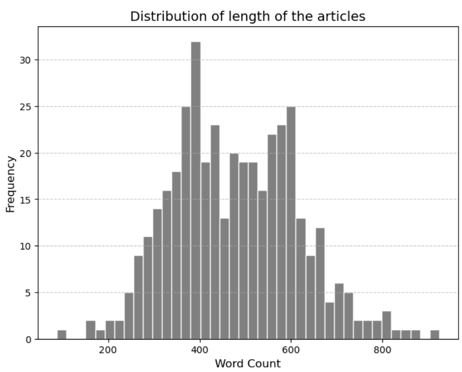

Figure 1: Distribution of article length

3 Explored Methods

In the course of the project, we tried to combine different embedding models with various classifier models. Based on the models we explored, we tried to get an understanding of which of these approaches worked the best with our limited Data.

3.1 Embeddings

Our first approach to generating meaningful Embeddings for our Articles was to use traditional approaches like TF-IDF (Jones, 1972), Word2Vec (Mikolov et al., 2013), or GloVe (Pennington et al., 2014). These embedding models can be trained by the corpus of articles we got for training. After that, we got vector representations for all words in the articles based on their context. To get a vector representation of the whole article, not only about single words, we used the technique of mean pooling, where the mean of all word embeddings is calculated to represent the whole article.

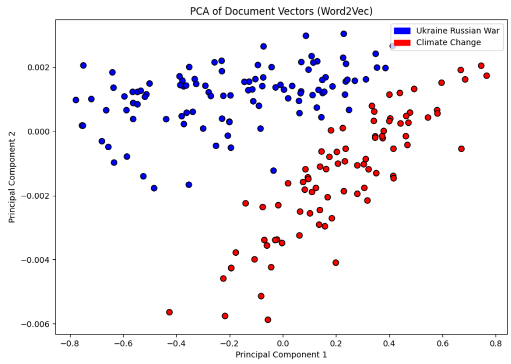

Figure 2: Mean pooled, 300-dimensional Embeddings generated by Word2Vec model using context window of size 10 showing the difference between "Climate Change" and "Ukraine Russian War" articles

As shown in Figure 2, Word2Vec with mean pooling already helps categorize articles into Climate Change and Ukraine-Russian war topics. However, training an embedding model solely on our dataset may lack sufficient contextual information. A potential solution is to use pre-trained embeddings like 'glove-wiki-gigaword-300', which are trained on 6 billion tokens. Yet, these embeddings do not capture word context, leading to ambiguity.

To address this, we used BERT, a Transformer-based model that generates contextualized embeddings, allowing words to have different representations depending on context. While mean pooling helps create article-level embeddings, it also averages out key semantic information, potentially degrading representation quality.

A solution to the problems raised by word embedding models is to use SBERT (Sentence-BERT) (Reimers and Gurevych, 2019) and the sentence-transformers library. Sentence embedding models like SBERT are pre-trained to understand sentence meanings as a whole. A big issue we faced using SBERT is its limited context size of 128 tokens. Therefore, we used the 'all-MiniLM-L6-v2' sentence transformer (Sentence-Transformers) that generates 384-dimensional embeddings with a maximum sequence input length of 256 tokens.

3.2 Classifiers

In the context of this work, we test different classifiers; this section will give a short overview of which classifiers were tested in this project.

With combinations of basic machine learning and more advanced deep learning models with several embedding approaches, as seen in 3.1, we try to gain an understanding of what classifiers work best with what type of embedding models.

Logistic Regression We experimented with multiple models for multi-label classification. Initially, we used logistic regression with TF-IDF features and fine-tuned Word2Vec embeddings, incorporating a hierarchical structure for sub-narrative classification. Later, we transitioned to sentence embeddings from all-MiniLM-L6-v2, which further improved results.

SVM We also implemented a Support Vector Machine (SVM) for the classification task. For the article embeddings, we used Word2Vec, which likely performs better than TF-IDF in this context, as it captures semantic relationships between words (Rahmawati and Khodra, 2016). However, SVM did not perform well for our task, which may be due to its incompatibility with multi-label classification, particularly in the presence of imbalanced labels (Fu et al., 2015).

Feed-Forward Neural Network (FNN) We also implemented a Feed-Forward Neural Network (FNN) for multi-label classification (Nam et al., 2014). Sentence-BERT embeddings were inputs, followed by three hidden layers with 256, 128, and 64 neurons, each with ReLU activation and dropout regularization (0.4). The output layer applied sigmoid activation to predict multiple labels. The model was trained with the Adam optimizer (learning rate = 0.01) and Binary Cross-Entropy (BCE) loss. After threshold tuning, the FNN achieved an F1 Macro score of 0.2593 for narratives and 0.0949 for sub-narratives, but it underperformed compared to XGBoost and logistic regression.

CNN We also tested a Convolutional Neural Network (CNN) for our task (Kim, 2014). We used a multi-label binary classifier for the labels and Word2Vec embeddings for the articles. We used two separate fully connected (FC) layers with sigmoid activation to address the multi-label classification, as the labels are hierarchical. For the loss function, we used Binary Cross-Entropy (BCE) loss. CNN didn't work as well as the other methods, maybe because it is not as suited as transformer-based methods (Soyalp et al., 2021).

LSTM We implemented a Long Short-Term Memory (LSTM) network, introduced by Hochreiter and Schmidhuber (Hochreiter and Schmidhuber, 1997), which is well-suited for capturing sequential dependencies in text and can improve classification performance. Our implementation utilized pre-trained GloVe (glove.6B.300d) embeddings with mean pooling. However, our experiments showed that an LSTM combined with a fully connected linear layer at the output performed only slightly better than the traditional machine learning classifiers mentioned earlier.

BERT Finally, we evaluated a transformer-based approach utilizing the bert-base-uncased model. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) is designed to capture deep contextual word representations, making it well-suited for text classification tasks. We fine-tuned the model on our dataset, employing a classification head with a sigmoid activation function for multi-label classification. While this model initially appeared promising, it exhibited poor performance on our limited training dataset. Despite extensive hyperparameter tuning, we were unable to achieve satisfactory results.

XGBoost Approach We trained an XGBoost-based multi-label classification model using fine-tuned Sentence-BERT's text embeddings as input features. We computed scale position weights to handle class imbalance and incorporated them into the model training process. The model was optimized using threshold tuning to select the optimal threshold, leading to improved F1 macro scores for both narrative and sub-narrative predictions (Tang et al., 2019) (Zhang and Li, 2020) (Awan et al., 2022).

3.3 Data Augmentation

To enhance the diversity of our dataset, we employed the following data augmentation methods:

Synonym Replacement: We used a RoBERTa-based (Liu et al., 2019) masked language model to generate contextually appropriate synonyms for nouns and verbs in the text (Wei and Zou, 2019). The process involved: (1) Identifying keywords using spaCy. (2) Replacing words with synonyms predicted by RoBERTa. (3) Filtering augmented samples based on cosine similarity to ensure semantic preservation. (4) Retaining only those samples with a cosine similarity score above 0.90 before adding them to the dataset.

This method aimed to increase the model's accuracy while maintaining data integrity but couldn't improve the performance due to the lack of diversity in our data. We had some labels with only one article associated with them and some labels with more than 10 articles associated with them. This made it difficult for the model to classify since it had seen only one type of article for some labels while training.

Conditional Text Generation using GPT-4o: Given the challenge of limited examples for unique narrative-sub-narrative combinations, we leveraged ChatGPT-4o to generate additional training samples (Keskar et al., 2019). The approach includes: (1) Prompting ChatGPT-4o with specific narratives and sub-narratives. (2) Generating three contextually relevant news articles for each input set. (3) Formatting the output as a CSV file with columns wrapped in double quotes for seamless data ingestion.

This method creates synthetic training data tailored to underrepresented narrative combinations to enhance model generalization. Still, this method did not improve the model performance since GPT-4o did not generate high-quality news articles. The use of artificially generated articles actually degraded the performance of the models.

4 Methodology

In this section, we present the three best-performing models and detail the methodology behind each approach. Our results indicate that high-quality sentence embeddings paired with simpler classifier models consistently achieved the highest F1 scores, outperforming more complex architectures such as FNN, LSTM-based neural networks, CNN, and BERT classifiers.

Logistic Regression with all-MiniLM-L6-v2 Embeddings This was our best performing model for multi-label classification. Initially, we employed a multi-label classification approach using logistic regression with class weighting to address class imbalance. The text data was first transformed into TF-IDF features and fine-tuned Word2Vec embeddings, which were used both independently and in an ensemble setup. A hierarchical structure was incorporated, where narrative predictions guided sub-narrative classification, and the model was evaluated using stratified k-fold cross-validation. However, this approach yielded limited performance, with F1 macro scores around 0.2 for sub-narratives when using Word2Vec embeddings.

To improve results, we replaced the ensemble approach with sentence embeddings generated by the all-MiniLM-L6-v2 transformer model, which increased the F1 macro score to 0.3. Surprisingly, this combination of sentence embeddings with a simple logistic regression classifier outperformed more complex deep learning approaches, such as BERT-based models.

XGBoost Classifier Our second-best approach was an XGBoost-based approach using Sentence-BERT embeddings as input features. The embeddings were generated from preprocessed news articles and then used to train two separate XGBoost classifiers: one for predicting narratives and another for sub-narratives. To address the class imbalance, we computed scale position weights based on positive-to-negative label ratios and integrated them into the training process. Each classifier was implemented as a multi-output model using XGBoost with hyperparameters set to a learning rate of 0.05, a maximum depth of 6, and 200 boosting iterations. The objective function was binary logistic regression, and we used 'logloss' as the evaluation metric. A threshold tuning strategy was applied to optimize classification performance, leading to improved F1 macro scores for both narrative and sub-narrative predictions. This approach aligns with previous studies exploring the effectiveness of gradient-boosting models in multi-label text classification (Tang et al., 2019) (Zhang and Li, 2020) (Awan et al., 2022).

Feed-Forward Neural Network (FNN) Our third-best approach was a Feed-Forward Neural Network (FNN). The FNN architecture consisted

Classifier	Embeddings	F1-Macro coarse	F1-Macro fine
XGBoost	all-MiniLM-L6-v2	0.43	0.15
Logistic Regression	all-MiniLM-L6-v2	0.46	0.30
Logistic Regression	Word2Vec + TF-IDF	0.40	0.20
Feed Forward Neural Network	all-MiniLM-L6-v2	0.26	0.10
Bert-base-uncased	BERT	0.12	0.05
LSTM + linear layer	glove.6B.300d	0.16	0.05
SVM	Word2Vec + mean pooling	0.17	0.03
CNN	Word2Vec + mean pooling	0.13	0.04

Table 3: Results in F1-Macro score on both narratives (coarse) and sub-narratives (fine), evaluated on different classifier and embedding combinations.

of an input layer matching the dimensionality of the embeddings, followed by three hidden layers with 256, 128, and 64 neurons, respectively, each utilizing ReLU activation functions. Dropout regularization with a rate of 0.4 was applied to each hidden layer to mitigate overfitting. The output layer had neurons corresponding to the number of labels, with sigmoid activation functions to produce probability scores for each label. The network was trained using the Adam optimizer with a learning rate of 0.01 and the Binary Cross-Entropy loss function. After training, the FNN achieved an F1 Macro score of 0.2472 for narratives and 0.0939 for sub-narratives using non-optimal thresholds. By adjusting the thresholds to 0.12 for narratives and 0.04 for sub-narratives, we observed improved F1 Macro scores of 0.2593 and 0.0949, respectively. This approach aligns with methodologies discussed in prior research on multi-label text classification using neural networks (Nam et al., 2014).

5 Results

5.1 Quality of Embeddings

To assess the quality of different embedding models, we selected the ten most frequent narratives from our dataset and retrieved the corresponding articles for embedding analysis. After generating document embeddings for all articles with the model under evaluation, we computed the centroid vectors for clusters of articles labeled with a specific narrative. The cosine similarity function was then applied across all narratives, resulting in a similarity matrix. The mean cosine similarity of this matrix served as an indicator of embedding model quality, where lower mean values suggest better performance in distinguishing between different narratives and sub-narratives.

Model	Mean Cosine-Similarity
all-MiniLM-L6-v2	0.51
SBERT	0.88
glove.6B.300d + mean pooling	0.93
DistilBERT + mean pooling	0.96
Word2Vec + mean pooling	0.99

Table 4: Average Cosine-Similarity of centroids of document embeddings with same narratives for different Models. Higher is better.

As shown in Table 4, sentence embedding models outperform static models like Word2Vec and glove.6B.300d (both using mean pooling) in distinguishing articles with different narratives. Notably, even DistilBERT with mean pooling over the last hidden layer does not improve embedding quality compared to non-transformer models.

5.2 Classifier Performance

To evaluate different classifiers for narrative classification, we tested a range of machine learning and deep learning models with various embeddings, summarized in Table 3. Logistic regression and XGBoost performed strongly when paired with sentence embeddings (e.g., all-MiniLM-L6-v2), achieving F1-macro scores of up to 0.46 for narratives and 0.30 for sub-narratives. In contrast, FNN, CNN, and LSTM performed poorly, indicating that simpler models with high-quality embeddings can outperform more complex architectures in this multi-label setting. SVM and fine-tuned BERT also yielded weaker results, likely due to label imbalance and limited data. Overall, these findings highlight the importance of embedding quality and suggest that methods like logistic regression offer a good balance of interpretability and classification power.

6 Conclusion

Our findings highlight the effectiveness of combining deep contextual embeddings with efficient classifiers, striking a balance between accuracy, interpretability, and computational efficiency. Future work should focus on expanding datasets to include high quality data, refining labeling taxonomies, and using large language models for improved narrative inference. Our results contribute to advancing multilingual news classification and the detection of manipulative narratives online using efficient classifiers.

7 Contributions

7.1 Leon Kogler

I implemented and analyzed LSTM with pre-trained GloVe embeddings, trained a random-forest classifier, fine-tuned a BERT-based model, and worked on logistic regression with sentence transformer embeddings (all-MiniLM-L6-v2). I conducted an embedding analysis of Word2Vec, GloVe, SBERT, and sentence transformers using cosine similarity metrics. Additionally, I contributed to exploratory data analysis (EDA), examining dataset structure, label distribution, and preprocessing challenges. I also worked on data augmentation, generating new articles for narratives using LLaMA 3.3.

7.2 Ines Theron

My contributions include the development of CNN using Word2Vec embeddings, as well as the implementation of an SVM classifier with Word2Vec embeddings. I also experimented with TF-IDF embeddings as an alternative to Word2Vec for both CNN and SVM models. Additionally, I worked on the initial BERT-base model. I conducted EDA to gain insights into the text characteristics, which informed our preprocessing. I also explored data augmentation with synonym replacement. In the report, I wrote the introduction and dataset sections, as well as the explanations of CNN and SVM in the classifier parts.

7.3 Raj Vasani

I implemented the initial logistic regression ensemble approach using TF-IDF and GloVe embeddings, as well as the Feed-Forward Neural Network (FNN) and XGBoost classifiers. Beyond developing these models, I conducted basic exploratory data analysis (EDA) and contributed to data augmentation by incorporating synonym replacement and conditional text generation with ChatGPT. Additionally, I fine-tuned a RoBERTa model to generate embeddings for comparative evaluation across various classifiers.

References

Muhammad Awan, Ahmad Zain, and Muhammad Sohaib. 2022. XGBoost-based Multi-Label Classification for News Articles. *IEEE Access*, 10:112345–112356.

Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. 2022. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems*, 33:12792–12804.

Di Fu, Bo Zhou, and Jinglu Hu. 2015. Improving svm based multi-label classification by using label relationship. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Karen Spärck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

Yoon Kim. 2014. Convolutional neural networks for sentence classification.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Marcin Michał Mirończuk and Jarosław Protasiewicz. 2018. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106:36–54.

Jinseok Nam, Jason Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. 2014. Large-scale multi-label text classification—revisiting neural networks. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2800–2806.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Dyah Rahmawati and Masayu Leylia Khodra. 2016. Word2vec semantic representation in multilabel classification for indonesian news article. In *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pages 1–6.

Duraisamy Ramyachitra and Parasuraman Manikandan. 2014. Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, 5(4):1–29.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Sentence-Transformers. all-minilm-l6-v2. Accessed: 2025-02-28.

Gokhan Soyalp, Artun Alar, Kaan Ozkanli, and Beytullah Yildiz. 2021. Improving text classification with transformer. In *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pages 707–712.

B. Tang, S. Kay, and H. He. 2019. Xgboost for multi-label classification.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks.

World Economic Forum. 2022. News consumption habits in europe shifting to online and social media. Accessed: 2025-02-17.

G. Zhang and Z. Li. 2020. Multi-label text classification with xgboost and bert embeddings. In *Proceedings of the 6th International Conference on Machine Learning and Data Engineering (iCMLDE)*, pages xxx–yyy.