Text Analytics Assessment

Contents

1 Task: Climate Sentiment

1.1 Evaluation of Three Methods for Climate Sentiment Classification

1.1.1 Modifications to the Naïve Bayes Classifier

The Naïve Bayes classifier was improved by cleaning the text more carefully. I used steps like turning all words into lowercase, breaking the text into words (tokenizing), removing common words (stop words), and reducing words to their basic form (lemmatization with WordNet). This helped remove extra noise and made the features better by treating similar words as the same. I also trained the model on this cleaned-up text and adjusted the settings (hyperparameters) to get the best results. I tried adding bigrams and using TF-IDF features to boost performance sci, but these changes didn't help much, probably because the texts were short or the data was too sparse.

1.1.2 Comparison of Results and Interpretation

Model	Accuracy	Macro F1	Notable Strengths	Notable Weaknesses
Naïve Bayes	0.771	0.76	High precision for class 0 (0.89)	Lower recall for class 2 (0.60)
ClimateBERT NN	0.741	0.75	Best recall for class 1 (0.76)	Lower precision for class 0 (0.81)
TinyBERT	0.766	0.77	Balanced performance overall (0.88)	Lower recall for class 1 (0.66)

Table 1: Climate Sentiment Model Performance

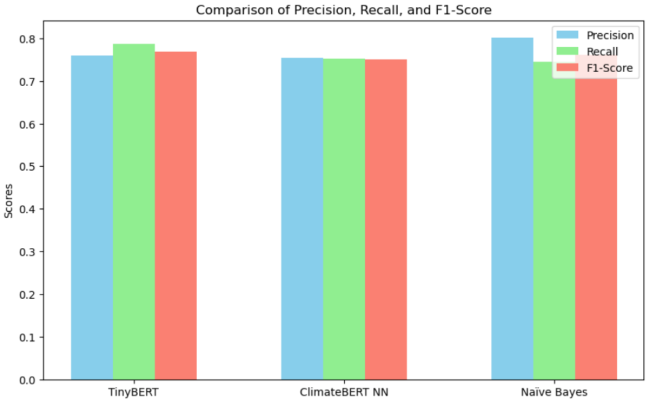


Figure 1: Comparison of Precision, Recall, and F1-score

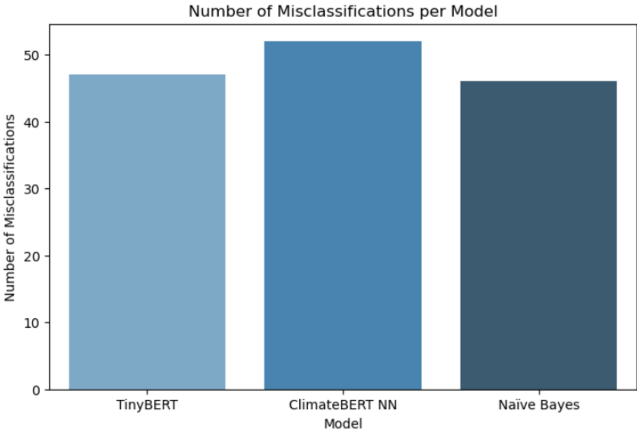
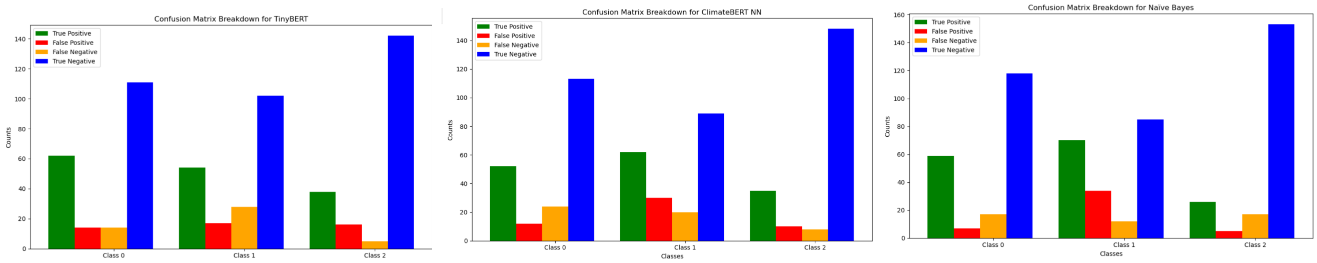


Figure 2: No. of Misclassifications per Model

- **Naïve Bayes**: Strong for neutral/operational class, but struggles with minority classes due to class imbalance.
- **ClimateBERT NN**: Uses transfer learning, improving recall for nuanced classes but sometimes confuses neutral and risk.
- **TinyBERT**: Most balanced, leveraging deep contextual representations, but still confuses risk and opportunity in overlapping contexts Devlin et al. [2019], Jiao et al. [2020], Webersinke et al. [2022].



((a)) TinyBERT

((b)) ClimateBERT NN

((c)) Naive Bayes

Figure 3: Confusion Matrix

1.1.3 Naïve Bayes Classifier: Mathematical Foundation

The Naïve Bayes classifier is based on Bayes' theorem and the assumption of conditional independence between features (words) given the class Manning and Schütze [1999], Zhang [2004]. The decision rule is:

$$\hat{c} = \arg\max_{c \in C} P(c) \prod_{f \in F} P(f|c) \tag{1}$$

Where:

- $P(c)$: Prior probability of class c
- $P(f|c)$: Probability of feature (word) f given class c
- F: Set of features (words) in the document

1.1.4 Misclassification Analysis & Interpretation

Looking at the mistakes the models made, it's clear they often mix up neutral or operational statements with risk or opportunity categories, especially when the text uses similar words (like "emissions" or "initiatives"). For example, TinyBERT sometimes predicts something as an "opportunity" when it's actually a "risk," probably because the same words show up in both types of texts. This shows how tricky it is to pick up on small differences in tone when analyzing climate-related documents.

1.1.5 Future Improvements

- Data augmentation and more granular annotation could help models learn finer distinctions.
- Ensemble methods may combine the strengths of different classifiers.
- Domain-specific pretraining (e.g., further tuning BERT on climate texts) could further improve contextual understanding.

1.2 Identifying Topics Associated with Climate-Related Risks or Opportunities

1.2.1 Methods for Topic Identification

Two main unsupervised topic modeling approaches were used:

1. **Latent Dirichlet Allocation (LDA)**: We used LDA on both Bag-of-Words and TF-IDF versions of the data. LDA finds topics by grouping together words that often appear together, making it easier to understand the main themes based on the most common words in each topic.

2. **Top2Vec**: Top2Vec is a method that groups documents by their meanings and finds topic words based on how similar they are Angelov [2020]. It's good at picking up more detailed, context-aware topics, which is especially helpful for short or technical texts.

1.2.2 Motivation and Limitations

- LDA is interpretable and widely used but assumes topics are multinomial over words and may struggle with short texts or overlapping vocabulary.
- Top2Vec leverages deep embeddings, capturing semantic similarity more effectively, but can be sensitive to hyperparameters and requires more computational resources.

1.2.3 Comparison of Variations

- **LDA with Bag-of-Words vs. TF-IDF**: Both approaches yielded similar topics, with key words like "energy," "risk," "emissions," and "climate" dominating the top lists. TF-IDF did not substantially alter topic coherence, likely due to the prevalence of domain-specific terms.
- **Top2Vec** found four main topics, with clear clusters related to "risk," "stakeholders," "emissions," and "sustainability." Searching for topics related to "risk" or "opportunity" yielded interpretable results: for "risk," top topics included words like "risk," "liabilities," and "compliance"; for "opportunity," topics featured "renewables," "sustainability," and "investments."

1.2.4 Results and Visualization

LDA Topics (Bag-of-Words/TF-IDF):

- Topic 0: “energy,” “risk,” “carbon,” “climate”
- Topic 1: “emissions,” “gas,” “coal,” “oil”
- Topic 2: “climate,” “risk,” “change,” “environmental”

Top2Vec Topics:

- Topic 0: “stakeholders,” “renewables,” “sustainability,” “emissions”
- Topic 1: “climate,” “risk,” “disasters,” “environment”
- Topic 2: “risk,” “liabilities,” “compliance”
- Topic 3: “emissions,” “pollution,” “regulatory”

Top topics related to ‘risk’:

- Topic 2: “risk,” “liabilities,” “compliance”
- Topic 1: “climate,” “risk,” “disasters”

Top topics related to ‘opportunity’:

- Topic 0: “renewables,” “sustainability,” “investments”

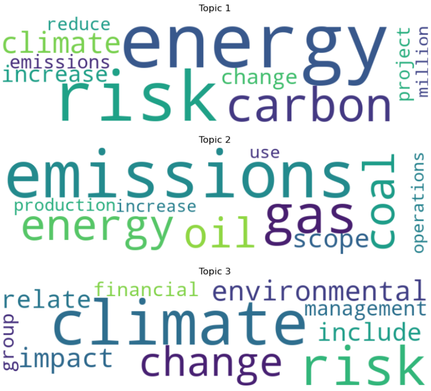


Figure 4: Word Cloud: TF-IDF

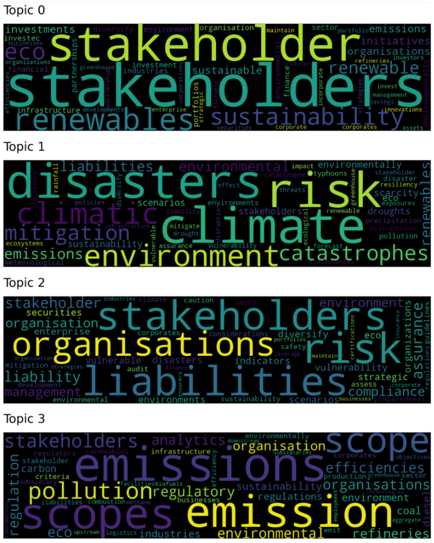


Figure 5: Word Cloud: Top2Vec

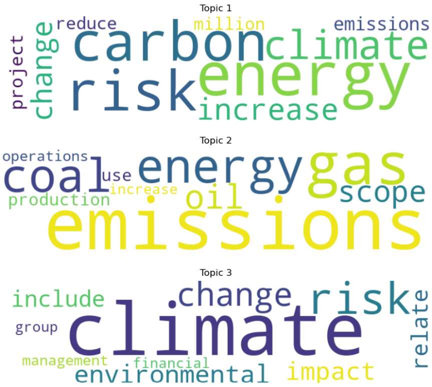


Figure 6: Word Cloud: LDA

Visualizations further highlighted the prominence of these themes and their relative document sizes.

1.2.5 Interpretation and Limitations

Both LDA and Top2Vec found groups related to climate risks (like regulations, liabilities, and disasters) and opportunities (such as renewables and sustainability) C et al. [2022], University [2023]. However, LDA sometimes combined different themes because it focuses on word frequency, while Top2Vec, which understands meaning better, sometimes mixed unrelated ideas because of limitations in how it groups words.

Summary of Limitations:

- LDA: May conflate topics with overlapping vocabulary; less effective for short texts.
- Top2Vec: Sensitive to parameter choices and may require large corpora for stable topics.
- General: Both methods are unsupervised and require manual interpretation; results depend on preprocessing quality and dataset representativeness.

1.3 Conclusion

Using transformer-based classification alongside embedding-driven topic modeling creates a strong approach for analyzing climate sentiment and identifying risk and opportunity themes. Future work should look into supervised topic models, training on domain-specific data, and adding external knowledge sources to gain deeper, more useful insights.

2 Task: Named Entity Recognition on Twitter

2.1 Design and Implementation of a Sequence Tagger

2.1.1 Chosen Method and Rationale

For Named Entity Recognition (NER) on the Broad Twitter Corpus (BTC), I implemented Conditional Random Fields (CRFs) as the sequence tagging method Lafferty et al. [2001]. Two models were built:

- **Basic CRF**: Uses only the current word as a feature.
- **Custom CRF**: Incorporates a rich set of linguistic and contextual features.

CRFs are probabilistic models used for tasks where labels are assigned to sequences. They calculate the likelihood of a sequence of labels based on the input, making them ideal for Named Entity Recognition (NER), where understanding the context is key.

Mathematical Formulation The CRF models the conditional probability of a label sequence $Y = (y_1, \ldots, y_n)$ given an input sequence $X = (x_1, \ldots, x_n)$ as:

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{t=1}^{n} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, X, t)\right) \tag{2}$$

where f_k are feature functions, λ_k are their learned weights, and $Z(X)$ is the normalization factor.

Strengths

- CRFs capture dependencies between labels (e.g., `I-PER` usually follows `B-PER`).
- They allow for flexible feature engineering, enabling the use of domain knowledge.

Limitations

- Require manual feature design (unlike neural models).
- Performance depends on quality and diversity of features.
- Do not leverage deep contextual embeddings unless combined with neural networks.

2.1.2 Tokenization and Tag Alignment

The BTC dataset already has data that's been tokenized and labeled with BIO tags. Since CRFs work with word-level tokens and don't need subword tokenization, no further alignment is needed.

2.1.3 Entity Span Encoding Example

Entities in the BTC dataset are encoded using the BIO scheme Ramshaw and Marcus [1995]:

- `B-PER`: Beginning of a person entity
- `I-PER`: Inside a person entity
- `O`: Outside any entity

Example:

```
Sample sentence tokens: ['How', 'did', 'Dorothy', 'Gale', 'come', 'back', '*', 'younger', '*', 'in'..]
Sample sentence tags:   ['O', 'O', 'B-PER', 'I-PER', 'O', ...]
Extracted entity spans: {'PER': [(2, 4, 0)]}  # "Dorothy Gale"
```

2.1.4 Feature Design and Rationale

Basic CRF:

- Only uses the current word as a feature.

Custom CRF:

- **Token-level features**: word, word shape (e.g., Xxxx, xxxx), capitalization, presence of digits, punctuation, suffixes/prefixes, character n-grams.

- **Contextual features**: previous/next words, previous/next tags.
- **Linguistic features**: POS tags.
- **Twitter-specific features**: detection of mentions (@), hashtags (#), and URLs.

Rationale:

- Word shape, capitalization, and POS help distinguish names and entities (e.g., "London" vs "london").
- Contextual features capture dependencies between entities (e.g., "of London" likely signals a location).
- Twitter-specific features are crucial due to the informal and noisy nature of tweets.
- Suffixes/prefixes and character n-grams help generalize to unseen words (e.g., "ington" in "Washington").

Hypothesis: Richer features will improve recall and precision, especially for entities with ambiguous or rare surface forms.

2.2 Evaluation, Interpretation, and Discussion

2.2.1 Performance Metrics

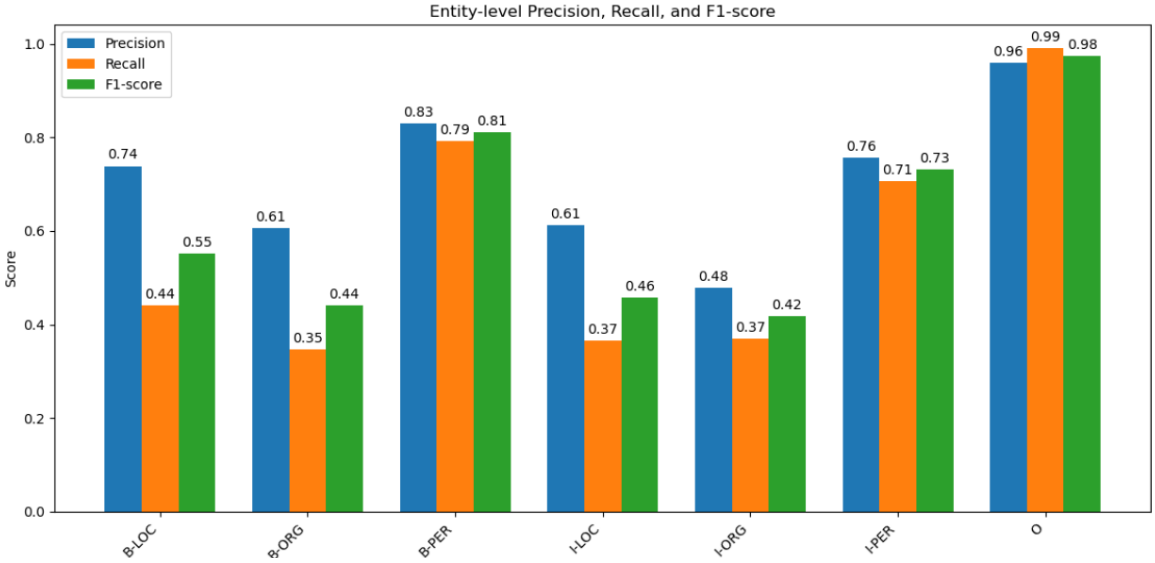


Figure 7: Entity-level precision, recall, and F1-score for each named entity class label, illustrating model performance across different entity types.

- **Token-level accuracy**: Fraction of correctly tagged tokens.
- **Precision, Recall, F1-score (per entity type)**: Standard for NER, reflecting both correctness and completeness.
- **Span-level F1**: Measures exact match of predicted entity spans, which is stricter and more informative for real NER tasks.

Limitations:

- Token-level metrics can be misleading (e.g., high accuracy due to many "O" tokens).
- Span-level F1 is more informative for practical NER use, but harder to optimize.

2.2.2 Testing Procedure

- **Data splits**: Used BTC's predefined train and test splits.
- **Training**: Both CRF models trained on the training set.
- **Evaluation**: All metrics reported on the test set.

2.2.3 Results

Token-level Accuracy

- Basic CRF: 0.8857
- Custom CRF: 0.9364

Table 2: Token-level metrics for Custom CRF

Entity	**Precision**	**Recall**	**F1**	**Support**
B-LOC	0.74	0.44	0.55	636
B-ORG	0.61	0.35	0.44	1090
B-PER	0.83	0.79	0.81	2648
I-LOC	0.61	0.37	0.46	208
I-ORG	0.48	0.37	0.42	246
I-PER	0.76	0.71	0.73	269
O	0.96	0.99	0.98	30317

Classification Report (Custom CRF) Macro-average F1: 0.63
Weighted-average F1: 0.93

Span-level F1

- PER: 0.73
- LOC: 0.54
- ORG: 0.41
- Macro-average span-level F1: 0.56

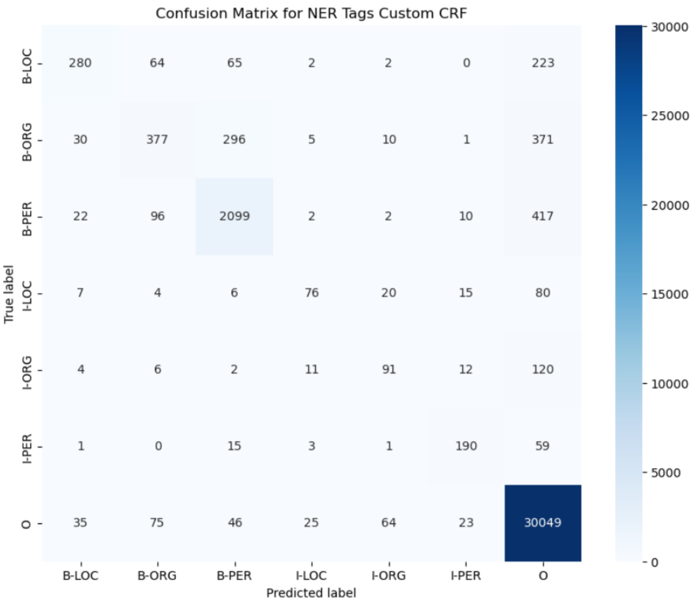


Figure 8: Confusion matrix of Custom CRF

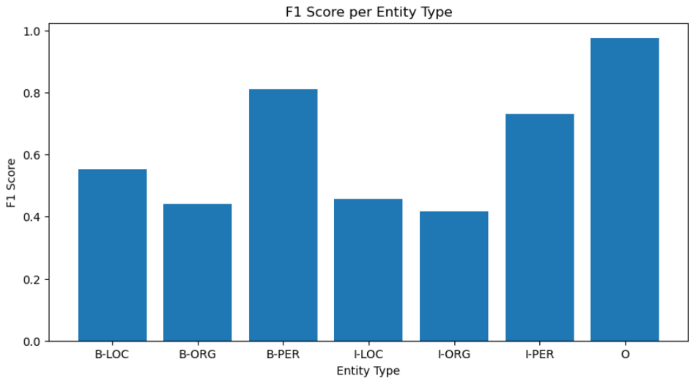


Figure 9: F1 score per Entity type

Plots

- Bar plots for F1 per entity type show “PER” entities are recognized best, “ORG” worst.
- Confusion matrices reveal most errors are confusions with the “O” label or between similar entity types.

2.2.4 Error Analysis

Common Errors:

- Confusing “ORG” with “LOC” or “O”, especially for organization names that resemble locations or common nouns.
- Missing multi-token entities (e.g., “Chinese Assoc of Vic” split incorrectly).
- Mislabeling ambiguous tokens (e.g., “Vic” as “B-LOC” instead of “I-ORG”).

Example Misclassifications:

```
Sentence: Enjoying the Chinese Assoc of Vic annual dinner right here in Aston !
Token: 'Chinese' | True: B-ORG | Pred: O
Token: 'Vic'     | True: I-ORG | Pred: B-LOC
```

The model splits the organization entity and mislabels “Vic” as a location.

Possible Improvements:

- Incorporate pre-trained word embeddings or contextual embeddings (e.g., BERT).
- Use neural sequence taggers (BiLSTM-CRF or transformers) to better capture context.
- Augment training data or use external gazetteers for rare entities.
- Further tune feature selection or regularization.

2.3 Conclusion

The CRF-based NER tagger works well with Twitter data, especially when using detailed, task-specific features, and it achieves good token-level and decent span-level F1 scores. Most errors come from unclear or multi-token entities, as well as the informal style of tweets. Future improvements could include using neural models and adding external knowledge to enhance performance.

References

scikit-learn: Machine learning in python. `https://scikit-learn.org/`. Accessed: 2025-04-28.

Dimo Angelov. Top2vec: Distributed representations of topics. `https://github.com/ddangelov/Top2Vec`, 2020.

Bala Priya C, Nishrin Kachwala, Anju Mercian, Debaditya Shome, Farhad Sadeghlo, and Hussein Jawad. Using topic modeling to understand climate change domains. `https://www.omdena.com/blog/topic-modeling-climate`, 2022.

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

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. In *Findings of EMNLP*, pages 4163–4174, 2020.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.

Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Lance A. Ramshaw and Mitchell P. Marcus. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*, pages 82–94, 1995.

Carnegie Mellon University. A topic modelling approach for developing a typology of climate change misinformation on twitter. `https://www.cmu.edu/ideas-social-cybersecurity/events/edb_a-topic-modelling-approach-for-developing-a-typology-of-climate-change-misinformation-on-twitter revised.pdf`, 2023.

Nils Webersinke, Nicolas Leippold, et al. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2211.06062*, 2022.

Harry Zhang. The optimality of naive bayes. *AAAI*, pages 562–567, 2004.