A Comparative Analysis of Machine Learning Methods for Sentiment Classification and Topic Modeling in Climate Change Discourse

April 27, 2025

1. Task 1 Methodology Evaluation

In this section, I briefly summarize the modifications made to the Naive Bayes classifier in Task 1.1d. I then compare the performance of three models using tables and figures, analyze their strengths and limitations, and discuss how the results reflect key concepts from the course. Finally, I present misclassified examples to illustrate model weaknesses and guide potential improvements.

1.1. Modifications to the Naive Bayes Classifier

In Task 1.1d, I enhanced the baseline Naive Bayes classifier to improve performance. First, I expanded the n-gram range from strict bigrams $(2, 2)$ to unigrams and bigrams $(1, 2)$, capturing both isolated words and short phrases. Second, I applied feature selection by setting `min_df=2` to remove rare tokens and `max_df=0.9` to exclude overly common terms. Standard English stopwords were also filtered out.

These modifications increased classification accuracy from 73% to 78%. The adjustments to `max_df`, `min_df`, and `ngram_range` significantly improved the Naive Bayes model's ability to balance feature selection and classification performance.[1]

Additionally, I evaluated TF-IDF transformation but found it reduced performance. As a probabilistic model, Naive Bayes works better with raw counts, since TF-IDF reweighting weakens high-frequency features that are important for classification in this task.

1.2. Comparison and analysis of method accuracy

To assess different modeling strategies for text classification, I compared three models: Naive Bayes, a Feed-Forward Neural Network (FFNN), and BERT-tiny. Table 1 reports the evaluation results, and Table 2 summarizes the optimal hyperparameters.

Model	Accuracy	Macro F1
Naive Bayes	0.780	0.780
Neural Network	0.480	0.478
BERT-tiny	0.775	0.778

Table 1: Performance comparison across models.

In terms of accuracy, both Naive Bayes and BERT-tiny outperformed the FFNN model. Considering both accuracy and macro F1 scores, Naive Bayes and BERT showed similar overall performance, clearly surpassing FFNN. Although BERT slightly underperformed Naive Bayes in accuracy (0.775 vs. 0.780), its F1 score was more balanced (0.778 vs. 0.780), indicating more even class distribution and stronger robustness to class imbalance.

Although Naive Bayes is a basic linear model, it demonstrated strong robustness in small-scale, sparse-feature scenarios when combined with n-gram features and proper text preprocessing (Lüdeling and Kytö, 2008; Řehůřek and Sojka, 2010). In contrast, the FFNN, despite its nonlinear modeling capacity, suffered from underfitting due to the lack of pretrained embeddings and failed to capture semantic nuances. The BERT-tiny model, leveraging transfer learning, introduced rich contextual knowledge, leading to steadily improving validation accuracy and more balanced classification across

Model	Optimal Hyperparameters
Naive Bayes	ngram=(1,2), min_df=2, max_df=0.8, stopwords='english'
FFNN	dropout=0.2, lr=0.0001, weight_decay=1e-6, clip=1.0, epoch=30
BERT-tiny	bert-tiny, max_len=64, lr=5e-5, drop=0.2, wd=0.01, bs=16, clip=1.0, epoch=15

Table 2: Optimal hyperparameters for each model.

[1] The full tuning process and accuracy visualization are available in the supplementary Jupyter notebooks `text_analytics_part1.ipynb` and `task2.1_Visualisation.ipynb`.

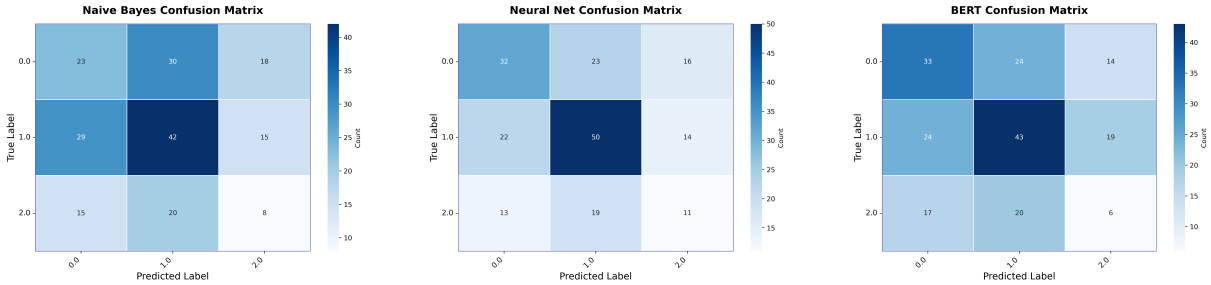

Figure 1: Confusion matrices of three models (from left to right): Naive Bayes, Feed-Forward Neural Network, and BERT-tiny.

classes.

1.3. Confusion Matrix Analysis of Class-Specific Errors

The confusion matrix analysis (Figure 1) reveals that all three models encounter varying levels of difficulty when predicting the *opportunity* class (label 2).

The Feed-Forward Neural Network (FFNN) shows the highest misclassification rate for this class, frequently labeling *opportunity* instances as *risk* (label 0). This pattern indicates the FFNN's limited ability to capture abstract semantic nuances.

Although Naive Bayes achieves reasonable overall performance, it often misclassifies *opportunity* samples as *neutral* (label 1). This tendency reflects the model's reliance on word frequency without contextual awareness (Bird et al., 2009; Jurafsky and Martin, 2020).

BERT-tiny outperforms the other models, displaying the most balanced prediction distribution across the three classes. Its pretrained Transformer architecture allows better modeling of contextual relationships. However, BERT-tiny still struggles with sentences that contain multiple sentiments or semantic reversals, leading to occasional misclassifications.

These findings highlight the varying capacities of different models in handling subtle and compound emotional expressions.

1.4. Semantic Analysis of Model Misclassifications

An analysis of misclassified samples highlights the strengths and weaknesses of different models in interpreting the three emotional categories: *risk*, *neutral*, and *opportunity*.

The Naive Bayes model, which relies on a bag-of-words approach, often misclassifies sentences with positive intentions but weak emotional cues as *neutral*. This reflects the model's reliance on word frequency and its inability to capture semantic-level sentiment.

In contrast, the Feed-Forward Neural Network (FFNN) shows more frequent misclassifications in the *opportunity* category, often mistaking it for *risk*. These patterns highlight the network's limitations in building semantic representations without pretrained embeddings.

The BERT model achieves the best overall performance, especially on texts requiring strong contextual understanding. However, it still faces challenges with compound sentiments, showing that despite advanced semantic modeling, BERT's sentiment interpretation remains influenced by syntactic structure and emotional signal placement (Mikolov et al., 2013; Hamilton et al., 2016).

1.5. Future Directions for Model Improvement

To address the identified shortcomings, several directions for model enhancement are proposed. For the Naive Bayes classifier, incorporating TF-IDF features and enriching input representations with domain-specific lexicons can improve its ability to capture implicit sentiment.

For the FFNN model, adopting pretrained word embeddings such as GloVe and exploring more expressive architectures like BiLSTM or CNN could enhance semantic modeling, especially in complex sentence structures.

Regarding the BERT model, replacing it with more advanced language models (e.g., RoBERTa) or integrating a Conditional Random Field (CRF) layer can potentially improve boundary recognition and emotion tagging accuracy. Additionally, designing post-processing rules informed by systematic error analysis—particularly for handling compound emotional expressions—offers a practical path to improving the system's overall performance (Ester et al., 1996; McInnes et al., 2017; Newman, 2010; Mihalcea and Tarau, 2004).

2. Task 2 Dataset Topic Analysis

In response to Task 2.2, I applied and compared multiple topic modeling techniques to iden-

tify climate-related risks and opportunities in the dataset. This section covers methodology explanation, model comparison and parameter optimization, results presentation, and limitations reflection.

2.1. Model Explanation and Rationale

To ensure robust modeling, I compared three topic modeling methods. The TF-IDF + K-means model performed poorly, with a silhouette score of 0.0176 and fuzzy cluster boundaries (Figure 2). LDA, with five topics, yielded low coherence (0.3341) and high keyword redundancy (Figure 3). In contrast, BERTopic achieved 0.6440 coherence and 0.9091 diversity even before optimization.

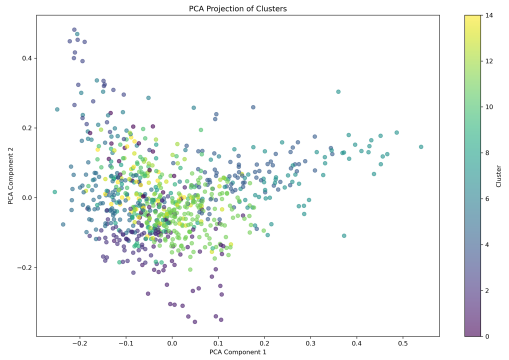

Figure 2: KMeans PCA clustering result.

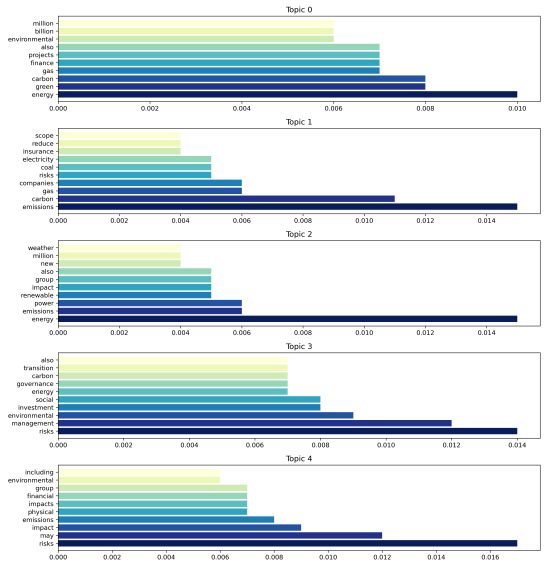

Figure 3: LDA topic keywords visualization.

Climate-related texts are often short and semantically overlapping, which limits the effectiveness of traditional models. K-means relies on distance metrics and lacks contextual understanding (Ester et al., 1996), while LDA struggles with short texts and produces vague topic divisions (Řehůřek and Sojka, 2010).

BERTopic addresses these challenges through semantic embeddings and density-based clustering, making it the most suitable modeling approach for this study (Mikolov et al., 2013; Hamilton et al., 2016; McInnes et al., 2017).

2.2. Selected Model Parameter Optimization

To improve BERTopic's performance, I optimized five key components:[2] the embedding model (S-BERT), UMAP parameters, HDBSCAN configuration, CountVectorizer settings, and topic representation method (KeyBERTInspired). Six sets of climate-related seed keywords were also introduced for guided modeling. I used a weighted scoring metric—0.7 × Topic Coherence + 0.3 × Topic Diversity—to balance semantic clarity and coverage. The final model incorporated all optimal settings.

2.3. Results and Topic Sentiment Identification

The final BERTopic model identified six topics covering policy risks, environmental governance, climate impacts, energy transition, and green technology. Based on sentiment classification, texts were labeled as "Risk," "Opportunity," or "Neutral." Table 3 summarizes the topic-wise distribution of texts and sentiment proportions.

Topic	Total	Risk	Opp.
Risk & climate	385	46.2%	21.8%
ESG governance	231	53.2%	38.1%
Climate impact	124	8.1%	0.0%
Wind & infrastructure	29	17.2%	72.4%
Arctic exploration	16	50.0%	6.2%
EVs & green energy	15	33.3%	53.3%
Total	800	41.1%	25.2%

Table 3: Text count and sentiment proportions (Risk and Opportunity) for each topic

The analysis reveals clear sentiment patterns across topics. Risk-oriented themes dominate, particularly those related to climate policy and corporate responsibility (e.g., ESG), often conveying concern or caution. In contrast, topics such as wind energy development and electric vehicles are framed more optimistically, highlighting innovation and opportunity, and evoking emotions like hope and confidence. Meanwhile, discussions of climate impacts—such as rising temperatures and sea levels—are largely factual and descriptive, thus exhibiting a neutral tone.

[2]The full tuning process and evaluation results are available in the supplementary Jupyter Notebook `Task2.2 Climate Change Text Topic Modeling.ipynb`.

2.4. Model Limitations and Future Work

Although BERTopic demonstrated strong performance in topic quality, clarity, and emotional interpretability, several limitations remain. The model relies heavily on the embedding quality and clustering parameters, which may lead to unstable boundaries or semantic confusion in cases of limited samples or topic overlap (McInnes et al., 2017). Its clustering hyperparameters strongly influence topic granularity but lack automatic tuning criteria (Ester et al., 1996). Seed word use in guided modeling introduces subjective bias, potentially restricting novel topic discovery (Mihalcea and Tarau, 2004). Additionally, emotion labeling based solely on sentence-level classifiers can misclassify ambiguous or ironic texts due to lack of contextual awareness (Newman, 2010).

Future work could incorporate context-aware emotion models (e.g., BERTweet or prompt-based classifiers) and explore hybrid guided strategies that combine unsupervised and semi-supervised learning, aiming to improve robustness and interpretability in semantically complex corpora.

3. Task 3 Named Entity Recognition Model Architecture

In Task 3, I built a BERT-based NER model for the BTC dataset, using BIO tagging and contextual embeddings. I explained the tagging process, tokenizer alignment, and feature choices. Performance was evaluated with precision, recall, and F1-score, alongside error analysis highlighting common misclassifications and improvement directions.

3.1. Model Design and Tagging Strategy

I implemented a BERT-based sequence labeling model for named entity recognition on the BTC dataset. The model consists of a `bert-base-cased` encoder for contextualized representations, followed by a dropout layer (rate = 0.1) and a linear classifier that outputs BIO-formatted entity labels for each token. Training utilized the AdamW optimizer, a linear learning rate scheduler with warmup, and gradient clipping to ensure stability.

To address BERT's subword tokenization, we used the `word_ids()` method from the fast tokenizer to map subwords to their original token indices. Entity labels were assigned only to the first subword of each token, while subsequent subwords were marked as `-100` and ignored during training via `CrossEntropyLoss(ignore_index=-100)`. This ensured proper label alignment under the BIO scheme (Jurafsky and Martin, 2020).

For example, in the sentence `["EPA", "is", "based", "in", "Washington"]`, with entity spans `[0:1, ORG]`, `[4:5, LOC]`, the corresponding label sequence would be `["B-ORG", "O", "O", "O", "B-LOC"]`, which served as the training supervision signal.

Input processing was handled by a custom `NER-Dataset` class, managing tokenization, label alignment, attention masks, and padding to ensure compatibility with the BERT architecture. The model relies entirely on contextual embeddings and does not incorporate manual linguistic features. The design was informed by prior work on CoNLL-2003, WNUT17, and noisy-text NER. Given the informal and noisy nature of tweets, a CRF layer was excluded to prevent overfitting on unstable label transitions; instead, a token-level softmax classifier was used for greater robustness.

3.2. Training Setup and Dataset Split

Experiments were conducted on the Broad Twitter Corpus (BTC) with predefined training, validation, and test splits, each balanced in entity distribution and mutually exclusive in samples.

We trained the model for 5 epochs with a batch size of 16 and learning rate of 5e-5. Dropout was disabled during evaluation. Training followed a consistent setup without early stopping, using a fixed random seed and preserving all hyperparameters for reproducibility.

3.3. Model Architecture and Design Considerations

The BertNER model adopts a standard BERT-based sequence labeling framework with contextual encoding, dropout regularization, and linear classification. This structure offers strong transferability and robustness for noisy, context-dependent inputs. However, limitations include the absence of explicit modeling for label transitions, which can cause BIO boundary mismatches, and reduced generalization to organizational entities, abbreviations, and variant spellings.

3.4. Evaluation and Results

I evaluated the NER model using standard token-level metrics: Precision, Recall, and F1-score, computed for each entity type and BIO tag. Despite an overall accuracy of 95%, it was not a reliable indicator due to the dominance of the "O" label. I therefore used macro-averaged F1 as the primary metric, which treats all classes equally and better reflects performance under label imbalance.

Table 4 shows that the model performed best on person entities (F1 = 0.85 for B-PER, 0.89 for I-PER). Location recognition was moderate (F1 <0.73), while organization entities remained chal-

lenging, with F1 scores of 0.56 (B-ORG) and 0.60 (I-ORG).

Label	Prec.	Rec.	F1	Supp.
B-LOC	0.78	0.69	0.73	636
B-ORG	0.71	0.46	0.56	1090
B-PER	0.93	0.79	0.85	2650
I-LOC	0.76	0.70	0.73	208
I-ORG	0.72	0.51	0.60	246
I-PER	0.90	0.88	0.89	269
O	0.96	0.99	0.98	30325
Accuracy	–	–	**0.95**	35424
Macro avg	**0.82**	**0.72**	**0.76**	35424
Weighted avg	**0.95**	**0.95**	**0.95**	35424

Table 4: Performance metrics for each entity type and position

To better illustrate class-wise performance, I visualized the results using a bar chart (Figure 4), which highlights the strong performance on PER labels and the relative weakness on ORG entities.

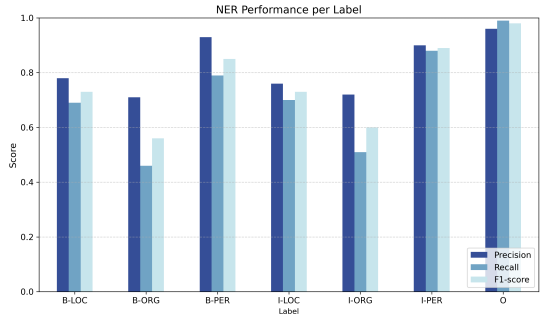

Figure 4: NER performance across different entity types

To further investigate model behavior, I analyzed a normalized confusion matrix based on prediction probability (Figure 5). The results show that ORG tokens are frequently misclassified as “O” or “PER,” and LOC tokens are occasionally labeled as ORG. These errors reflect the inherent ambiguity in organizational name recognition and indicate challenges in boundary identification, particularly in noisy or context-dependent inputs.

3.5. Error Analysis and Discussion

To better understand model limitations, I conducted error analysis based on confusion matrices and misclassified examples. Three major error types emerged: entity omission, category confusion, and boundary misalignment.

The most severe issue was entity omission, particularly for organization names. In *“President Obama met with Apple in Washington,”* “Apple” was mislabeled as LOC, and “Washington” as I-ORG, with 31.96% of B-ORG tokens incorrectly tagged as “O,” leading to low ORG recall.

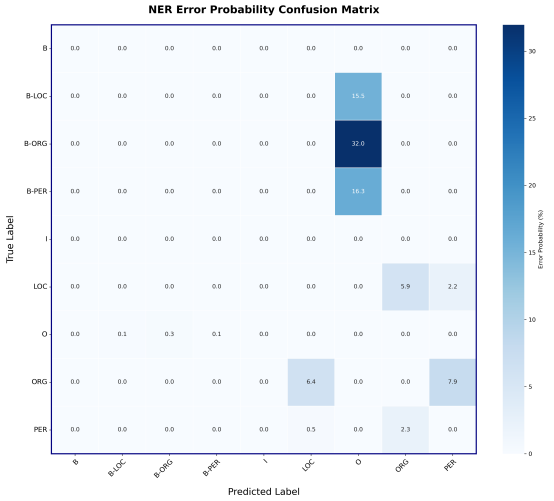

Figure 5: Normalized confusion matrix based on prediction probabilities for the NER model.

The second issue was category confusion, especially between ORG and LOC or PER. Ambiguous terms like “WHO” or “Apple” were often misclassified, as seen in *“The WHO released a new report,”* where “WHO” was labeled as “O” instead of B-ORG.

Lastly, boundary errors in BIO tagging were observed: 20 B-tags predicted as I, and 31 I-tags as B, indicating limited structural learning without a CRF layer (Jurafsky and Martin, 2020).

To mitigate these issues, future improvements could include adding a CRF layer to enforce label transitions, adopting BERTweet or RoBERTa for social media text, and incorporating external resources like gazetteers. Class reweighting or hard example mining may also help distinguish commonly confused entity types.

While the model performs well on frequent entities like persons, it struggles with noisy, ambiguous, or rare cases. Enhancing contextual understanding and structural coherence remains essential for robust NER in informal texts.

4. References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O’Reilly Media.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases

with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Daniel Jurafsky and James H. Martin. 2020. *Speech and Language Processing*, 3rd edition. Pearson.

Anke Lüdeling and Merja Kytö. 2008. *Corpus Linguistics: An International Handbook*. De Gruyter.

Leland McInnes, John Healy, and James Melville. 2017. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mark EJ Newman. 2010. *Networks: An Introduction*. Oxford University Press.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.