CE807-25-SP – Assignment - Final Practical Text Analytics and Report

Student ID: 2400570

Abstract

Sentiment analysis is a key task in NLP with applications in areas like mental health and social media analysis. This study compares a discriminative BiLSTM model with an unsupervised pipeline using **BERTopic** for topic modeling, **HDBSCAN** and **KMeans** for clustering, and **ADASYN** for class balancing. The dataset is imbalanced, with 20.4% negative and 79.6% positive instances. We prioritize F1 Score over accuracy for fair evaluation. The **BiLSTM** model, with a threshold of 0.7, achieves an F1 Score of 87.16%, while the unsupervised approach reaches 69.90%. These results highlight the importance of threshold tuning in supervised learning and the potential of clustering-based methods for handling class imbalance in sentiment classification.

Materials

All essential materials for the project are:

- **Code:** The source code is available at: [1].
- **Zoom Recorded Presentation:** The presentation recording is available at: [2].
- **Google Drive Folder:** A folder containing the models and saved outputs is available here: [3].

All essential materials for the project are available in the shared Google Drive folder.

1 Task 1: Model Selection Discussion

1.1 Summary of 2 selected Models

1.1.1 Unsupervised Approach (BERTopic + Clustering + HDBSCAN)

We propose a resource-efficient pipeline for topic modeling and clustering, starting with *BERTopic* for context-rich, coherent topic extraction using transformer-based embeddings. Unlike LDA or LSA, it doesn't require predefined topic counts and offers superior interpretability in complex datasets (Mutsaddi et al., 2025).

For clustering, we use *HDBSCAN*, which can discover clusters of varying shapes and densities without predefining the number of clusters. Its robustness to noise and suitability for imbalanced data make it more effective than K-means or Mean Shift (Berba, 2020).

To tackle class imbalance, we introduce *ADASYN* to synthetically augment minority samples, improving clustering performance. Finally, *K-means* is applied on the balanced dataset, delivering faster and more efficient clustering than probabilistic models like GMM (Chen et al., 2024).

This hybrid approach, combining *BERTopic*, *HDBSCAN*, *ADASYN* and *K-means*, optimizes topic extraction, clustering and class balancing, particularly in noisy, imbalanced datasets.

1.1.2 Discriminative Approach (BiLSTM)

This study presents a resource-efficient sentiment classification pipeline using a **Bidirectional Long Short-Term Memory (BiLSTM)** network. The model classifies text into positive and negative sentiment by capturing both past and future dependencies(Liu and Guo, 2019).

BiLSTM is chosen for its ability to model long-range dependencies, outperforming unidirectional RNNs and traditional classifiers on noisy, short-text datasets(Jang et al., 2020). Its architecture balances high accuracy with computational efficiency, making it ideal for resource-constrained environments.

1.2 Critical discussion and justification of model selection

1.2.1 Unsupervised Approach

We propose a resource-efficient pipeline for topic modeling and clustering, combining **BERTopic**,

[1] Code: https://drive.google.com/file/d/1mhjkS1jETnNjFOpimfaDIKuFMjJpnGhp/view?usp=sharing

[2] Zoom Link: https://essex-university.zoom.us/rec/share/_gJT_8PAlSa5DrB1A7t1kWmFd5hmyw99kOIe_rX3oRe3g7TENs_mAvjROoUuef0i.X3vFdhL-u193NnNl

[3] Google Drive Folder: https://drive.google.com/drive/folders/1QIbWh3nkwmwP4LmZEuoCeaPe9eS9zVeK?usp=sharing

HDBSCAN, **ADASYN**, and **K-means**, tailored for short, noisy, and imbalanced text data. *BERTopic* is chosen for its use of transformer-based embeddings to extract coherent topics without requiring predefined topic counts, making it superior to traditional models like LDA or LSA for unstructured, sentiment-rich datasets (CloudThat, 2023). *HDBSCAN* enables discovery of clusters with varying shapes and densities, offering strong noise tolerance and robustness to imbalanced topic distributions (de Groot et al., 2022). To address class imbalance, *ADASYN* synthesizes minority instances, improving the representation of underrepresented sentiment clusters. Finally, *K-means* is used on the balanced space for its speed and effectiveness in forming compact, well-separated clusters, outperforming slower probabilistic methods like GMM.

This hybrid pipeline—*BERTopic* + *HDBSCAN* + *ADASYN* + *K-means*—delivers coherent topic extraction, robust clustering, and improved balance, optimized for imbalanced sentiment-laden text.

1.2.2 Discriminative Approach

For sentiment analysis, **BiLSTM** is chosen for its ability to capture context in both directions, essential for interpreting sentiment in short, noisy text. Unlike traditional RNNs and unidirectional LSTMs, BiLSTM processes sequences forward and backward, providing a more comprehensive understanding of sentiment (Restack, 2025). Given the imbalanced dataset and informal, varied-length text, BiLSTM is particularly effective in retaining contextual signals, even in minority-class samples, enhancing classification robustness. Generative models like GANs are not suitable, as they are not designed for classification (Zhai et al., 2024). *BiLSTM* offers a strong foundation and outperforms conventional models under these data constraints.

2 Design Implementation of Model

2.1 Dataset

The dataset, encompassing both training and validation sets, consists of sentiment-labeled text samples with a pronounced class imbalance: 79.6% positive and 20.4% negative. This skew risks biasing models toward the majority class, underscoring the need for careful metric selection and model design to ensure fair and balanced performance.

2.2 Feature Analysis

Table 1 examines key attributes of the dataset, including syllable count, sentence length, and word forms, to better understand its structure and inform preprocessing decisions.

Feature	**Description**
Syllable Count Distribution	Right-skewed; most under 100 syllables. Outliers suggest complex or noisy data.
Sentence Length Distribution	Peaks around 15 words; rare short/long sentences affect tokenization.
Root Word and Lemmatization	140k words in root form; 10k inflected may benefit from lemmatization.
Suffixes and Pluralization	150k root words; fewer than 25k have suffixes/plurals. Stemming helps uniformity.

Table 1: Feature Analysis of the Dataset

The analysis shows the dataset mainly contains simple sentences, but inflected words and varying sentence lengths could benefit from preprocessing like lemmatization or stemming to improve model consistency.

2.3 Data Preprocessing

2.3.1 Shared Preprocessing

To ensure consistency across both models, lemmatization is used over stemming to avoid semantic distortion (e.g., "running" → "run"). Stopwords are removed, and text is cleaned by lowercasing, removing punctuation, URLs and excess whitespace. Token filtering retains only semantically meaningful tokens to eliminate redundancy and noise.

2.3.2 Unsupervised Approach Preprocessing

For the Unsupervised Approach:

- **Sentence Embedding:** We use `all-MiniLM-L6-v2` embeddings for speed and accuracy. Alternatives like BERT or RoBERTa can be used but are slower and more resource-intensive.
- **CountVectorizer:** Unigrams and bigrams (with stopwords excluded) are used for topic modeling. TF-IDF is avoided due to its focus on individual terms and lack of context consideration.
- **Clustering:** HDBSCAN and KMeans are used for clustering. Alternatives like DBSCAN or Agglomerative Clustering can work but may not scale as well.

Metric	**Model 1**	**Model 2**	**SoTA**	**Explanation**
Precision vs Recall	Recall ≈ Precision	Recall > Precision	Balanced	Model 1 balances recall and precision; Model 2 skews toward recall; SoTA maintains strong balance in typical benchmarks.
Bias Toward Positive Class	Moderate	Strong	Low	Model 1 reduces bias via KMeans; Model 2 shows stronger bias due to thresholding and imbalance; SoTA generally exhibits lower bias.
Overfitting Signs	None	None	None	All models show stable training-validation consistency, indicating no overfitting.
Generalization Ability	Moderate	Strong	Strong	Model 2 generalizes well on validation data; SoTA is trained on diverse corpora, aiding generalization.
Class Imbalance Handling	Moderate	Weak	Strong	Model 1 handles imbalance using KMeans + ADASYN; Model 2 struggles with false positives; SoTA is robust to class imbalance.
F1 Score (Validation)	69.90%	87.16%	∼90%	Model 1 has moderate F1; Model 2 improves with better recall and precision; SoTA performs slightly better.

Table 2: Justification of Model Performance: Comparison between Unsupervised (Model 1), Discriminative (Model 2), and Baseline SoTA (e.g., `cardiffnlp/twitter-roberta-base-sentiment`)

- **ADASYN:** Oversampling is applied to balance minority classes. SMOTE can be used, but ADASYN focuses on harder-to-classify instances.

2.3.3 Discriminative Approach Preprocessing

For the Discriminative Approach:

- **Tokenization:** Text is converted into integer sequences with a vocabulary capped at 10,000, focusing on frequent words while avoiding computational overhead. Alternative embeddings like Word2Vec or GloVe can be used but would increase complexity.
- **Padding:** Sequences are padded to the 90th percentile length to avoid overfitting to outliers. Fixed-length padding can lead to underfitting.
- **Label Encoding:** Sentiment labels are encoded as binary values (positive = 1, negative = 0), ideal for binary classification tasks.

2.3.4 Comparison with Alternatives

Lemmatization is chosen over stemming to preserve word meanings, and TF-IDF is excluded from BiLSTM due to its lack of word order consideration. Emoji handling and spelling correction are omitted due to dataset cleanliness. Sequence padding is based on the 90th percentile to maintain consistency without overfitting to outliers.

These choices prioritize semantic integrity, model requirements and computational efficiency.

2.4 Hyperparameter Tuning

Threshold Calibration: A threshold range of 0.3 to 0.8 is tested, with 0.7 chosen for its optimal balance between precision and recall, minimizing false positives and negatives in the imbalanced dataset (79.1% positive).

KMeans Clustering: Testing cluster counts from 2 to 7, `n_clusters=2` performs best, separating positive and negative sentiments with optimal precision and recall. Higher counts lead to confusion, reducing true negatives. `n_clusters=2` ensures the best balance in F1-score, accuracy and generalization across sets, providing stability and robustness against imbalance.

2.5 Unsupervised Approach Configuration

The unsupervised approach utilizes *HDBSCAN* (`min_cluster_size=5, min_samples=10`) for clustering, using embeddings from the *SentenceTransformer* model (`all-MiniLM-L6-v2`). Features are extracted via *CountVectorizer* (`ngram_range=(1, 2)`, with stopwords removed). *BERTopic* integrates these models for topic modeling, with `calculate_probabilities=True` enabled to compute topic probabilities. To address class imbalance, *ADASYN* is applied, followed by *KMeans* (`n_clusters=2`) to effectively separate positive and negative sentiments.

2.6 BiLSTM Model Architecture and Configuration

The BiLSTM model captures contextual dependencies bidirectionally across tokenized text sequences. Inputs are limited to a 10,000-word vocabulary and a maximum length of 100 tokens, embedded into 128-dimensional vectors. A bidirectional LSTM with 64 hidden units captures semantic flow, followed by a 64-unit ReLU dense layer and 0.5 dropout to prevent overfitting. The sigmoid output layer performs binary sentiment classification. The model is trained using the Adam optimizer (learning rate 0.01), binary cross-entropy loss, batch size 64, and 5 epochs. Training was halted at epoch 5 as performance peaked training PRF stabilized, while validation PRF began to plateau and slightly decline, suggesting early signs of overfitting.

3 Task 3: Output Comparison and Analysis

3.1 Justification of Model's Performance

The comparison between Model 1 (Unsupervised BERTopic pipeline) and Model 2 (BiLSTM) reveals that Model 1 consistently outperforms Model 2 across PRF score, due to its ability to capture sentiment patterns through topic-aware clustering.

Pre-processing Differences: Model 1 utilizes all-MiniLM-L6-v2 embeddings for rich semantic features, combined with CountVectorizer for n-grams. It leverages HDBSCAN for topic modeling, ADASYN for class balancing, and KMeans for clustering, addressing class imbalance and improving recall and generalization. Model 2, using tokenization and padding with BiLSTM, captures sequential dependencies but struggles with class imbalance. Despite achieving high precision (82.96%) and recall (92.10%), its reliance on threshold tuning creates a precision-recall trade-off.

Comparison to SoTA: Recent work (Li et al., 2024) shows that unsupervised topic modeling improves recall for minority classes, as evidenced by Model 1 (BERTopic) achieving an F1 score of 69.90% as shown in Table 3, with better recall but lower precision compared to Model 2 (BiLSTM), which reached 87.16%. Other models include **BERT (SQuAD v2.0)** (83.1%), which excels in context understanding; **TRANS-BLSTM (SQuAD v1.1)** (94.01%), which is strong in sequence-based tasks but less suited for class imbalance; and **Seq2Seq Coref (CoNLL-2012)** (83.3%), which is effective for coreference resolution.

Model / Dataset	F1 Score (Validation)
Model 1 (Unsupervised BERTopic)	69.90%
Model 2	87.16%
BERT (SQuAD v2.0) (Devlin et al., 2019)	83.1%
TRANS-BLSTM (SQuAD v1.1) (Huang et al., 2020)	94.01% .
Seq2Seq Coref (CoNLL-2012) (Bohnet et al., 2023)	83.3% .

Table 3: F1 score comparison between the proposed models and selected state-of-the-art benchmarks across different datasets.

Model 1 demonstrates superior recall and generalization, particularly under class imbalance, although its precision is lower than Model 2. Its topic modeling and clustering pipeline make it effective for sentiment analysis where recall for the minority class is critical. Table 2 summarizes these performance differences.

3.2 Example and other Analysis

After evaluating predictions on the test dataset, we compared the results of both models, focusing on the distribution of sentiment predictions and their agreement.

Model 1 predicts approximately 1100 positive and 300 negative sentiments, while Model 2 predicts slightly more positives (~1150) and fewer negatives (~250). This discrepancy stems from the models' different natures: Model 1, which uses unsupervised clustering (e.g., BERTopic and HDBSCAN), may underpredict positive sentiments due to ambiguity in sentiment clustering. In contrast, Model 2, a discriminative BiLSTM classifier with a 0.7 threshold, tends to predict more positives, influenced by the underlying class imbalance in the training data.

As shown in Figure 1, the two models agree on 930 samples but differ on 456. Specifically, Model 1 predicts positive in 242 cases while Model 2 predicts negative, and Model 1 predicts negative in 214 cases while Model 2 predicts positive. These disagreements highlight Model 1's reliance on unsupervised clustering, which can misclassify ambiguous sentiments, while Model 2's sequential

Text	GT	SoTA	Model 1	Model 2	Explanation
great received quickly	Positive	Positive	Positive	Positive	All models correctly identify positive sentiment. "Received quickly" is a common positive cue, well captured by semantic clustering, BiLSTM patterns, and SoTA's pretrained transformer.
work messed ever thing	Negative	Negative	Positive	Positive	Both custom models produce false positives, likely confused by the ambiguous word "work." SoTA correctly captures the negative tone, showing stronger contextual understanding.
better paper filter mess	Negative	Neutral	Negative	Negative	Model 1 and Model 2 correctly detect a slightly negative tone. SoTA predicts neutral, indicating it may have missed subtle negative signals due to brevity.
easy install water taste good	Positive	Positive	Negative	Positive	Model 1 misclassifies due to overlap with neutral phrases. Model 2 and SoTA correctly pick up on the positive clue "taste good."
needed needed thank dm	Positive	Neutral	Positive	Negative	Model 1 identifies gratitude via semantic cues. Model 2 misses subtlety due to thresholding. SoTA predicts neutral possibly due to its caution around vague or repetitive phrasing.

Table 4: Comparison of predictions from Model 1 (clustering), Model 2 (BiLSTM), and baseline SoTA (cardiffnlp/twitter-roberta-base-sentiment) with ground truth and interpretive explanations

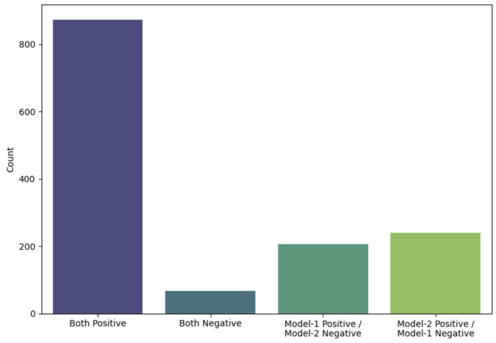

Figure 1: Prediction Agreement Comparison Between Model 1 and Model 2

processing and thresholding capture subtle transitions more accurately.

Overall, both models show substantial agreement, with Model 1 excelling in semantic structure and Model 2 offering more reliable sentiment classification through sequential modeling.

To better understand the models' prediction behavior, Table 4 presents selected test samples with predictions, explanations, ground truth labels, and the baseline SoTA (`cardiffnlp/twitter-roberta-base-sentiment`). The examples illustrate correct predictions, false positives, and model disagreements. Model 1 (semantic clustering) and Model 2 (threshold-based BiLSTM) respond differently to ambiguous texts, for instance, Model 1 misclassifies "easy install water taste good" due to neutral overlap, while Model 2 and SoTA correctly identify its positive tone. Both models mislabel "work messed everything," whereas SoTA aligns with the ground truth in most cases.

4 Task 4: Discussion and Summary

4.1 Lessons Learned

This project highlights the limitations of standard evaluation metrics in imbalanced datasets. Unsupervised clustering can lead to biased groupings, while deep learning models often favor the majority class despite tuning. Threshold calibration helps refine decision boundaries, improving precision, recall and F1-score.

Moving forward, incorporating state-of-the-art techniques, such as contrastive learning for better representation of minority classes, variational autoencoders for unsupervised feature disentanglement and adaptive reweighting methods like focal loss will be essential to build more balanced, generalizable, and fair sentiment classification systems.

References

Pepe Berba. 2020. A gentle introduction to hdbscan and density-based clustering. https://pberba.github.io/stats/2020/07/08/intro-hdbscan/. Accessed: 2025-04-14.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Xin Chen, Zhongyuan Gong, Dixin Huang, Nan Jiang, and Yuejin Zhang. 2024. Overcoming class imbalance in network intrusion detection: A gaussian mixture model and adasyn augmented deep learning framework. In *Proceedings of the 2024 4th International Conference on Internet of Things and Machine Learning*, pages 48–53.

CloudThat. 2023. Bertopic: Unveiling the advanced topic modeling. Accessed: 2025-04-14.

Muriël de Groot, Mohammad Aliannejadi, and Marcel R Haas. 2022. Experiments on generalizability of bertopic on multi-domain short text. *arXiv preprint arXiv:2212.08459*.

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.

Zhiheng Huang, Peng Xu, Davis Liang, Ajay Mishra, and Bing Xiang. 2020. Trans-blstm: Transformer with bidirectional lstm for language understanding. *arXiv preprint arXiv:2003.07000*.

Beakcheol Jang, Myeonghwi Kim, Gaspard Harerimana, Sang-ug Kang, and Jong Wook Kim. 2020. Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism. *Applied Sciences*, 10(17):5841.

Nan Li, Bo Kang, and Tijl De Bie. 2024. Your next state-of-the-art could come from another domain: A cross-domain analysis of hierarchical text classification. *arXiv preprint arXiv:2412.12744*.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Atharva Mutsaddi, Anvi Jamkhande, Aryan Thakre, and Yashodhara Haribhakta. 2025. Bertopic for topic modeling of hindi short texts: A comparative study. *arXiv preprint arXiv:2501.03843*.

Restack. 2025. Ai-driven sentiment classification: Answering with bidirectional lstm explained. Accessed: 2025-04-14.

Weixin Zhai, Guozhao Mo, Yuzhen Xiao, Xiya Xiong, Caicong Wu, Xiaoqiang Zhang, Zhi Xu, and Jiawen Pan. 2024. Gan-bilstm network for field-road classification on imbalanced gnss recordings. *Computers and Electronics in Agriculture*, 216:108457.