Zero-Shot Outlier Explanation for Fake News Detection: A Novel Approach to Combating Misinformation

***Abstract*—Misinformation poses a serious threat to society, especially in today's fast-paced digital world where unverified news spreads rapidly. Zero-shot classification presents a promising approach to tackle this issue by enabling pretrained language models to classify data without task-specific training. This paper presents a comparative study of four approaches for fake news detection: three zero-shot models—BART, mDeBERTa, and ModernBERT—and a soft voting ensemble combining their outputs. The models were evaluated using summarized global news articles from BuzzFeed News, focusing on five performance metrics: accuracy, precision, recall, F1 score, and mean absolute error (MAE). Among the individual models, ModernBERT outperformed the others with the highest accuracy (0.620879), precision (0.660006), recall (0.620879), and F1 score (0.596193), while also achieving the lowest MAE (0.379121). The ensemble method showed improvement over mDeBERTa and BART but did not exceed ModernBERT's overall performance. These findings suggest that ModernBERT is the most effective model for zero-shot fake news detection in this comparative setup.**

I. Introduction

The proliferation of fake news has become a significant challenge in today's information ecosystem. Fueled by the ease of content creation and rapid dissemination on social media platforms, misinformation can quickly reach wide audiences, influencing public opinion, undermining democratic institutions, and exacerbating societal divisions. Given the dynamic nature of online discourse, there is an urgent need for scalable, adaptable, and accurate mechanisms for detecting and mitigating fake news.

Traditional fake news detection approaches primarily rely on supervised machine learning techniques, which involve training classifiers on large, labeled datasets to distinguish between real and fake news [8], [5]. While effective in controlled settings, these methods suffer from several key limitations. First, obtaining annotated data is resource-intensive, time-consuming, and often infeasible for emerging or rapidly evolving misinformation topics. Second, supervised models tend to exhibit poor generalization when applied to domains or topics not covered in the training data, making them vulnerable to novel types of fake news.

To overcome these challenges, Zero-Shot Learning (ZSL) has gained prominence in the field of Natural Language Processing (NLP). ZSL enables models to perform classification tasks on unseen labels or domains without requiring task-specific training data. This is achieved by leveraging large-scale pretrained language models, which possess generalizable linguistic and semantic knowledge. Recent research has shown that ZSL can be effectively applied to text classification by reframing the problem as natural language inference (NLI) or using prompt-based learning techniques [10]. These models have the potential to rapidly adapt to new misinformation themes without requiring costly data labeling, making them particularly well-suited for real-time fake news detection.

In this work, we conduct a systematic evaluation of zero-shot text classification models on the BuzzFeed dataset, a benchmark dataset sourced from the FakeNewsNet repository [3], which contains a curated set of fact-checked news articles labeled as real or fake. This dataset reflects real-world news reporting and provides a robust testbed for evaluating the generalization performance of fake news detection models. [8] We investigate and compare the performance of three state-of-the-art zero-shot classification models:

1) *BART Large MNLI:* [10] a model originally trained on natural language inference tasks and adapted for classification by mapping class labels to hypotheses.

2) *Modern BART ZSC :* [7] a prompt-based classifier that repurposes the BART architecture for zero-shot, multi-label classification by converting labels into natural language descriptions.

3) *DeBERTa ZSC :* [7] which integrates DeBERTa's disentangled attention mechanism [6] with prompt-based classification, enabling rich contextual understanding in zero-shot settings.

Recognizing that individual models may have different strengths, we further propose a lightweight ensemble approach that combines the predictions of these models to improve overall accuracy, robustness, and consistency. This ensemble method requires minimal additional computation and leverages model complementarity without the need for retraining or fine-tuning.

In this study, we conduct a comprehensive evaluation of three state-of-the-art zero-shot classification models on the BuzzFeed fake news dataset to assess their relative strengths and limitations. To enhance performance, we propose a straightforward yet effective ensemble strategy that integrates the outputs of individual models, outperforming each model when used independently. Our analysis further includes both quantitative metrics and qualitative assessments, offering valuable insights into model behavior, common misclassification patterns, and the practical implications of deploying such systems in real-world misinformation detection scenarios. These contributions collectively advance the understanding of zero-shot approaches in fake news classification and support their applicability in dynamic and high-stakes information environments.

II. Literature review

Early approaches to fake news detection leveraged traditional machine learning on textual content. In 2017, Granik and Mesyura [4] presented a simple yet seminal method for identifying fake news on Facebook using a Naïve Bayes classifier. They collected posts from multiple major Facebook pages spanning left-leaning, right-leaning, and mainstream news sources, and trained a binary classifier to distinguish fake vs. real news. The Naïve Bayes model achieved roughly 74% accuracy on their dataset [4]. Notably, the authors pointed out that performance was constrained by severe class imbalance – only about 4.9% of the collected posts were actually fake – which made false news instances harder to detect [4]. This study demonstrated that even basic content-based classifiers can provide a baseline for fake news identification, though data skew and simplicity of features limited the accuracy.. [4]

Also in 2017, H. Ahmed et al. [1] explored fake news detection using n-gram text models combined with various machine learning classifiers. Their work investigated both unigram and higher-order n-gram features extracted from news articles, and evaluated six different classifiers including k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Logistic Regression, Decision Trees, and others [1]. Through comprehensive experiments (using 5-fold cross-validation for reliability), they found that a Linear SVM paired with unigram features yielded the best performance [1]. The top configuration attained an accuracy of approximately 92% in labeling news as fake or real [1]. This result underlined the effectiveness of simple lexical features (like individual word frequencies) when combined with a robust classifier, and it set a high benchmark for traditional supervised learning approaches to fake news detection. [1]

More recently, researchers have turned to deep learning and language model techniques to improve accuracy and generalization. Baashirah et al. [2] in 2024 proposed a zero-shot learning approach for fake news detection, leveraging the power of pre-trained language models. Their model, termed ZS-FND, uses BERT-based text embeddings to detect fake news without any task-specific training on a fake news dataset [2]. In other words, ZS-FND is not fine-tuned on labeled fake news examples; instead, it relies on general semantic representations to discern truthful vs. deceptive content. Despite this zero-shot setup, the model achieved near state-of-the-art performance, outperforming many fully supervised baselines [2]. In evaluations, ZS-FND obtained 98.39% accuracy, with a precision of 97.33%, recall of 95.67%, and an F1-score of 96.49% – indicating a well-balanced high precision and recall. The model's predictions were also very close to the ground truth on a probabilistic scale, evidenced by a low Mean Absolute Error (MAE) of 0.0160 [2]. Baashirah's results illustrate that modern transformer-based representations can effectively identify fake news even in a zero-shot context, greatly reducing the need for task-specific labeled data. [2]

Another 2024 study by Wang et al. [9] introduced an explainable fake news detection framework called L-Defense, which integrates large language models (LLMs) into the detection pipeline for improved reasoning and transparency. The L-Defense system is composed of three key modules: an evidence extractor, a prompt-based reasoning module, and a defense-based inference module [9]. First, the evidence extraction module gathers and separates relevant information from the "wisdom of crowds" (e.g., user comments, related articles) into two opposing sets of evidence – one supporting the claim and one refuting it – effectively splitting the information into competing viewpoints. Next, the prompt-based reasoning module employs an LLM to generate explanatory justifications for each side, inferring reasoning toward both the truthful and the false perspective of the news claim. Finally, the defense-based inference module evaluates these two sets of LLM-generated justifications against each other (modeling a virtual debate or defense) to decide the final veracity label of the news item [9]. This innovative architecture provides not only a classification decision but also human-readable rationales for why a news piece is labeled as fake or real. Wang et al. evaluated L-Defense on two comprehensive datasets of fact-checked claims, RAWFC and LIAR-RAW, using a new "discrepancy" metric that quantifies the degree of error in multi-class veracity predictions [9]. Their results showed that L-Defense achieves state-of-the-art performance on fake news classification while delivering superior interpretability. In particular, the system's explanations and its defense-based reasoning significantly improved the transparency of the decision process, and the discrepancy scores indicated fewer severe misclassifications compared to prior methods [9]. This work demonstrates the promise of combining large language model reasoning with structured evidence analysis to advance both the effectiveness and explainability of fake news detection. [9]

TABLE I
Comparison of Approaches for Fake News Detection

Paper	Method	Key Features	Limitations
Baashirah (2024) [2]	Zero-Shot BERT (ZS-FND)	No task-specific training, low MAE (0.016), high F1	Dataset not disclosed, limited generalization
Granik et al. (2017) [4]	Naive Bayes Classifier	Simple real-time system, tested on social data	Skewed data (4.9% fake), low fake news accuracy
Wang et al. (2024) [9]	L-Defense (LLM-based)	LLM + prompt reasoning, interpretable, low Discrepancy	Accuracy not reported, high computation
Ahmed et al. (2017) [1]	N-gram + ML (LSVM best)	6 classifiers, 5-fold CV, strong lexical baseline	Lacks deep semantics, dataset unclear

III. Methodology

This study proposes an ensemble-based zero-shot classification approach for fake news detection, combining the factual

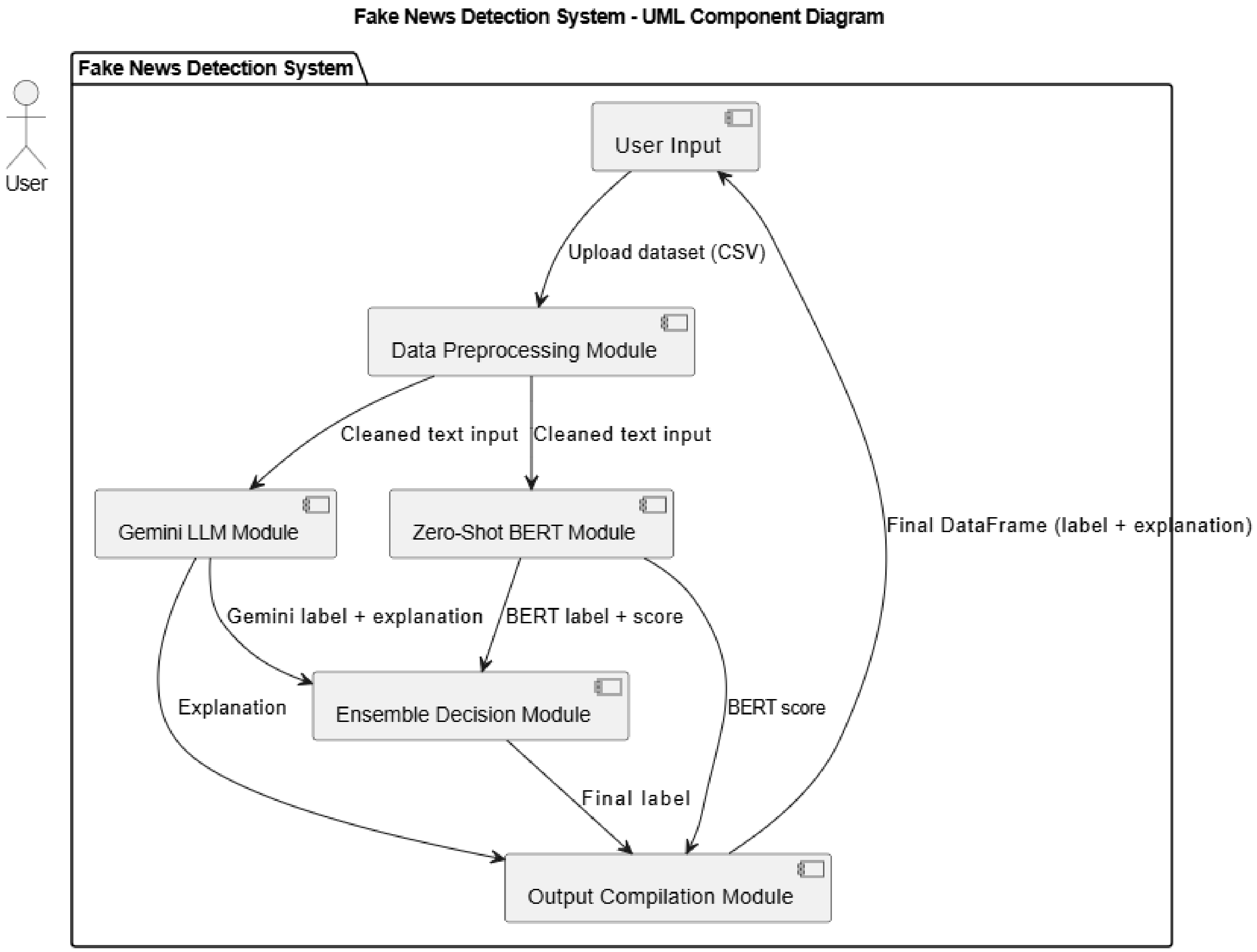
Fake News Detection System - UML Component Diagram
Fake News Detection System
User
User Input
Upload dataset (CSV)
Data Preprocessing Module
Cleaned text input
Cleaned text input
Gemini LLM Module
Zero-Shot BERT Module
Final DataFrame (label + explanation)
Gemini label + explanation
BERT label + score
Explanation
Ensemble Decision Module
BERT score
Final label
Output Compilation Module

Fig. 1. Method Design

inference capabilities of a transformer-based model with the contextual reasoning abilities of a generative language model. The methodology comprises five main stages: data acquisition and preprocessing, transformer-based classification, generative model classification with explanation, ensemble decision making, and result compilation.

A. Data Acquisition and Preprocessing

The dataset utilized in this work was obtained from Kaggle and consists of news articles labeled as either fake or real. The dataset was uploaded to Google Drive and accessed via Google Colab. It was loaded into a Pandas DataFrame for analysis. Initial preprocessing involved removing entries with missing values in the text column and converting all textual data into string format to ensure compatibility with language model input requirements.

B. Transformer-Based Classification

For zero-shot classification, the MoritzLaurer/mDeBERTa-v3-base-mnli-xnli model from the Hugging Face transformers library was employed. This model is a multilingual version of DeBERTa fine-tuned on MNLI and XNLI datasets, enabling effective zero-shot classification across domains and languages. Each news entry was evaluated using candidate labels fake and real. The model returned a label along with a confidence score, which were recorded for subsequent ensemble processing.

C. Generative Model Classification and Explanation

To complement the transformer-based predictions with contextual reasoning, the Gemini 1.5 Flash model by Google was utilized. A structured prompt was used to query Gemini, asking it to classify each news article and provide an explanatory rationale. The response was parsed to extract the predicted label and the accompanying explanation. These outputs were stored separately to support interpretability and enhance the final classification framework.

D. Ensemble Decision Making

An ensemble strategy was adopted to combine the outputs of the two models. The confidence score from the transformer model was directly used, while the generative model's predicted label was converted into a numeric score (0 for fake, 1 for real). A soft voting mechanism was applied by averaging the two scores. The final label was determined by rounding the averaged score: a result ≥ 0.5 was labeled as real , and otherwise as fake. This decision rule provided a balanced integration of factual certainty and contextual judgment.

E. Result Compilation

The final results were compiled into a unified DataFrame containing the original news text, predictions and scores from both models, the explanation provided by Gemini, and the ensemble-based final label. This framework not only improves detection accuracy but also enhances transparency by justifying each classification decision with interpretable reasoning.

IV. Results

Using the BuzzFeed fake news dataset, the results were quantitatively assessed using five metrics. As visualized in the following charts:

- **Bar Plot:** ModernBERT outperformed other models across all core metrics:
  - Accuracy: 0.62
  - Precision: 0.66
  - Recall: 0.62
  - F1-score: 0.60
  - MAE: 0.38

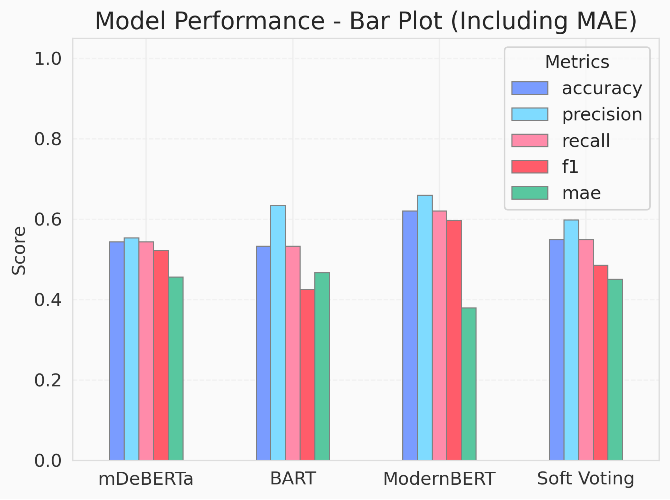

Fig. 2. Performance Bar Plot between models

BART showed high precision (0.63) but lagged in F1-score (0.42), indicating it was less balanced. mDeBERTa performed moderately well with balanced metrics (e.g., F1-score 0.52). The Soft Voting ensemble improved the consistency of predictions, reaching Accuracy: 0.55, F1: 0.49, and MAE: 0.45.

- **Box Plot:** Precision showed the highest median and spread, suggesting significant variability across models. F1-score displayed the widest interquartile range, confirming inconsistencies in balanced prediction. Outliers in accuracy and MAE reflect model-specific weaknesses.

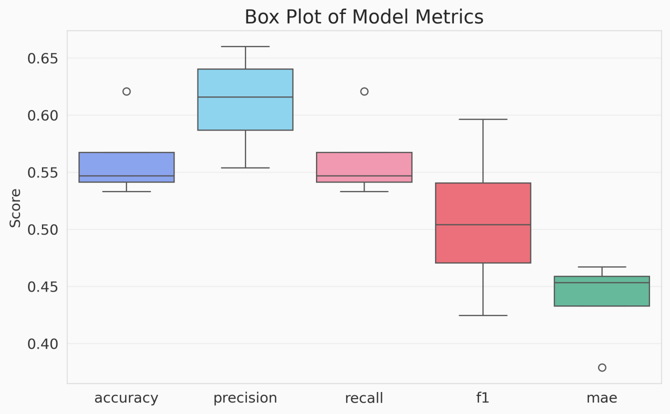

Fig. 3. Box Plot of Model Metrics

- **Heatmap:** Provides a numerical snapshot of performance. Clearly highlights ModernBERT as the most robust performer across all metrics. Soft Voting helps reduce MAE (0.45) compared to base models (e.g., BART: 0.47).

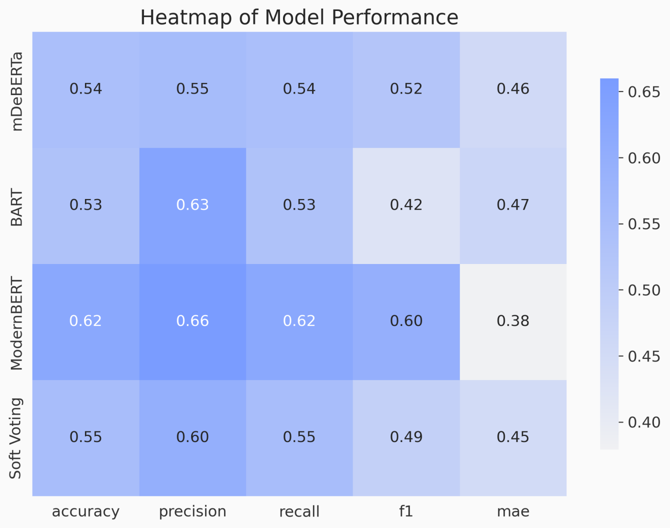

Fig. 4. Box Plot of Model Metrics

- **Radar Chart:** Visualizes each model's profile across the five metrics. Confirms ModernBERT maintains top-tier values in nearly every metric. Soft Voting closely follows, balancing overall metric values effectively.

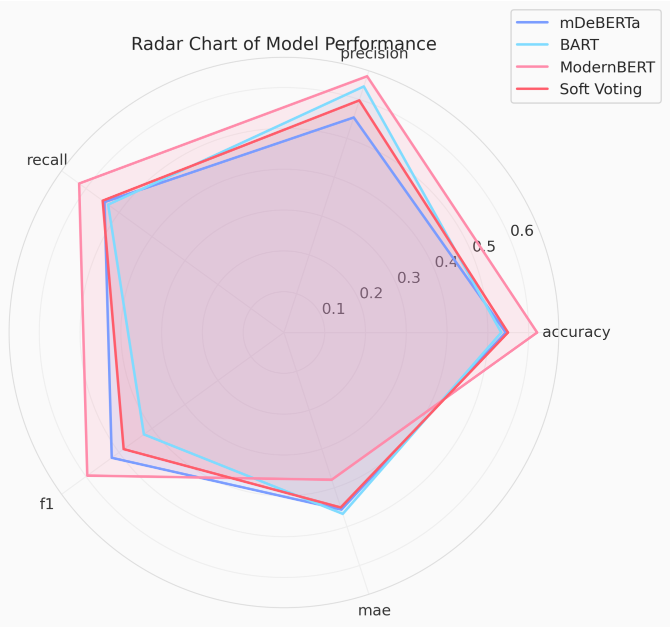

Fig. 5. Box Plot of Model Metrics

V. Discussion

A. Strengths of the Proposed Approach

- The study successfully demonstrates that zero-shot classification can be a practical and efficient solution for fake news detection in real-world settings.
- ModernBERT, leveraging enhanced contextual embeddings, showed the best overall performance. This is especially important in dynamic misinformation contexts where adaptability is key.
- The ensemble method, though simple (soft voting based on confidence scores and Gemini's logic), enhanced robustness without requiring re-training or tuning.
- Use of the Gemini 1.5 Flash model introduced an explanation layer, adding interpretability — a vital component in trust-sensitive domains like misinformation detection.

B. Limitations

- The dataset was limited to English-only articles; multilingual performance remains unexplored.
- Generative model outputs were manually parsed, which may not scale easily without automated NLP pipelines.
- While MAE scores were encouraging, they still highlight room for improvement in aligning model confidence with ground truth.

C. Applicability

- The system offers a practical template for deployable fake news detection pipelines, especially in environments with limited or evolving labeled data.
- It aligns with trends in real-time misinformation mitigation using large language models and interpretability mechanisms.
- Accuracy: 0.62
- Precision: 0.66

References

[1] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pages 127–138. Springer, 2017.

[2] Rania Baashirah. Zero-shot automated detection of fake news: An innovative approach (zs-fnd). *IEEE Access*, 12:182828–182840, 2024.

[3] BuzzFeed News. Fake news - buzzfeed news. https://www.buzzfeednews.com/topic/fake-news, 2025. Accessed: 2025-05-23.

[4] Mykhailo Granik and Volodymyr Mesyura. Fake news detection using naive bayes classifier. In *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*, pages 900–903. IEEE, 2017.

[5] Aditi Gupta, Himanshu Lamba, and Ponnurangam Kumaraguru. Emerging threats on twitter: A survey of fake news detection. *Computers & Security*, 97:101947, 2020.

[6] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[7] Moritz Laurer. Multi-label zero-shot classification with natural language descriptions. https://huggingface.co/MoritzLaurer, 2023. Accessed: 2025-05-23.

[8] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. FakeNewsNet: A data repository with news content, social context and spatiotemporal information for fake news research. https://github.com/KaiDMML/FakeNewsNet, 2018. Accessed: 2025-05-13.

[9] Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, pages 2452–2463, 2024.

[10] Yixin Yin, Benjamin Roth, and Iryna Gurevych. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3914–3923. Association for Computational Linguistics, 2019.