EMATM0067 Text Analytics Coursework

April 26, 2025

1. Task 2: Climate Sentiment – Report

2.1a Naive Bayes Modifications (3%)

In 1.1d, I tried several modifications to improve the Naïve Bayes classifier. While the original model used `ngram_range=(1,2)` with both unigrams and bigrams acheived good accuracy (74%), the best result (75.5%) was achieved using unigrams only. This reduced feature sparsity and better aligned with the model's assumption of conditional independence, as individual words like *risk* and *transition* already provided sufficient sentiment signal. Using only bigrams decreased accuracy (73.5%), probably because they were too specific and led to overfitting. Removing English stop words also decreased performance (65.5%), since sentiment-related words like *not* and *never* carry strong polarity in this domain, but were excluded. Finally, replacing `CountVectorizer` with `TfidfVectorizer` significantly reduced accuracy (50%), as the Naïve Bayes classifier relies on raw count estimates; TF-IDF reweighting disrupted its probabilistic assumptions and suppressed useful signals.

2.1b Method Comparison and Interpretation (10%)

Model	Validation Accuracy	Test Accuracy
Naïve Bayes	0.755	0.790
FFNN	0.550	0.490
TinyBERT	0.750	0.785

Table 1: Performance comparison on validation and test sets

As shown in Table 1, Naïve Bayes achieved the highest accuracy overall, benefiting from a unigram representation that effectively identified high-frequency sentiment-related keywords. Despite the assumption of feature independence, it performed well on structured and keyword-rich texts.

In contrast, the FFNN performed worst, likely due to its randomly initialized embeddings and limited model capacity, which constrained its ability to capture non-linear semantic relationships and contextual nuances. TinyBERT is a compact pretrained transformer model, demonstrated strong generalization even in this low-resource setting, revealed the advantages of transfer learning in NLP.

Error analysis showed that Naïve Bayes, FFNN, and TinyBERT misclassified 49, 89, and 50 samples, respectively. Misclassifications commonly involved semantically ambiguous or neutral texts. For example, "Insights or commitments we have gained from the TCFD process..." (positive) and "An internal analysis of the generation fuel mix..." (neutral) were likely misclassified due to weak lexical signals or domain-specific terminology. Even clearly positive examples like "IFC is helping reverse that decline." were misclassified due to indirect phrasing. In terms of error patterns: Naïve Bayes was highly sensitive to high-frequency emotional keywords; FFNN struggled with contextual reasoning; and TinyBERT, while generally context-aware, still showed limitations with subtle language and domain shifts.

Future improvements could include pretrained embeddings (e.g., GloVe or static BERT) to improve FFNN's semantic representation, along with syntax- or discourse-based features to capture structural cues. Since both FFNN and TinyBERT struggled with ambiguous or neutral texts, better class balancing (e.g., Focal Loss) could help address this. For TinyBERT, further fine-tuning, soft-label strategies, or data augmentation may enhance its robustness on borderline cases.

2.2 Topic Identification (25%)

To identify climate-related risks and opportunities in ESG reports, we used BERTopic—an unsupervised topic modeling method that combines contextual BERT embeddings with UMAP for dimensionality reduction and HDBSCAN for clustering. This approach was chosen over traditional models like LDA with TF-IDF because it better captures semantic similarity in complex, domain-specific language. The use of BERT embeddings helps preserve contextual meaning in climate-related narratives, while HDBSCAN avoids the need to predefine the number of topics, making the model more flexible. BERTopic also provides built-in visualization and topic probability outputs, which support interpretation and analysis. However, it has some limitations: results can be sensitive to parameter setting, topic boundaries are sometimes vague, and topic labeling is still done manually, which introduces a degree of subjectivity.

To evaluate topic modeling performance on ESG-related climate texts, we compared three unsupervised methods: LDA, KMeans with BERT embeddings, and BERTopic. These represent classical probabilistic modeling, vector-based clustering, and modern embedding-driven frameworks, respectively. All models were applied to the same 1,000-document corpus, with LDA and KMeans extracting 18 topics to align with BERTopic's output.

As shown in Table 2, BERTopic outperformed the other two methods in coherence and diversity. LDA groups words based on co-occurrence but lacks contextual understanding. It produced overlapping topics with frequent keyword repetition (e.g., "climate," "risk") and a moderate coherence score of 0.4363. KMeans with MiniLM embeddings captured contextual semantics, but due to its hard clustering mechanism, failed to represent nuanced topic boundaries, reflected by a low silhouette score of 0.0195. BERTopic, which combines BERT embeddings with UMAP and HDBSCAN, automatically determined the number of topics and used `c_TF_IDF` for topic generation. It achieved the highest coherence (0.5821) and diversity (0.7450), and better differentiated between risk and opportunity themes. While its output can vary due to random initialization, fixing the seed ensures reproducibility.

Method	Coherence	Diversity	Silhouette
LDA	0.4363	—	—
KMeans + BERT	—	—	0.0195
BERTopic	0.5821	0.7450	—

Table 2: Comparison of three topic modeling methods

To further assess parameter sensitivity, we compared two BERTopic models with different `min_topic_size` values (10 vs. 30), which affect topic granularity. As shown in Figure 1, the model with `min_topic_size=10` produced 17 well-separated topics with clearer boundaries, suggesting higher semantic resolution. In contrast, the `min_topic_size=30` model generated only 4 broader topics with overlapping clusters, which reduced interpretability.

The smaller-size model also achieved higher coherence (0.5685 vs. 0.5462) and diversity (0.7833 vs. 0.7200), as shown in Table 3, indicating better internal consistency and topical variety.

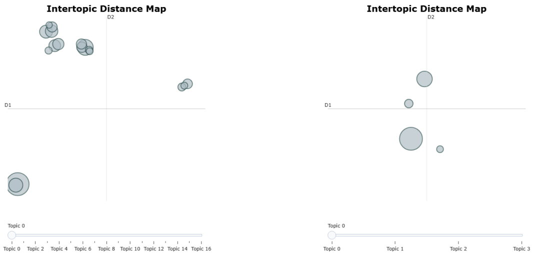
Parameter Setting	Topics	Coherence	Diversity
min_topic_size = 10	17	0.5685	0.7833
min_topic_size = 30	4	0.5462	0.7200

Table 3: Effect of `min_topic_size` on BERTopic results

Overall, while smaller topic sizes may increase fragmentation, they produce more interpretable and semantically rich themes—especially valuable for identifying nuanced ESG risks and opportunities. This shows that topic granularity settings significantly affect interpretability, and careful tuning is important when analyzing complex ESG narratives.

Topic	Total	% Opp.	% Risk	% Neutral	Purity	Entropy	Dominant Category
1	151	96.0	3.3	0.7	96.0	0.267	Opportunity
2	73	21.9	57.5	20.5	57.5	1.408	Risk
3	57	73.7	26.3	0.0	73.7	0.831	Opportunity
4	48	0.0	97.9	2.1	97.9	0.146	Risk
5	45	6.7	73.3	20.0	73.3	1.053	Risk
6	40	0.0	12.5	87.5	87.5	0.544	Neutral
7	37	0.0	29.7	70.3	70.3	0.878	Neutral
8	32	6.2	78.1	15.6	78.1	0.947	Risk
9	30	13.3	80.0	6.7	80.0	0.906	Risk
10	30	36.7	43.3	20.0	43.3	1.518	Risk
11	27	3.7	3.7	92.6	92.6	0.455	Neutral
12	18	11.1	22.2	66.7	66.7	1.224	Neutral
13	17	41.2	52.9	5.9	52.9	1.253	Risk
14	15	0.0	13.3	86.7	86.7	0.567	Neutral
15	14	100.0	0.0	0.0	100.0	0.000	Opportunity
16	14	0.0	21.4	78.6	78.6	0.750	Neutral
17	12	41.7	41.7	16.7	41.7	1.483	Opportunity

Table 4: Label distribution and evaluation metrics for BERTopic (min_topic_size = 10)



(a) min_topic_size = 10 (b) min_topic_size = 30

Figure 1: Intertopic distance maps for two BERTopic models with different granularity settings

To evaluate how well the identified topics align with the original task objective, we mapped each topic from the BERTopic model (`min_topic_size=10`) to the climate sentiment labels (Opportunity, Risk, Neutral). For each topic, we calculated the label distribution and used two evaluation metrics—Purity (the proportion of the dominant label) and Entropy (label dispersion)—to assess semantic consistency. The results are shown in Table 4.

Most topics demonstrate a clear alignment with a dominant label. For example, Topic 1 (keywords: *physical*, *risk*, *climate change*) contains 96% Opportunity-labeled texts with low entropy (0.267), indicating a coherent semantic cluster. In contrast, Topics 2 and 17 show more mixed label distributions ($entropy > 1.4$), suggesting ambiguous topic boundaries. Topics 6, 11, and 12 are dominated by Neutral texts, likely reflecting vague or context-dependent content where BERT embeddings struggle to capture clear sentiment.

Overall, BERTopic shows strong capability in identifying semantically coherent themes relevant to climate risks and opportunities. The use of class label distribution, Purity, and Entropy provided a structured way to assess topic quality, and most clusters were focused and interpretable.

However, several limitations still remain. First, BERTopic is unsupervised and lacks contextual disambiguation—terms like “targets” or “disclosure” may shift meaning across contexts, leading to semantic drift. Second, results are sensitive to parameters like `min_topic_size`, affecting topic granularity. Finally, because labels were only used post-hoc, the method lacks task-level guidance, and interpretation still relies on human judgment, introducing subjectivity.

2. Task 3:Named Entity Recognition on Twitter

2.1. Sequence Tagger Construction

2.1.1. Description of Model Structure Design

We adopted a Transformer-based NER approach using `vinai/BERTweet-base` model [1], a BERT encoder pre-trained on large-scale Twitter data, making it well suit the noisy, informal style of the Broad Twitter Corpus (BTC).

Tokens are processed with BERTweet's subword tokenizer. Due to the lack of fast tokenizer support, manual label alignment was implemented by assigning labels only to the first sub-token, ignoring others during loss computation.

Model outputs are generated by projecting contextual embeddings through a linear layer, trained with CrossEntropyLoss to predict BIO tags.

This setup demonstrates strong contextual modeling and domain adaptation, but increases engineering complexity and struggles with rare entities or heavily misspelled tokens.

2.1.2. Token and Label Alignment

The BTC dataset provides entity tags at the token level, while BERTweet's byte-level BPE tokenizer may split tokens into subwords. To align labels, we:

1. Tokenize input with `is_split_into_words=True` to preserve word boundaries.

2. Track sub-token counts for each original token.

3. Assign labels only to the first sub-token; mask others with `-100` during loss computation.

This produced a label sequence that matches the length of the model input. Here are some examples:

Token	Subword(s)	Label(s)
Angela	Angela	B-PER
Merkel	Mer`@@`, kel	I-PER, -100
met	met	O
in	in	O
Berlin	Ber`@@`, lin	B-LOC, -100

This method ensures label alignment with model inputs, enabling accurate entity boundary learning while slightly increasing implementation complexity.

2.1.3. Entity Span Encoding

Entity labels follow the BIO tagging scheme: `B-XXX` marks the beginning, `I-XXX` marks inside an entity, and `O` marks non-entity tokens.

For example, in *"Angela Merkel met in Berlin"*: "Angela" is tagged `B-PER` (start of a person entity), "Merkel" as `I-PER` (continuation), "Berlin" as `B-LOC` (start of a location entity), and "met" and "in" as `O` (non-entity tokens).

This scheme clearly defines entity boundaries and supports multi-token entity modeling.

2.1.4. Feature Selection and Hypotheses

To address the noise in Twitter text, we selected the following input features:

Subword tokenization: Captures internal morphology, improving recognition of misspellings and abbreviations, and reducing out-of-vocabulary (OOV) issues.

Contextualized embeddings: Transformer-based embeddings use surrounding context to resolve noisy or incomplete sentences and capture long-range dependencies.

BIO tagging: Explicitly marks entity spans, improving multi-token entity recognition and reducing boundary errors.

Domain-adapted vocabulary: BERTweet's vocabulary includes Twitter-specific expressions and informal expressions , enhancing recall over general models like `bert-base-cased`.

2.2. Method Evaluation and Result Discussion

2.2.1. Performance Metrics and Limitations

NER performance is evaluated using Precision, Recall, and F1-score at the token level, macro-averaged across entity types (PER, LOC, ORG) to avoid bias from the dominant O tag. F1-score serves as the primary metric for balancing precision and recall.

However, Token-level metrics cannot fully capture entity boundary correctness or differentiate error types. Therefore, we complement them with error sample analysis and confusion matrices for deeper evaluation.

2.2.2. Testing Procedure

We used the Broad Twitter Corpus (BTC), pre-split into train, dev, and test sets with balanced label distributions and are mutually exclusive.

Model was optimized on the training set and monitored on the dev set. Early stopping based on dev F1-score was abandoned due to the observed stable improvements.

Final evaluation was performed on the test set using the best dev checkpoint. Metrics reported include Precision, Recall, and F1-score. We used HuggingFace's Trainer API with `evaluation_strategy="epoch"`, selected by `eval_f1`, and fixed the random seed (`seed=42`) for reproducibility.

Cross-validation was not used, due to BTC's predefined split and the high computational cost of fine-tuning.

2.2.3. Results and Performance Analysis

Table 5 summarizes model performance on the BTC test set:

Model	**P**	**R**	**F1**	**Loss**
BERTweet	0.8074	0.7975	0.8004	0.1348

Table 5: Evaluation metrics on BTC test set

Compared to `bert-base-cased` (F1 = 0.2177), BERTweet achieved a 58-point more F1 gain, confirming the benefit of domain-specific pretraining.

Figure 2 shows dev set metrics by epoch. Performance steadily improved, with epoch 3 yielding the best results used for final testing.

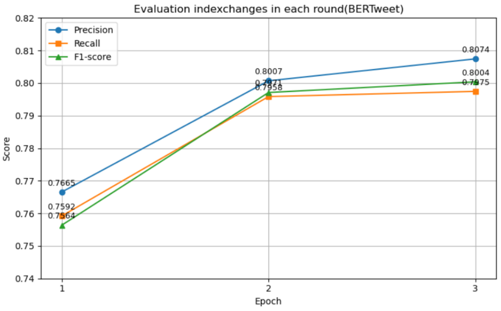


Figure 2: Evaluation scores per epoch (dev set)

Figure 3 illustrates F1-scores by entity type. Recognition was highest for PER, then LOC, with ORG lowest, likely due to data scarcity and naming ambiguity.

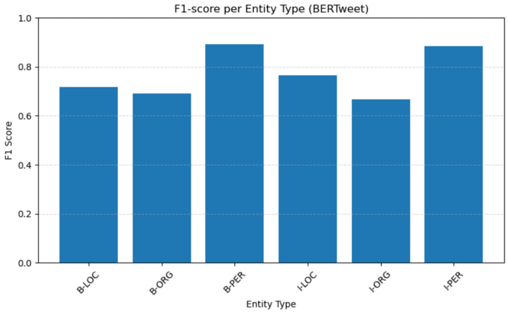


Figure 3: F1-score by entity type (test set)

2.2.4. Error Analysis and Future Improvements

Despite strong performance of BERTweet, typical errors include partial recognition of multi-word entities (e.g., recognizing only "Angela" in "Angela Merkel"), confusion between ORG and PER when organization names resemble person names, and omission of entities due to abbreviations or misspellings.

These errors revealed challenges in subword alignment, label ambiguity, and noise handling. Future work may explore character embeddings, spelling normalization, entity prompts, and data augmentation to improve robustness, especially more fine-grained, noise-tolerant strategies for rare or informal expressions.

REFERENCES

[1] Nguyen, D. Q., Vu, T., and Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14. https://aclanthology.org/2020.emnlp-demos.2/