Team INSAntive at SemEval-2025 Task 10: Hierarchical Text Classification using BERT

Yutong Wang, Diana Nurbakova, Sylvie Calabretto

Team INSAntive at SemEval-2025 Task 10: Hierarchical Text Classification using BERT

Yutong Wang **Diana Nurbakova** **Sylvie Calabretto**
INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205,
69621 Villeurbanne, France
{fistname.lastname}@insa-lyon.fr

Abstract

We propose a BERT-based hierarchical text classification framework to address the challenges of training multi-level multi-class text classification task. As part of the SemEval-2025 Task 10 challenge (Subtask 2), the framework performs fine-grained text classification by training dedicated sub-category classifiers for each top-level category. Experimental results demonstrate the feasibility of the proposed approach for such a task.

1 Introduction

With the rapid development of Natural Language Processing (NLP), extracting narratives from online news has attracted widespread attention in both academia and industry (Piskorski and Yangarber, 2013). A deep understanding of how entities are presented in news articles and the identification of underlying narratives are crucial for media analysis, misinformation detection, and socio-political research (Vosoughi et al., 2018).

This study focuses on the cross-lingual, multi-label, and multi-category document classification task, which involves automatically identifying and assigning narrative labels and sub-narrative labels to news articles based on a two-level narrative labeling system within a specific domain. Specifically, each article may contain one or more narrative labels, with each narrative further subdivided into sub-narrative labels. The main objective of this study is to accurately assign all applicable narrative labels and their corresponding sub-narrative labels to each article. The task covers five languages (Bulgarian, English, Hindi, Portuguese, and Russian) and aims to evaluate narrative classification performance under cross-lingual conditions.

SemEval-2025 Task 10 (Piskorski et al., 2025; Stefanovitch et al., 2025) focused on multilingual characterization and narrative extraction from online news, divided into three independent subtasks. We participated in the Subtask 2 on the narrative classification. In this task, our model achieved a mid-to-upper range ranking on the final leaderboard, outperforming the baseline systems while still leaving room for further improvement.

Our approach combines large-scale pre-trained language models, a hierarchical classification strategy, and entity framing analysis to automatically identify and classify narratives across different languages and topics, providing a solution for narrative extraction from multilingual news texts.

In this study, we first perform translation-based data augmentation on the raw text data to ensure label consistency and accuracy across multiple languages. We then fine-tune a Transformer-based language model on the augmented dataset. To identify sub-narrative labels, we train separate classification models for each sub-narrative label and combine these models into an ensemble model. After completing the narrative label classification, we calculate the probability distribution of each sub-narrative label under its corresponding main narrative label and select the sub-narrative label with the highest probability as the classification result.

2 Background

Narrative classification is of great importance for extracting and identifying narrative structures from various types of texts, and it plays a crucial role in fields such as computational linguistics, NLP, and information retrieval (IR). In recent years, researchers have shifted their focus from traditional personal narratives (Langellier, 1989) to more diverse text types, especially informational texts (e.g., news reports, meeting minutes, and case analyses), and have made significant progress in developing computational methods for identifying narratives.

Classical narrative theory was initially proposed

by Labov and Waletzky (Labov and Waletzky). Building on this work, Swanson et al. (Swanson et al., 2014) manually annotated texts related to personal stories, categorizing clauses into three narrative types (orientation, evaluation, and action) and developed corresponding feature-based models. This endeavor laid an important foundation for subsequent research in narrative classification.

As the demand for automatic recognition of narrative structures continues to grow, there has been increasing interest in integrating narrative theory with machine learning models, aiming at achieving more efficient and accurate narrative classification across a variety of text types. Saldias and Roy (Saldias and Roy, 2020) employed convolutional neural networks (CNNs) to classify sentences in personal-story texts, automatically labeling each sentence according to the three narrative types proposed by Swanson et al. Meanwhile, Levi et al. (Levi et al., 2022) introduced NEAT (Narrative Elements AnnoTation), which uses multiple supervised learning models to distinguish highly interrelated narrative categories. Hatavara et al. (Hatavara et al., 2024) further developed a rule-based and computational approach to systematically extract narratives from parliamentary records and oral history interviews, demonstrating its feasibility on large datasets.

Recent surveys have provided comprehensive overviews of current methods and challenges in narrative extraction (Santana et al., 2023; Norambuena et al., 2023). Meanwhile, large-scale pre-trained language models, especially BERT (Devlin et al., 2019) with their pre-training on bidirectional language models (Rogers et al., 2020), are capable of better understanding contextual information within sentences and have thus been widely applied to narrative classification tasks (Gao et al., 2019; Hu et al., 2022; Purificato and Navigli, 2023). However, challenges still persist due to data scarcity and the inherent complexity of narrative structures, motivating researchers to explore the use of large language models (LLMs) for data augmentation (Conneau et al., 2020). Although both GPT (OpenAI et al., 2023) and BERT are large-scale pre-trained language models, GPT, as a generative model, has the ability to produce coherent and contextually relevant text. Thus, it offers novel solutions to address data insufficiency, particularly in the generation of high-quality narrative texts (Bartalesi et al., 2024) and translating narrative records (Hendy et al., 2023).

3 System overview

3.1 Framework Overview

Due to the limited volume of available training data and the large number of label categories, directly training deep learning models often fails to produce satisfactory results. To address this issue, this study proposes a multistage text translation and classification framework designed to efficiently handle multilingual texts and convert them into a standardized format suitable for deep learning model training. It comprises the following key steps:

(1) *Data Augmentation via Text Translation*: First, all articles written in languages other than the target language are translated into the required training language to ensure data consistency. For articles that exceed a certain length, a segmented translation strategy is employed to overcome API limitations while preserving textual integrity and readability.

(2) *Narrative Classification*: After translation, a global classification step is performed to assign articles to specific themes or categories based on their overall content. This step utilizes a pre-trained Transformer model (e.g., BERT) combined with supervised learning to optimize the classification process.

(3) *Sub-narrative Classification*: Next, texts under the same narrative label are further classified into subcategories to capture fine-grained semantic information. A hierarchical classification method is adopted to allow the model to recognize various narrative structures, thereby improving classification accuracy.

This framework offers several advantages: (1) Automation: It enables an end-to-end automated process from multilingual translation to classification, minimizing manual intervention and improving data-processing efficiency. (2) Adaptability: It supports multilingual inputs and can be adapted for various text classification tasks. (3) Computational Resource Optimization: The framework improves computational efficiency through segmented translation, dynamic model loading, parallel processing, and other optimization strategies.

3.2 Text Translation Framework

This study employs an automated text translation framework designed to batch-process and translate lengthy articles, ensuring the textual content

is comprehensively and efficiently converted into the target language. Built on OpenAI's GPT-4o, the system adopts a segmented translation strategy to address the challenges posed by extremely long texts. The framework consists of the following key steps: (1) *Segmentation and Preprocessing*: Long documents are divided into smaller segments to circumvent API length limitations. This segmentation method retains readability and allows for efficient parallel processing. (2) *Machine Translation Integration*: Each segment is translated into the target language using a LLM. This step ensures linguistic uniformity, which is crucial for downstream tasks such as classification and semantic analysis. (3) *Data Standardization*: The translated text is converted into a standardized format suitable for subsequent model training, facilitating organized data storage and retrieval.

Given a collection of source language text files, some of which may exceed the maximum output length supported by the OpenAI API (4,096 tokens), we propose a segmentation-based translation approach to address this limitation. This approach involves three main steps for each original text: (1) **Text Segmentation**: The input text is divided into smaller segments to ensure each falls within the API's token limit. (2) **API-based Translation**: Each segment is translated individually using the OpenAI translation API. (3) **Translation Merging**: All translated segments are then concatenated to reconstruct the complete translated text.

For texts exceeding the length limit, this process ensures translation segment by segment and reassemblage into a coherent full translation. By integrating segmentation, translation, and standardized output, this framework produces high-quality multilingual data for further classification and semantic analysis. Its modular design also enables flexible adaptation to different languages, domains, and model architectures, thus enhancing scalability and robustness in multilingual NLP pipelines.

3.3 Multi-label Text Classification

After augmenting the training articles via text translation, this study employs a BERT-based multi-label text classification framework. Specifically, for text data containing multiple narrative labels, a pre-trained Transformer model (`BERT-base-uncased` (Devlin et al., 2019)) is used for text representation learning. We then train on the augmented text corpus. The objective of this task is to perform multi-label classification on the input text, where each article can belong to multiple categories.

Given a dataset $D = \{(x_i, Y_i)\}_{i=1}^{N}$, where x_i represents the text and $Y_i \subseteq C$ denotes the set of labels assigned to that text (C is the set of all possible classes). The goal is to train a model F such that, for an input text x_i, it predicts the most appropriate set of class labels: $\hat{Y}_i = F(x_i) = \{\hat{y}_i^1, \hat{y}_i^2, \ldots, \hat{y}_i^m\}, \quad \hat{y}_i^j \in \{0, 1\},$, where $\hat{y}_i^j$ indicates the probability that x_i belongs to class c_j. The model architecture is given by: $F(x) = \text{Sigmoid}(\text{BERT}(x)),$, where: **BERT** acts as the backbone network, outputting logits that serve as class prediction scores, **Sigmoid** converts the logits into class probabilities: $p(y_i) = \frac{1}{1+e^{-z_i}},$, where z_i is the prediction score for class i. We optimize the Binary Cross-Entropy loss: $L = -\sum_{i=1}^{m} [y_i \log p(y_i) + (1 - y_i) \log(1 - p(y_i))]$,, where m denotes the total number of classes, y_i is the ground truth label for class i, and $p(y_i)$ is the model-predicted probability for that class.

3.4 Sub-narrative Label Classification

For sub-narrative labels, we use a BERT-based hierarchical text classification framework designed to perform multi-level classification. In this task, each text is first categorized into one or more top-level labels, and then further subdivided into sub-labels associated with each top-level category. To improve classification accuracy and generalization, a dedicated sub-label classifier is trained separately for each top-level label.

For each top-level category c_j, we train a dedicated subcategory classifier. The number of neurons in its output layer is equal to the number of subcategories under c_j. Formally, we denote this classifier as: $M_j = \text{BERT}_\theta + \text{FC}(h, |C_j^{\text{sub}}|),$, where BERT_θ represents the BERT model parameterized by θ, $\text{FC}(h, |C_j^{\text{sub}}|)$ is a fully connected layer that takes the hidden representation h as input and outputs a vector of length $|C_j^{\text{sub}}|$.

Given a text dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i represents the text and y_i represents the corresponding class label, let the set of top-level categories be denoted by: $C_{\text{top}} = \{c_1, c_2, \ldots, c_m\}$, and let the set of subcategories associated with a specific top-level category c_j be denoted by: $C_j^{\text{sub}} = \{s_j^1, s_j^2, \ldots, s_j^n\}$. The goal is to learn a function F such that, given a known top-level category c_j, it can predict the subcategory label for a text x_i as follows: $\hat{s}_i = F(x_i \mid c_j)$.

4 Experimental Setup

Dataset Description. The SemEval-2025 Task 10 Subtask 2 dataset consists of news articles in five different languages: English (EN), Portuguese (PO), Russian (RU), Bulgarian (BU), and Hindi (HI). Each article is annotated with one or more narrative labels from a predefined set of 21 top-level narratives, and each narrative is further associated with one or more sub-narratives from a total of 91 possible sub-narratives.

For our experiments, we used the multilingual corpus with varying document distributions across languages (see Figures 1-2). Prior to data augmentation, the training set contained 399 EN articles, 400 PO articles, 133 RU articles, 401 BU articles, and 366 HI articles. The development set consisted of 41 EN articles, 35 PO articles, 32 RU articles, 35 BU articles, and 35 HI articles. Our test set included 101 EN articles, 100 PO articles, 60 RU articles, 100 BU articles, and 99 HI articles.

The dataset exhibits significant class imbalance at both narrative and sub-narrative levels, as shown in Figures1 and 2. Some narratives like "*Criticism of Institutions and Authorities*" and "*Discrediting Ukraine*" appear frequently, while others like "*Climate Change is Beneficial*" and "*Controversy about green technologies*" are rarely represented. This imbalance presents a challenge for classification models, particularly in identifying and correctly classifying minority classes.

To address the imbalance in training data across languages, particularly the limited number of RU articles, we implemented a translation-based data augmentation strategy. After augmentation, each language in the training set contained 1,699 articles, creating a balanced training corpus across all five languages. This augmentation approach enabled our model to learn more robust cross-lingual patterns and improved overall performance, especially for languages with fewer original training samples.

Implementation Details. We implemented our model using PyTorch (1.10.0) and the Hugging Face Transformers library (4.16.2). For the text translation framework, we employed OpenAI's GPT-4o via the official API. Documents were segmented into chunks of approximately 1,000 tokens to stay within API limits while maintaining context coherence. For our augmentation process, we translated non-English articles to English and vice versa to ensure balanced representation across languages.

We set the following hyperparameter values for BERT-based multi-label classification model: (a) pre-trained model: `bert-base-uncased`, (b) maximum sequence length: 128 tokens, (c) batch size: 8 (narrative classifier) / 3 (sub-narrative classifiers), (d) learning rate: 2e-5 with AdamW optimizer, (e) weight decay: 0.01, (f) training epochs: 30 (early stopping with patience of 5) for narrative classifier / 3 for sub-narrative classifiers, (g) dropout rate: 0.1, (h) classification threshold: 0.2 (optimized on validation set) for narrative classifier.

To validate the effectiveness of our data augmentation approach, we conducted experiments both with and without the augmented data. The results demonstrate significant performance improvements when using the augmented dataset, particularly for languages with fewer original training samples like Russian (see Table 3).

The training process for the narrative and sub-narrative classifiers was structured as follows. We first trained the top-level narrative classifier on the full augmented dataset of 1,699 articles per language. For each narrative category, we filtered the dataset to include only articles with that narrative label. We then trained a dedicated sub-narrative classifier for each narrative category using a single-label classification approach with LabelEncoder. During inference, we first predicted the narrative labels using the top-level classifier, then employed the corresponding sub-narrative classifiers to predict the fine-grained labels.

For the sub-narrative classification, we organized articles by their top-level narrative labels and trained separate BERT-based models for each top-level category. Unlike the multi-label approach used for narrative classification, each sub-narrative classifier was trained as a standard single-label classification model using cross-entropy loss. This approach was chosen due to the hierarchical nature of the labels and to reduce complexity in the sub-narrative prediction task.

All sub-narrative models were saved with their corresponding tokenizer and label encoder to enable efficient inference. During prediction, once the top-level narrative was identified, the corresponding sub-narrative model was loaded to predict the specific sub-category. This hierarchical approach allowed us to effectively handle the large number of potential sub-narratives while maintaining computational efficiency.

To address the class imbalance issue in the top-level narrative classification, we implemented class weighting in the loss function, where weights were

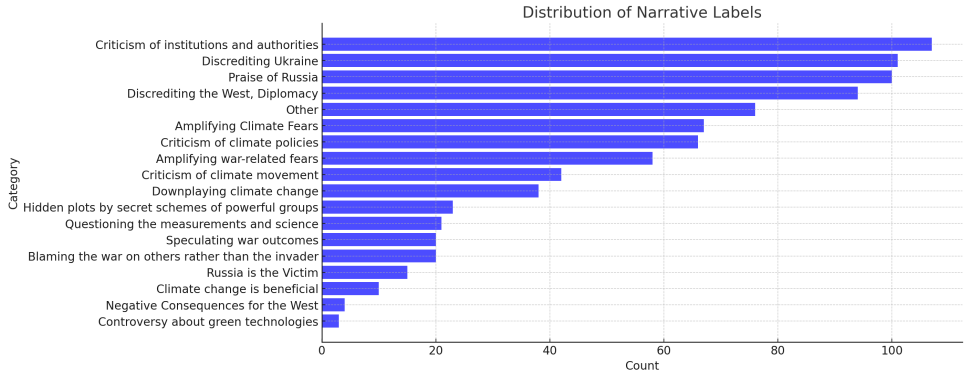

Figure 1: Distribution Narrative Labels

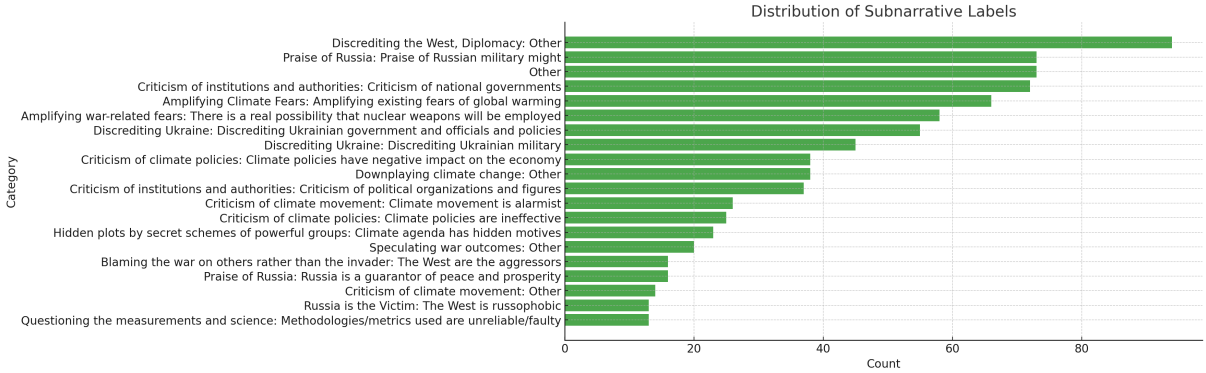

Figure 2: Distribution Subnarrative Labels

inversely proportional to class frequencies in the training set. This approach improved the model's ability to identify minority classes without significantly degrading performance on majority classes.

Evaluation Metrics. Following the official SemEval-2025 Task 10 evaluation criteria, we measured our model's performance using the following metrics: (a) **F1 Macro**: The unweighted mean of F1 scores for each class, giving equal importance to all classes regardless of their frequency. This metric is particularly important for evaluating performance on imbalanced datasets, as it prevents the model from being overly biased toward majority classes. (b) **F1 Samples**: The F1 score calculated for each instance and then averaged, which accounts for the multi-label nature of the task. This metric provides insights into the model's ability to correctly predict all relevant labels for each document.

Additionally, we report the standard deviation (St.Dev) for both metrics to analyze the stability of our model's performance across different classes and samples. A lower standard deviation indicates more consistent performance across all categories, which is desirable for robust classification systems.

5 Results and Analysis

In this study, we performed multilevel classification on text data in different languages and computed the F1 Macro Coarse and F1 Samples metrics to evaluate the model's classification performance across different language datasets. The experimental results are presented in Table 3, where F1 Macro Coarse measures the overall balance of classification performance between categories, and F1 Samples focuses on the performance of the model in individual samples.

Table 1: F1 Scores on Test set

Lang	F1 Macro	F1 St.Dev	F1 Sample	F1 St.Dev Smp
EN	0.443	0.380	0.281	0.352
PO	0.491	0.275	0.245	0.204
RU	**0.554**	**0.328**	**0.323**	**0.342**
BU	0.523	0.366	0.324	0.360
HI	0.365	0.440	0.365	0.414

To further validate the effectiveness of the proposed model, we conducted a systematic comparison with baseline methods on different language-specific datasets. As presented in Table 3, our

model consistently outperforms the baseline across all evaluated languages. The F1 Macro and F1 Sample scores demonstrate substantial improvements, reflecting both better overall classification balance and stronger sample-level performance.

Table 2: F1 Scores Comparison with Baseline Models

	Baseline		**Proposed Model**	
	F1 Macro	F1 Sample	F1 Macro	F1 Sample
EN	0.030	0.013	**0.443**	**0.281**
PO	0.037	0.014	**0.491**	**0.245**
RU	0.065	0.008	**0.554**	**0.323**
BU	0.040	0.039	**0.523**	**0.324**
HI	0.081	0.000	**0.365**	**0.365**

The comparative analysis reveals remarkable improvements over the baseline models. For English (EN), our model achieves an F1 Macro score of 0.443 compared to the baseline's 0.030, representing a 14.8× improvement. Similarly, Portuguese (PO) shows a 13.3× improvement, Russian (RU) an 8.5× improvement, Bulgarian (BU) a 13.1× improvement, and Hindi (HI) a 4.5× improvement in F1 Macro scores. The most dramatic improvement is observed in the F1 Sample metric for Hindi, where our model achieves 0.365 compared to the baseline's 0.000, indicating the baseline completely failed to correctly classify individual samples in this challenging language.

Table 3: F1 Macro Coarse Comparison

Lang	No Augmentation	With Augmentation
EN	0.329	**0.443**
PO	0.220	**0.491**
RU	0.224	**0.554**
BU	0.188	**0.523**
HI	0.301	**0.365**

Our model performs best on the Russian (RU) dataset, reaching an F1 Macro Coarse of 0.554 and an F1 Samples of 0.323, suggesting relatively high accuracy both at the global category level and for individual samples. In contrast, when applied to the Hindi (HI) dataset, it exhibits the lowest classification performance, with an F1 Macro Coarse of only 0.365, implying that this language poses a greater classification challenge—potentially due to data quality or linguistic factors. Meanwhile, Portuguese (PO) and Bulgarian (BU) show comparable results at 0.491 and 0.523, respectively, indicating relatively stable model generalization for these languages. Regarding the standard deviation (St. Dev.), Hindi's F1 St.Dev. Coarse is as high as 0.440, with an F1 St.Dev. Samples of 0.414, suggesting large variability in its classification performance—likely stemming from data imbalance or label inconsistencies. In contrast, Portuguese has an F1 St.Dev. Coarse of only 0.275, implying more stable classification outcomes, making it suitable for more fine-grained text classification tasks. The distribution of classification results is not uniform: a few high-frequency categories (e.g., "*Criticism of Institutions and Authorities*", "*Slandering Ukraine*," "*Praise for Russia*") occupy a relatively large portion of the corpus, whereas other categories (e.g., "*Questioning Scientific Measurements and Indicators*", "*Climate Change is Beneficial*") have significantly fewer samples. This imbalance not only reflects real-world differences in the frequency with which various narratives appear but also potentially affects the model's discriminatory power: when high-frequency categories dominate the dataset, the model tends to learn their features more effectively, while its ability to recognize low-frequency categories weakens accordingly.

6 Conclusion

In the multi-label setting, the framework integrates a BERT-based text classification method, using automated data processing, optimized training workflows, and memory management strategies. Our proposed framework additionally provides a range of functional modules (segmentation, automated translation, and standardized output) that facilitate the generation of high-quality multilingual data for subsequent classification and semantic analysis. Experimental results show that our method performs well in handling large-scale, multilingual text data and achieves high accuracy in hierarchical classification tasks. Future research directions include: (1) further optimizing parallel processing strategies to improve overall training efficiency; (2) enhancing the accuracy of sub-category classification; and (3) exploring more powerful multilingual pre-trained models to strengthen system robustness and generalization capabilities.

Acknowledgments

Yutong Wang is supported by the China Scholarship Council scholarship for Ph.D. program at INSA Lyon, France. File No. 202308120039.

References

David Bamman, Brendan O'Connor, and Noah Smith. 2012. Censorship and deletion practices in chinese social media.

Valentina Bartalesi, Emanuele Lenzi, and Claudio De Martino. 2024. Using large language models to create narrative events. 10:e2242.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. 2019. Target-dependent sentiment classification with BERT. 7:154290–154299.

Mari Hatavara, Kirsi Sandberg, Mykola Andrushchenko, Sari Hälikkö, Jyrki Nummenmaa, Timo Nummenmaa, Jaakko Peltonen, and Matti Hyvärinen. 2024. Computational recognition of narratives: Applying narratological definitions to the analysis of political language use.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation.

Yongjun Hu, Jia Ding, Zixin Dou, and Huiyou Chang. 2022. Short-text classification detector: A bert-based mental approach. 2022:1–11.

William Labov and Joshua Waletzky. Narrative analysis. oral versions of personal experience. In *Essays on the Verbal and Visual Arts: Proceedings of the American Ethnological Society*, volume 12-44. Seattle: American Ethnological Society.

Kristin M. Langellier. 1989. Personal narratives: Perspectives on theory and research. *Text and Performance Quarterly*, 9:243–276.

Effi Levi, Guy Mor, Tamir Sheafer, and Shaul Shenhav. 2022. Detecting narrative elements in informational text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1755–1765. Association for Computational Linguistics.

Sebastian Michelmann, Manoj Kumar, Kenneth A. Norman, and Mariya Toneva. 2023. Large language models can segment narrative events similarly to humans.

Iraklis Moutidis and Hywel T. P. Williams. 2020. Complex networks for event detection in heterogeneous high volume news streams.

Brian Norambuena, Tanushree Mitra, and Chris North. 2023. A survey on event-based news narrative extraction. *ACM Computing Surveys*, 55(14s).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak,

Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 technical report.

Jakub Piskorski, Tarek Mahmoud, Nikolaos Nikolaidis, Ricardo Campos, Alípio Jorge, Dimitar Dimitrov, Purificação Silvano, Roman Yangarber, Shivam Sharma, Tanmoy Chakraborty, Nuno Ricardo Guimarães, Elisa Sartori, Nicolas Stefanovitch, Zhuohan Xie, Preslav Nakov, and Giovanni Da San Martino. 2025. SemEval-2025 task 10: Multilingual characterization and extraction of narratives from online news. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

Jakub Piskorski and Roman Yangarber. 2013. Information extraction: Past, present and future. In Thierry Poibeau, Horacio Saggion, Jakub Piskorski, and Roman Yangarber, editors, *Multi-source, Multilingual Information Extraction and Summarization*, pages 23–49. Springer Berlin Heidelberg.

Antonio Purificato and Roberto Navigli. 2023. APatt at SemEval-2023 Task 3: The Sapienza NLP System for Ensemble-based Multilingual Propaganda Detection. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 382–388, Toronto, Canada. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. 8:842–866.

Belen Saldias and Deb Roy. 2020. Exploring aspects of similarity between spoken personal narratives by disentangling them into narrative clause types. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 78–86, Online. Association for Computational Linguistics.

Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. 2023. A survey on narrative extraction from textual data. *Artificial Intelligence Review*, 56(8):8393–8435.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2023. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data.

Nicolas Stefanovitch, Tarek Mahmoud, Nikolaos Nikolaidis, Jorge Alípio, Ricardo Campos, Dimitar Dimitrov, Purificação Silvano, Shivam Sharma, Roman Yangarber, Nuno Guimarães, Elisa Sartori, Ana Filipa Pacheco, Cecília Ortiz, Cláudia Couto, Glória Reis de Oliveira, Ari Gonçalves, Ivan Koychev, Ivo Moravski, Nicolo Faggiani, Sopho Kharazi, Bonka Kotseva, Ion Androutsopoulos, John Pavlopoulos, Gayatri Oke, Kanupriya Pathak, Dhairya Suman, Sohini Mazumdar, Tanmoy Chakraborty, Zhuohan Xie, Denis Kvachev, Irina Gatsuk, Ksenia Semenova, Matilda Villanen, Aamos Waher, Daria Lyakhnovich, Giovanni Da San Martino, Preslav Nakov, and Jakub Piskorski. 2025. Multilingual Characterization and Extraction of Narratives from Online News: Annotation Guidelines. Technical Report JRC141322, European Commission Joint Research Centre, Ispra (Italy).

Reid Swanson, Elahe Rahimtoroghi, Thomas Corcoran, and Marilyn Walker. 2014. Identifying narrative clause types in personal stories. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 171–180. Association for Computational Linguistics.

Adane Nega Tarekegn. 2024. Large language model enhanced clustering for news event detection.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Shanshan Yu, Jindian Su, and Da Luo. 2019. Improving BERT-based text classification with auxiliary sentence and domain knowledge. 7:176600–176612.