informatics

MDPI

Article

Which Model Feels Better? A Comparison of Computational Approaches to Emotion Detection in Social Media with Imbalanced Data

Anastasia Vishnevskaya [1,*], Vasileios Pavlopoulos [2,*], Jerry Stott [3] and Porismita Borah [4]

[1] College of Media & Communication, Texas Tech University, Lubbock, TX 79409, USA
[2] Department of Information Systems, Supply Chain, and Analytics, University of Alabama in Huntsville, Huntsville, AL 35899, USA
[3] Department of Political Science, University of Utah, Salt Lake City, UT 84112, USA
[4] Edward R. Murrow College of Communication, Washington State University, Pullman, WA 99164, USA
* Correspondence: anastasia.vishnevskaya@ttu.edu (A.V.); vasileios.pavlopoulos@uah.edu (V.P.)

Abstract

Emotion detection in social media remains challenging, particularly in polarized public debates where emotional expression is often imbalanced across categories. Addressing this challenge, this study compares the performance of multiple computational approaches using a gold-standard dataset of tweets about an ongoing geopolitical conflict. The dataset reflects the authentic, skewed distribution of emotions observed in real-world online discourse. We evaluated lexicon-based methods, classical machine-learning classifiers, deep-learning architectures, transformer models in both fine-tuned and zero-shot configurations, and a zero-shot large language model to assess their effectiveness in capturing both frequent and less frequently expressed emotions. Across approaches, transformer models, especially those fine-tuned for contextual emotion recognition, demonstrated the strongest overall performance, with emotion-specific fine-tuning offering a particular advantage for detecting rare emotion categories. These findings emphasize the importance of evaluating emotion detection methods under realistic class imbalance and highlight both the comparative strengths and limitations of widely used modeling strategies in applied social media research. This study advances emotion analysis and computational social science by offering practical guidance for selecting appropriate emotion detection methods in complex, imbalanced social media contexts.

Keywords: emotion analysis; computational methods; machine-learning; transformers; BERT

Academic Editors: Patricia Anthony and Jing Zhou

Received: 29 January 2026
Revised: 2 May 2026
Accepted: 4 May 2026
Published: 22 May 2026

1. Introduction

Emotions play an important role in political communication. For example, studies show that affective responses significantly impact political engagement (e.g., [1,2]), participation (e.g., [3–6]), perceptions of political issues (e.g., [7]), and behaviors toward candidates and parties (e.g., [8,9]). Research also shows that emotional responses significantly influence people's attitude formation (e.g., [10]), information sharing (e.g., [11,12]), and collective actions in politically charged contexts (e.g., [13–15]).

With the growth of social media as a major forum for online political discourse (e.g., [16,17]), understanding emotional dynamics has become increasingly important for communication scholars. In these online spaces, emotions shape the tone and direction of

political discussion. Research consistently shows that discrete emotional appeals significantly influence user engagement on Facebook (e.g., [18,19]), Twitter/X (e.g., [20]), and Instagram (e.g., [21]). Emotionally charged messages on social media tend to be shared more frequently and rapidly than neutral ones (e.g., [20,22]). These emotional dynamics in online spaces can also shape public perceptions and values and reinforce or challenge existing social and political institutions [23]. Studies such as these highlight the importance and the growing need for discrete emotion analysis (e.g., [24]) rather than the limited identification of positive or negative valence within sentiment analysis (e.g., [25–29]).

Sentiment analysis, or a two-dimensional approach to emotion analysis (e.g., [18]), has been widely used in communication scholarship (e.g., [30–32]). However, in the social sciences, sentiment analysis lacks a consistent conceptualization and operationalization [33]. In communication research, the terms *emotions* and *sentiments* are often used interchangeably because both refer to affective responses, feelings, and evaluations. This usage, however, can be misleading. Although emotions and sentiments are closely related forms of affective experience, they are meaningfully distinct [34,35].

Emotions are intense, short-lived, and directly experienced affective states, such as *anger*, *joy*, *fear*, or *sadness* [34]. They are often characterized by distinct physiological responses and behavioral tendencies associated with particular stimuli [36,37]. Emotions can also be classified into a limited set of categories, drawing on the concept of "basic" emotions. For example, Ekman (1992) identified six basic emotional categories: *happiness*, *surprise*, *fear*, *sadness*, *anger*, and *disgust* [38]. Sentiments, on the contrary, are longer-lasting and more persistent attitudes or opinions that individuals hold toward objects, people, or ideas [35].

From a methodological standpoint, sentiments are commonly classified into positive, negative, and neutral categories, reflecting the overall opinion expressed in a text or a message rather than the specific emotions that may be present (e.g., [32,39,40]). Although sentiment analysis is widely applied in communication studies (e.g., [33]), it is limited in its ability to capture distinctions among discrete emotions that have unique psychological properties and communication effects (e.g., [34,35]). As a result, greater attention to emotion analysis is needed, particularly in studies examining how specific emotional responses shape communication processes and outcomes.

Identifying discrete emotions in text is a complex task (e.g., [37]). For instance, related work in adjacent natural language processing (NLP) domains, such as systematic reviews of hate speech detection, highlights the growing reliance on transformer-based (a class of deep-learning architectures) and large language models (LLMs) for classifying emotionally charged text, while also pointing to persistent challenges related to linguistic bias, ambiguity, and computational cost [41]. Other studies emphasize that emotion detection is not only technically challenging but also subjective (e.g., [42]), which can contribute to low inter-coder agreement in manual coding approaches (e.g., [43]). At the same time, as with sentiment analysis (e.g., [33]), a wide range of computational models is available for emotion analysis (e.g., [44–47]).

These models allow researchers to examine large-scale conversations and their societal implications across a range of contexts, including social media communication (e.g., [48,49]), interpersonal communication (e.g., [50,51]), and health communication (e.g., [52]), among others. Yet questions remain about the reliability of these tools for emotion analysis. Existing studies frequently rely on individual models (e.g., [18,20,53]), making it difficult to systematically compare and evaluate their performance. Even when multiple computational emotion analysis models are used, evaluations of their relative performance are often limited or absent (e.g., [54]).

To address this gap, our study provides a comprehensive methodological evaluation of computational emotion analysis models in the context of conflict-related social media discourse. We compared several categories of computational emotion detection techniques: (1) dictionary-based methods, (2) classical machine-learning algorithms, (3) deep-learning models, and (4) transformer-based models, including both fine-tuned classifiers and a zero-shot instruction-tuned model representing the prompting paradigm associated with large language models. This comparative analysis allowed us to systematically assess the performance, strengths, and limitations of each computational approach against validated human-coded emotion annotations.

Our study makes two primary contributions to communication scholarship. First, it emphasizes the theoretical need to differentiate discrete emotions from broad sentiment classifications. Second, rather than proposing a new classification framework or multi-step pipeline, this study provides a systematic comparative evaluation of existing computational approaches. In doing so, it identifies the strengths and limitations of diverse methods when applied to real-world, emotionally charged social media data and offers practical guidance for interdisciplinary communication research.

2. Data and Methods

2.1. Data

We used a secondary dataset of English-language Twitter/X posts about an ongoing international geopolitical conflict to evaluate these models. The data were collected in April 2022 and made available through the Institute of Electrical and Electronics Engineers (IEEE) DataPort. According to the dataset description, it is suitable for sentiment analysis and other NLP-related research. Importantly, this dataset is not only relevant because of its format and availability, but also because of the emotionally charged context in which the posts were produced.

Early April 2022 was marked by major conflict-related events that triggered intense public reactions. Reports of civilian atrocities during this period generated widespread shock and outrage [55]. Twitter/X discourse during this period also reflected a wide range of emotions, including anger, fear, sadness, disgust, surprise, and occasional joy (e.g., [56]). Because Twitter/X is commonly used by various actors to express real-time opinions during conflicts (e.g., [57]), the posts in this dataset provide an emotionally complex, highly expressive, and often contested context for evaluating computational emotion analysis models.

For our analysis, we selected the subset of tweets in English. To protect user privacy, we removed all personally identifiable information from the dataset before any analysis or annotations were made. To maintain ecological validity and assess model performance under realistic conditions, we retained the full range of identified emotions and their naturally imbalanced distribution, rather than applying multi-step classification strategies such as pre-filtering the neutral class. Although such approaches are common in sentiment analysis to improve classification balance [58], our focus on discrete, multi-class emotion detection required evaluating models on authentically skewed data. This approach allowed us to capture the challenges of detecting nuanced emotions in real-world social media discourse.

2.2. Gold Standard

A "gold standard" refers to a human-coded dataset that serves as a benchmark for evaluating automated text classification methods [59]. Such datasets are typically developed through manual content analysis, in which trained coders classify texts into predefined categories. Once acceptable intercoder reliability is established, the resulting dataset

provides a reference baseline for evaluating automated tools by measuring the extent to which their outputs align with human-coded labels.

Establishing a gold standard is, thus, essential for evaluating the reliability, precision, and overall performance of computational approaches (e.g., [60]). For the purpose of this study, we created such a standard by developing a codebook, recruiting and training coders, and assessing intercoder reliability.

First, we developed our emotion classification codebook based on the comprehensive taxonomy introduced in the GoEmotions dataset [61]. This taxonomy includes 27 fine-grained emotion categories in addition to one *neutral* category and has been extensively validated for capturing complex emotional expressions in text. We adapted the GoEmotions codebook to fit the context of the present study, resulting in a final coding scheme of 25 fine-grained emotion categories plus one *neutral* category. The recruited coders were then trained to identify emotions according to the developed codebook.

2.3. Intercoder Reliability

We recruited two coders with advanced-level graduate degrees, both native English speakers. The two coders completed four rounds of coding, each using randomly selected subsets of tweets: 300, 300, 400, and 250 tweets, respectively. The coders were instructed to evaluate the tweets based on the original authors' emotional expression and to avoid relying on their own political knowledge or contextual assumptions when assigning emotion labels. Each round was followed by detailed discussions to resolve disagreements between coders.

Initial intercoder reliability, measured using Cohen's κ [62], was substantial, ranging from 0.72 to 0.80 across coding rounds. After coder discussions and disagreement resolution, Cohen's κ exceeded 0.90. The finalized labels served as the gold standard against which the computational emotion classification models were evaluated.

2.4. Data Preparation

The initial manually annotated dataset consisted of 1250 tweets, each labeled with one of 26 categories (25 plus one *Neutral*). Python v. 3.9 was used for data analysis.

Given the emotionally charged and complex nature of online discourse surrounding the ongoing geopolitical conflict, we focused our analysis on content directly related to the conflict. As part of the data cleaning process, we systematically excluded tweets coded as *Irrelevant*, that is, tweets that did not pertain to the geopolitical conflict under study (n = 245). This step allowed us to improve analytical clarity and reduce noise in the dataset.

The remaining emotion labels were then grouped into seven coarse-grained categories: *Joy/Happiness*, *Sadness*, *Anger*, *Fear*, *Surprise*, *Disgust*, and *Neutral.* We aligned this categorization with Ekman's taxonomy of basic emotions [38,61] by collapsing the fine-grained categories into six basic emotions: *Anger*, *Disgust*, *Fear*, *Joy*, *Sadness*, and *Surprise*. The *Neutral* category was retained as a non-emotional baseline category. The rationale for restructuring our data was twofold.

First, reducing the data to six emotion categories offers both theoretical and practical alignment. Ekman's [38] emotion classification is validated and widely used across disciplines (e.g., [20]). Employing this established framework enhances the comparability of our findings with existing studies and increases the broader applicability of our results. Second, fine-grained emotion categories naturally aggregate into broader categories through hierarchical clustering analysis [61]. For instance, emotions such as *Anger*, *Annoyance*, and *Disapproval* consistently cluster together, providing both conceptual and empirical support for collapsing these categories. Based on these justifications, our final emotion groupings are presented in Table 1.

It is important to note that, after restructuring the data, the distribution of emotion labels remained highly imbalanced. Certain categories, such as *Neutral*, dominated the dataset, whereas others, such as *Surprise*, were rare. This imbalance poses technical challenges for classification because models may overfit to dominant categories and generalize poorly to minority emotions. We therefore retained the natural class distribution to evaluate model performance under realistic conditions.

Table 1. Mapping of Fine-Grained Emotions to Coarse-Grained Emotion Categories.

Coarse-Grained Category	Mapped Fine-Grained Emotions
Joy/Happiness	Admiration, Amusement, Caring, Desire, Excitement, Gratitude, Joy, Love, Optimism, Pride, Relief
Sadness	Sadness, Disappointment, Embarrassment, Grief, Remorse, Guilt, Shame
Anger	Anger, Annoyance
Fear	Fear, Nervousness
Surprise	Surprise, Realization
Disgust	Disgust
Neutral	Neutral

To analyze the emotional content of the tweets, we employed a range of computational approaches: (i) lexicon-based models (i.e., NRC and Empath), which rely on pre-compiled dictionaries that map words to specific emotions; (ii) classical machine-learning classifiers, including Naive Bayes (NB) and Support Vector Machine (SVM), which use such features as term frequency representations to classify emotions; (iii) deep-learning models, specifically Convolutional Neural Networks (CNNs) trained on word embeddings (i.e., GloVe), to capture more complex linguistic patterns; and (iv) transformer-based architectures, including BERT, RoBERTa, GoEmotions, DistilRoBERTa-Emotion, and DeBERTa-v3-base, which represent the state-of-the-art in computational text analysis. We additionally included Flan-T5-base, an instruction-tuned language model applied in a zero-shot prompting setting, to assess how this paradigm performs relative to fine-tuned classifiers on short, emotionally charged social media text.

To ensure an unbiased final evaluation, we adopted a differentiated assessment strategy tailored to each model family. For models involving training and checkpoint selection (CNN baseline, GloVe CNN, BERT, RoBERTa, Twitter-RoBERTa, and GoEmotions Fine-Tuned), we employed a three-way stratified split of the annotated dataset: 70% for training ($n = 703$), 15% for validation ($n = 151$), and 15% for final testing ($n = 151$). The validation set was used exclusively to monitor early stopping and select the best model checkpoint. Final performance metrics were computed on the held-out test set, which was not accessed during any stage of model training or selection. This design prevents the optimistic bias that can arise when checkpoint selection and final evaluation are performed on the same data partition [63]. Classical machine learning models (Naive Bayes and SVM) were evaluated using stratified 5-fold cross-validation applied to the full dataset ($n = 1005$), which is standard practice for models without checkpoint selection. Lexicon-based methods (NRC and Empath) and the GoEmotions zero-shot model required no training; NRC and Empath were evaluated on the same held-out test set ($n = 151$) as the trained models to ensure consistency, while GoEmotions zero-shot was applied to the full dataset. It should be noted that the relatively small size of the held-out test partition ($n = 151$) means that per-class metrics for rare emotion categories such as *Surprise* ($n = 2$ test instances) and *Disgust* ($n = 8$) should be interpreted with caution, as single misclassifications can produce large metric swings for these classes.

Prior to discussing how each category of models was developed and validated, we also clarified the evaluation metrics used to assess their performance.

2.5. Evaluation Metrics

To evaluate the performance of our emotion classification models, we used a set of standard metrics commonly applied in natural language processing. Each metric offers a different perspective on model effectiveness, particularly in multiclass settings where class imbalance is prevalent. In line with existing studies [33,64], we focused on *Precision*, *Recall*, *F1-score*, *Accuracy*, *Cohen's* κ, and *Macro-* and *Weighted-Average* scores. These metrics allowed us to assess both overall model performance and class-specific classification patterns in automated emotion detection. We discuss each metric below.

2.5.1. Precision

Precision measures the proportion of correctly predicted instances among all instances that were labeled as a given class by the model. It is formally defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

where TP is the number of true positives (correctly predicted instances of a class), and FP is the number of false positives (instances incorrectly predicted as that class). In emotion detection, this reflects the reliability of the model when it assigns a particular emotional label to a tweet. A high Precision value indicates that the model rarely makes false-positive errors. Precision is especially important when the cost of false positives is high, such as misclassifying neutral content as emotionally charged. Lexicon-based approaches often show high Precision but may sacrifice Recall [33].

2.5.2. Recall

Recall, also known as sensitivity, is the proportion of true positive predictions out of all actual instances of a given class. It is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

where FN is the number of false negatives (instances of a class that were not correctly predicted). It answers the question: "Of all the tweets that truly belong to an emotion class, how many did the model correctly identify?" This metric is especially valuable when it is important not to miss relevant emotional instances. Models with high Recall are better at covering the emotional spectrum, and low Recall often results from dictionary-based methods that fail to capture nuanced or less frequently occurring terms.

2.5.3. F_1-Score

The F_1-score is the harmonic mean of Precision and Recall, offering a balanced assessment of a classifier's performance:

$$F_1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

It is particularly useful when there is an uneven class distribution, which is common in real-world emotion datasets. The F_1-score provides a single number that balances both false positives and false negatives, and it is used as a comprehensive metric, especially when evaluating models that show trade-offs between Precision and Recall.

2.5.4. Accuracy

Accuracy refers to the overall proportion of correct predictions across all classes:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TN is the number of true negatives. While Accuracy is easy to interpret, it can be misleading when classes are imbalanced. A model may achieve high overall Accuracy by correctly predicting the majority class while performing poorly on less frequent emotion categories. Therefore, model performance in emotion classification tasks should be evaluated using a combination of Precision, Recall, and F_1-scores.

2.5.5. Cohen's κ

Cohen's κ is a statistical measure that evaluates the degree of agreement between predicted labels and ground truth labels while adjusting for the level of agreement that could occur by random chance [62]. It is calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed agreement, and p_e is the expected agreement by chance. Unlike raw accuracy, which simply computes the proportion of correct predictions, κ provides a more nuanced view by penalizing coincidental matches. The metric ranges from -1 to 1, where a score of 1 indicates perfect agreement, 0 denotes agreement equivalent to random chance, and negative values suggest systematic disagreement between the predictions and actual labels. This makes Cohen's κ especially valuable in contexts with imbalanced class distributions or subjective labeling criteria.

2.5.6. Macro Average

The Macro Average computes the arithmetic mean of the metric (e.g., F_1-score) independently for each class, treating all classes equally regardless of their frequency. It is calculated as:

$$\text{Macro-}F_1 = \frac{1}{C} \sum_{i=1}^{C} F_{1i}$$

where C is the number of emotion classes. This makes it suitable for imbalanced datasets where some emotion categories are underrepresented. We recommend using Macro Average to make sure that model performance is fairly assessed across all categories and to detect systematic model weaknesses in rarer emotion classes.

2.5.7. Weighted Average

The Weighted Average also computes the metric for each class but weights it by the number of true instances in each class (i.e., support). It is defined as:

$$\text{Weighted-}F_1 = \frac{\sum_{i=1}^{C} s_i \cdot F_{1i}}{\sum_{i=1}^{C} s_i}$$

where s_i denotes the number of true instances (support) for class i. This provides a more realistic picture of overall performance by accounting for class imbalance. According to the literature (e.g., [64]), Weighted Averages tend to be more aligned with Accuracy in highly imbalanced datasets, since they reflect performance on the dominant categories. We believe that Weighted Averages often mask poor performance in minority classes; thus, there is a need to consider both macro and weighted metrics.

2.6. Lexicon-Based Emotion Detection

Despite their known limitations, lexicon-based approaches in this study serve as a crucial baseline, allowing us to highlight the relative advancements achieved by classical machine-learning, deep-learning, and transformer-based models in the complex task of emotion analysis.

2.6.1. NRC Lexicon

The NRC Emotion Lexicon [65] is a widely used lexicon-based resource for emotion and sentiment analysis. It links English words to basic emotions based on crowd-sourced associations. The lexicon includes eight fundamental emotions (*Anger*, *Fear*, *Anticipation*, *Trust*, *Surprise*, *Sadness*, *Joy*, and *Disgust*), which can be flexibly mapped to broader emotional taxonomies. Unlike supervised or neural models, the NRC Lexicon does not require labeled training data or optimization procedures. Instead, it relies on the presence of emotionally charged words in the input text to infer sentiment or emotion categories.

To apply the NRC Lexicon, we used the *nrclex* Python library, which provides access to emotion associations for each token in a given text. Tweets were first cleaned by converting them to lowercase, removing URLs, mentions, hashtags, and non-alphabetic characters. We then lemmatized each word using NLTK's *WordNetLemmatizer* to improve alignment with the lexicon entries. The dominant emotion in each tweet was identified using the raw emotion scores returned by the lexicon and mapped to one of our seven coarse-grained categories. This allowed us to produce one predicted label per tweet and evaluate the results against the ground truth (*gold standard*).

2.6.2. Empath

Empath [66] is a lexicon-based tool that allows for large-scale text analysis across a broad range of predefined categories, including several psychological and emotional themes. Unlike traditional sentiment lexicons, Empath uses vector space representations learned from a large corpus to expand and validate its categories, allowing it to detect nuanced emotional expressions. While not originally developed exclusively for emotion classification, Empath includes categories such as *Joy*, *Sadness*, *Anger*, *Fear*, *Disgust*, and *Surprise*, which makes it suitable for aligning with common affective taxonomies.

To apply Empath to our tweets, we used the *empath* Python package. Tweets were first preprocessed by converting them to lowercase, removing special characters, and applying lemmatization using NLTK's *WordNetLemmatizer* [67] to ensure better alignment with Empath's vocabulary. We then extracted scores for each of the six emotional categories and assigned to each tweet the category with the highest normalized score. If no category received a nonzero score, the tweet was labeled as *Neutral*. This method required no training or fine-tuning, making it directly applicable in a zero-shot fashion.

2.7. Traditional Machine-Learning Models

2.7.1. Naive Bayes

Naive Bayes [68] is a probabilistic machine-learning method based on Bayes' theorem with the assumption of feature independence. It is widely used in text classification tasks due to its simplicity, scalability, and strong performance with high-dimensional data. Specifically, the Multinomial Naive Bayes variant is particularly effective for discrete features such as word counts or term frequencies. Despite its simplifying assumptions, Naive Bayes often performs competitively, especially on small- to medium-sized text corpora [69].

In our implementation, we used the *MultinomialNB* [70] classifier from *scikit-learn*, coupled with a *TfidfVectorizer* that transforms each tweet into a weighted bag-of-words

representation using unigrams and bigrams (i.e., 1–2 word sequences). To evaluate the model's performance, we employed stratified 5-fold cross-validation to ensure balanced label representation in each fold. The pipeline was fit on the training splits and evaluated on the validation folds, with performance metrics aggregated across folds. This approach allowed us to assess the generalization ability of the Naive Bayes classifier without requiring a separate test set or any additional tuning.

2.7.2. Support Vector Machine (SVM)

Support Vector Machines are supervised learning algorithms [71] that aim to identify an optimal hyperplane that separates classes with the maximum possible margin. They are particularly effective in high-dimensional spaces, making them well-suited for text classification tasks where each feature corresponds to a term or *n*-gram. The Linear SVM variant, which uses a linear kernel, is commonly adopted for its computational efficiency and strong generalization performance, especially in sparse text settings.

In this study, we used a *LinearSVC* model from *scikit-learn* with balanced class weights to account for the uneven distribution of emotion categories. Tweets were vectorized using a *TfidfVectorizer* [70] with unigrams and bigrams (1–2 word sequences) and a maximum of 5000 features. To evaluate the model robustly, we performed stratified 5-fold cross-validation, ensuring that each fold maintained the overall class distribution. For each split, the pipeline was trained on the training data and tested on the validation fold, and the final results were aggregated across all folds.

2.8. Deep-Learning Models

2.8.1. Baseline CNN

Convolutional Neural Networks (CNNs), though originally developed for image data, have shown effectiveness in text classification tasks due to their ability to detect local patterns such as *n*-grams. CNNs apply convolutional filters to word embeddings, capturing discriminative phrases indicative of emotional content. In emotion classification, CNNs offer an efficient deep-learning alternative to more computationally expensive transformer models, especially when used with fixed-length sequences and relatively small datasets.

For our baseline model, we implemented a multi-filter CNN using *Keras* [72]. Tweets were cleaned and tokenized using a *Tokenizer*, converted into padded sequences of maximum length 50, and embedded via a trainable embedding layer of dimension 100. The model used three separate convolutional layers with kernel sizes 3, 4, and 5, followed by global max pooling, concatenation, dropout, and a final softmax classification layer. The model was compiled with categorical cross-entropy loss and trained using the Adam optimizer [73]. Following the evaluation protocol described in Section 2.4, the dataset was partitioned into stratified training, validation, and test sets (70/15/15). The validation set was used exclusively for early stopping, and final metrics were reported on the held-out test set. This baseline does not use pretrained embeddings or class balancing.

2.8.2. GloVe CNN

To enhance the representational power of the baseline CNN, we incorporated pretrained word embeddings from the Global Vectors for Word Representation (GloVe) model. GloVe embeddings provide semantically meaningful vector representations for words, capturing context based on global word co-occurrence statistics. This approach allows the CNN to benefit from prior linguistic knowledge during training, improving its generalization ability, especially with small labeled datasets.

Our rationale for testing the Glove-CNN model was as follows. Although CNN models were not originally designed specifically for emotion recognition, they provide a powerful architecture for learning local features and patterns in text. A baseline CNN

with randomly initialized embeddings learns these patterns from scratch, which can be suboptimal when training data are limited. To overcome this limitation, we integrated GloVe embeddings, pretrained on a large corpus, to provide the model with richer semantic word representations. This allowed the network to begin with linguistically informed word vectors rather than learning representations entirely from the available training data. Incorporating GloVe was therefore a principled step toward improving emotion classification performance without substantially increasing model complexity.

We used the *glove.6B.100d.txt* corpus [74], which is a pretrained file, mapping words from the tokenized tweets to their corresponding 100-dimensional GloVe embeddings. These embeddings were loaded into a non-trainable embedding layer in our CNN architecture. The model retained the multi-filter design from the baseline CNN, using convolutional layers with kernel sizes 3, 4, and 5, followed by global max pooling, concatenation, dropout, and a final softmax layer. Additionally, to address class imbalance, we applied random oversampling [75] to the training set and included class weights during training. Model training used the stratified 70/15/15 train/validation/test split introduced earlier, with the validation set guiding early stopping and the held-out test set used for final evaluation (see Section 2.4).

2.9. Transformers

2.9.1. BERT

Transformer-based models have become state-of-the-art in natural language processing due to their ability to model long-range dependencies and context-sensitive representations [76]. Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained language model that has demonstrated strong performance across various downstream tasks, including sentiment and emotion classification. Its bidirectional attention mechanism enables it to capture nuanced relationships in textual data, making it well-suited for analyzing informal, context-rich content such as tweets.

To adapt BERT [63] for our seven-category emotion classification task, we fine-tuned the *bert-base-uncased* model using the Hugging Face Transformers library [77]. We first tokenized the tweet texts using the associated BERT tokenizer and padded/truncated them to a maximum length of 50 tokens. The dataset was partitioned into stratified training, validation, and test sets (70/15/15) (see Section 2.4). We computed class weights to address class imbalance in the labeled emotions and integrated these weights into the training process by defining a custom loss function using the *CrossEntropyLoss* with weighting. A subclassed *Trainer* was implemented to override the default loss computation. We trained the model for up to 15 epochs with early stopping (patience = 3), monitoring validation accuracy to select the best checkpoint. Final metrics were reported on the held-out test set, which was not accessed during training or model selection.

2.9.2. RoBERTa

Robustly Optimized BERT Approach (RoBERTa) [78] is an advanced transformer model built upon the BERT architecture, designed to enhance performance through extensive pretraining on larger corpora and improved training procedures. Unlike BERT, RoBERTa removes the Next Sentence Prediction (NSP) objective and employs dynamic masking, making it more robust across downstream tasks. Given its enhancements and proven effectiveness in nuanced text classification, RoBERTa offers a competitive alternative for emotion detection in social media data.

To apply RoBERTa in our study, we fine-tuned the *roberta-base* model using the Hugging Face Transformers library [77]. We tokenized tweets using RoBERTa's tokenizer and padded/truncated them to a maximum of 50 tokens. Following the evaluation protocol described in Section 2.4, the dataset was partitioned into stratified training, validation, and

test sets (70/15/15). Class imbalance was addressed by computing label frequencies and incorporating them as weights into the loss function via a subclassed *Trainer*. We trained the model for up to 15 epochs with early stopping (patience = 3), using the validation set exclusively for checkpoint selection. As with previous models, final metrics were reported on the held-out test set, which was not accessed during training or model selection.

2.9.3. Twitter-RoBERTa

Twitter-RoBERTa is a transformer-based language model pre-trained on large volumes of Twitter data [79,80]. Unlike generic language models, it is optimized to better capture the informal, abbreviated, and often noisy language typical of social media. This makes it particularly well-suited for tasks involving tweet classification and sentiment or emotion detection. Given its alignment with our dataset source, we included Twitter-RoBERTa to assess how a domain-specific model performs relative to general-purpose transformers like BERT and RoBERTa.

We fine-tuned the *cardiffnlp/twitter-roberta-base* model using the Hugging Face Transformers library [77]. Tweets were tokenized using the model's tokenizer and padded to a maximum length of 50 tokens. As in previous transformer-based methods, class imbalance was addressed by computing class weights and integrating them into a weighted cross-entropy loss within a custom *Trainer* class. A stratified 70/15/15 split was applied to create training, validation, and test subsets (see Section 2.4). The model was trained for up to 15 epochs with early stopping (patience = 3), using the validation set exclusively for checkpoint selection. Final metrics were also reported on the held-out test set.

2.9.4. GoEmotions, Zero-Shot

GoEmotions is a fine-grained emotion classification dataset encompassing 27 emotion labels plus a *Neutral* class [61]. For this study, we used *monologg/bert-base-cased-goemotions-original*, a BERT-based model trained on the original GoEmotions labels, without additional fine-tuning. This allowed us to evaluate the model's out-of-the-box performance on our manually labeled tweet dataset and assess how well a pre-trained emotion classifier generalizes to conflict-related social media discourse.

To apply the model, we preprocessed the tweets using standard cleaning steps and tokenized them with the GoEmotions tokenizer. The model output, consisting of sigmoid-activated probabilities for each of the 28 classes, was thresholded to identify active emotion labels. These GoEmotions labels were then mapped to our seven coarse-grained emotion categories (see Table 1). When multiple labels exceeded the threshold for a tweet, we used majority voting to select the dominant emotion; otherwise, we selected the highest-scoring label. This setup allowed us to assess the generalizability of GoEmotions in a zero-shot setting, without training it on our specific dataset. As no training or checkpoint selection was involved, this model was applied to the full dataset ($n = 1005$) rather than the held-out test partition, consistent with the evaluation protocol described in Section 2.4.

2.9.5. GoEmotions, Fine-Tuned

Although the off-the-shelf GoEmotions model demonstrated some transferability by leveraging its training on fine-grained emotion categories, its performance may have been constrained by the lack of task-specific adaptation. Therefore, we fine-tuned the *monologg/bert-base-cased-goemotions-original* model directly on our dataset. This allowed the model to update its learned representations and decision boundaries based on the characteristics of the current data and improve alignment between the input distribution and the target categories.

To adapt the GoEmotions model to our classification task, we first replaced its original 28-class output layer with a custom classification head containing seven output neurons

corresponding to our emotion categories (Table 1). This modification was necessary to avoid shape-mismatch errors during training. We used the Hugging Face Transformers library [77] and defined a custom *Trainer* instance that fine-tuned all model parameters on our manually labeled tweets. Again, as outlined in Section 2.4, we split the dataset into stratified training, validation, and test sets (70/15/15). Model performance was monitored using standard metrics such as Accuracy, Macro-F_1 score, and Cohen's κ. The final model was selected based on early stopping (patience = 3). The best validation accuracy and the final metrics were reported on the held-out test set.

2.9.6. DistilRoBERTa-Emotion

DistilRoBERTa-Emotion [81] is a compact transformer model based on the DistilRoBERTa architecture and specifically fine-tuned for emotion classification on social media text. It covers seven emotion categories that align closely with Ekman's taxonomy, making it a natural fit for our classification task. Unlike general-purpose transformers such as BERT and RoBERTa, this model brings domain-specific emotional knowledge from its pretraining, which may benefit performance on short, informal text such as tweets.

To apply the model to our task, we replaced its original classification head with a new seven-class output layer corresponding to our emotion categories and fine-tuned the full model on our labeled dataset. The performance of this model was evaluated using the stratified train/validation/test split (see Section 2.4). Class imbalance was addressed by computing class weights and incorporating them into a weighted cross-entropy loss via a custom *Trainer* class. The model was trained for up to 15 epochs with early stopping (patience = 3), using the validation set exclusively for checkpoint selection. Final metrics were reported on the held-out test set.

2.9.7. DeBERTa-v3-Base

DeBERTa-v3-base [82] is a transformer model that improves upon BERT and RoBERTa through a disentangled attention mechanism, which separately encodes word content and positional information. This architectural refinement allows DeBERTa to capture more precise relationships between words and their positions in a sequence, and it has demonstrated strong performance on a wide range of NLP benchmarks. We included DeBERTa-v3-base as a representative of more recent general-purpose transformer architectures beyond those originally evaluated in this study.

We fine-tuned the model on our seven-class emotion dataset following the same protocol used for BERT and RoBERTa. The dataset was split into stratified training, validation, and test sets (70/15/15). Class weights were computed from the training partition and incorporated into the loss function. The model was trained for up to 15 epochs with early stopping (patience = 3), and final metrics are reported on the held-out test set.

Although DeBERTa-v3-base is a powerful general-purpose transformer model with an architecture designed to improve contextual language representation, its advantages may be less pronounced in short, noisy social media texts. In our dataset, tweets were truncated to a maximum of 50 tokens, limiting the amount of contextual information available to the model. This provides important context for interpreting its performance relative to RoBERTa.

2.9.8. Flan-T5-Base, Zero-Shot

Flan-T5-base [83] is an instruction-tuned language model developed by Google, trained to follow natural language instructions across a wide range of tasks. Unlike the fine-tuned classifiers described above, Flan-T5 operates in a zero-shot prompting setting: rather than being trained on our labeled dataset, it receives a natural language prompt describing the classification task and produces a text response indicating the predicted

emotion category. This makes it representative of the prompting paradigm associated with large language models such as GPT-4, allowing us to assess how instruction-following models perform relative to supervised fine-tuning approaches without requiring API access or significant computational resources.

To classify each tweet, we constructed a prompt that described the seven emotion categories and asked the model to assign the single most appropriate label. The model's text output was then matched to the nearest valid emotion category. As no training or checkpoint selection was involved, Flan-T5 was evaluated on the same held-out test set as the fine-tuned models ($n = 151$), ensuring a consistent and fair comparison of final predictions across all approaches.

3. Results

Overall, fine-tuned transformer-based models generally outperformed lexicon-based and classical machine-learning approaches across the main evaluation metrics (Table 2; Figure 1). Among all models, RoBERTa achieved the highest overall Accuracy and agreement ($\kappa = 0.461$), while DistilRoBERTa-Emotion achieved the highest Macro-averaged F_1-score (0.418), suggesting that emotion-specific pretraining may improve performance across less frequent categories. Most models performed best on the majority class, *Neutral*, but struggled to detect minority emotion categories such as *Disgust*, *Fear*, and *Surprise*. This pattern likely reflects both the inherent difficulty of discrete emotion detection and the challenges of working with a relatively small, highly imbalanced dataset. In contrast, lexicon-based methods, including NRC and Empath, and traditional classifiers, such as Naive Bayes and SVM, showed more limited effectiveness, particularly in balancing performance across multiple emotion classes.

Table 2. Comparison of Emotion Classification Models Across Evaluation Metrics.

Model	Accuracy	Cohen's κ	Macro F1	Joy	Sadness	Anger	Fear	Surprise	Disgust	Neutral
NRC Lexicon [a]	0.331	0.018	0.153	0.129	0.182	0.128	0.098	0.000	0.000	0.532
Empath [a]	0.510	0.055	0.192	0.000	0.091	0.065	0.000	0.500	0.000	0.692
Naive Bayes [b]	0.542	0.021	0.109	0.000	0.000	0.062	0.000	0.000	0.000	0.703
SVM [b]	0.578	0.324	0.269	0.235	0.184	0.482	0.115	0.000	0.093	0.775
CNN (baseline) [a]	0.616	0.310	0.188	0.000	0.000	0.537	0.000	0.000	0.000	0.781
GloVe CNN [a]	0.616	0.345	0.265	0.167	0.167	0.507	0.000	0.000	0.222	0.791
BERT [a]	0.583	0.396	0.328	0.235	0.182	0.533	0.200	0.000	0.333	0.813
RoBERTa [a]	0.656	0.461	0.364	0.526	0.308	0.491	0.000	0.000	0.375	0.845
Twitter-RoBERTa [a]	0.616	0.428	0.348	0.182	0.222	0.508	0.267	0.000	0.417	0.841
GoEmotions (Zero) [c]	0.575	0.226	0.278	0.294	0.330	0.214	0.122	0.176	0.067	0.745
GoEmotions (Fine) [a]	0.616	0.348	0.254	0.105	0.167	0.486	0.000	0.000	0.222	0.800
DistilRoBERTa-Emotion [a]	0.609	0.420	0.418	0.167	0.522	0.467	0.308	0.333	0.316	0.815
DeBERTa-v3-base [a]	0.570	0.357	0.318	0.278	0.316	0.385	0.000	0.000	0.455	0.792
Flan-T5-base (Zero) [a]	0.172	0.054	0.108	0.258	0.000	0.314	0.138	0.000	0.000	0.048

[a] Evaluated on a held-out test set ($n = 151$, 15% of the full dataset), not seen during training or model selection. [b] Evaluated via stratified 5-fold cross-validation on the full dataset ($n = 1005$); no checkpoint selection is involved in these models. [c] No training involved; evaluated on the full dataset ($n = 1005$). Values in blue reflect updated results following the revised evaluation protocol (see Section 2.4).

Emotion-wise performance varied substantially across models, particularly for less frequent categories (Figure 2). Among the five best-performing models ranked by Macro F_1-score, RoBERTa performed strongly on *Joy* and *Anger*, while DistilRoBERTa-Emotion showed more balanced performance across several less frequent categories, including *Sadness*, *Fear*, and *Surprise*. Twitter-RoBERTa also performed well on *Anger* and *Disgust*, but struggled with *Surprise*. These patterns show that even the strongest models differed in their ability to detect specific emotions.

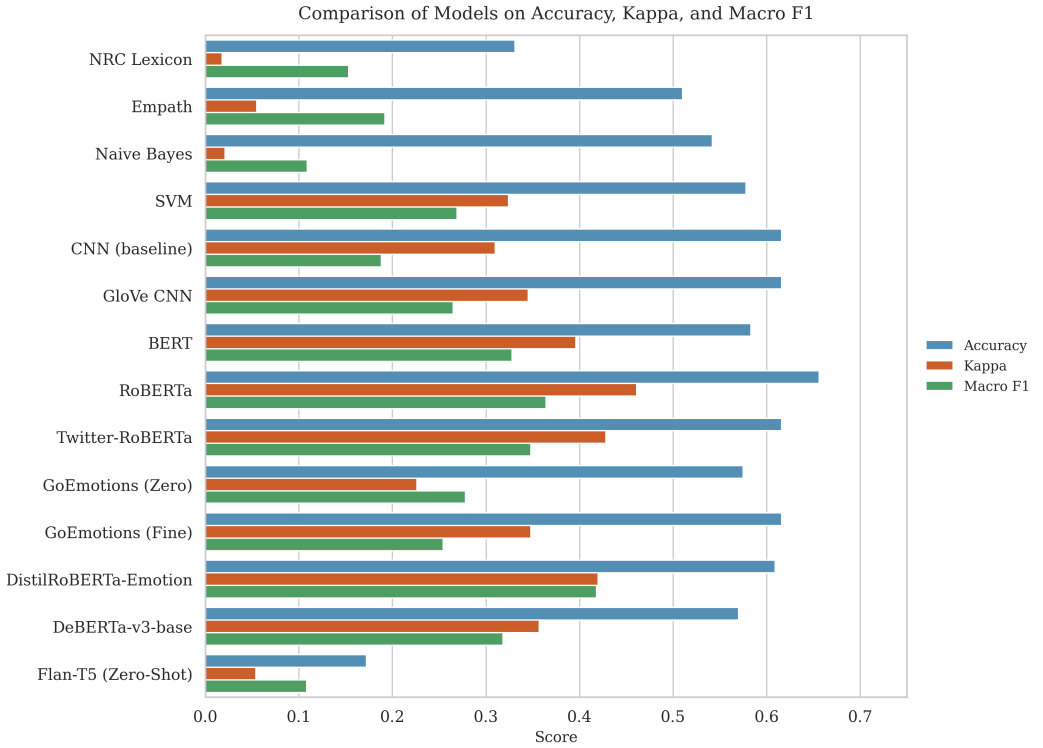

Figure 1. Comparison of Classification Performance Across All 14 Models Using Accuracy, Cohen's κ, and Macro F1-score.

The full per-emotion comparison across all models further reinforces this pattern (Figure 3). Across model families, *Neutral* and *Anger* were generally easier to detect, whereas *Fear*, *Surprise*, and, in some cases, *Disgust* remained more challenging. This finding highlights the importance of examining class-specific performance rather than relying solely on overall Accuracy, Cohen's κ, or aggregate F_1-scores.

Performance Results by Model

The NRC and Empath lexicons yielded the weakest performance overall. The NRC Lexicon reached only 33.1% Accuracy and a Macro F_1-score of 0.153, with a Cohen's κ of 0.018, indicating near chance-level agreement with the ground truth (*gold standard*). Empath performed slightly better, with 51.0% Accuracy and Macro F_1 of 0.192. Both lexicons performed reasonably on *Neutral* tweets but failed to capture nuance in emotionally rich content. These methods rely on predefined word lists and cannot account for context or subtle variations in meaning, making them particularly ill-suited for detecting emotions like *Disgust* or *Surprise* that often require contextual understanding.

Traditional models, such as Naive Bayes and SVM, demonstrated modest improvements over lexicon-based approaches. Naive Bayes achieved 54.2% Accuracy but had a Macro F_1 of only 0.109, as it overwhelmingly predicted the dominant class. The SVM classifier performed better with 57.8% Accuracy and a Macro F_1 of 0.269. SVM showed improved performance in recognizing *Anger* and *Joy*, but still lacked the sensitivity to detect rare emotions. These models benefited from learning *n*-gram patterns but lacked the deep contextual understanding offered by more recent approaches.

The baseline CNN achieved 61.6% Accuracy and a Macro F_1 of 0.188, mostly by overfitting to the *Neutral* class. Incorporating pre-trained GloVe embeddings led to notable gains: the GloVe CNN reached 61.6% Accuracy and a Macro F_1 of 0.265, with more balanced performance across emotions such as *Anger* and *Disgust*. The results suggest that incor-

porating external semantic information, even from general-purpose word embeddings, enables better generalization in detecting non-neutral sentiments.

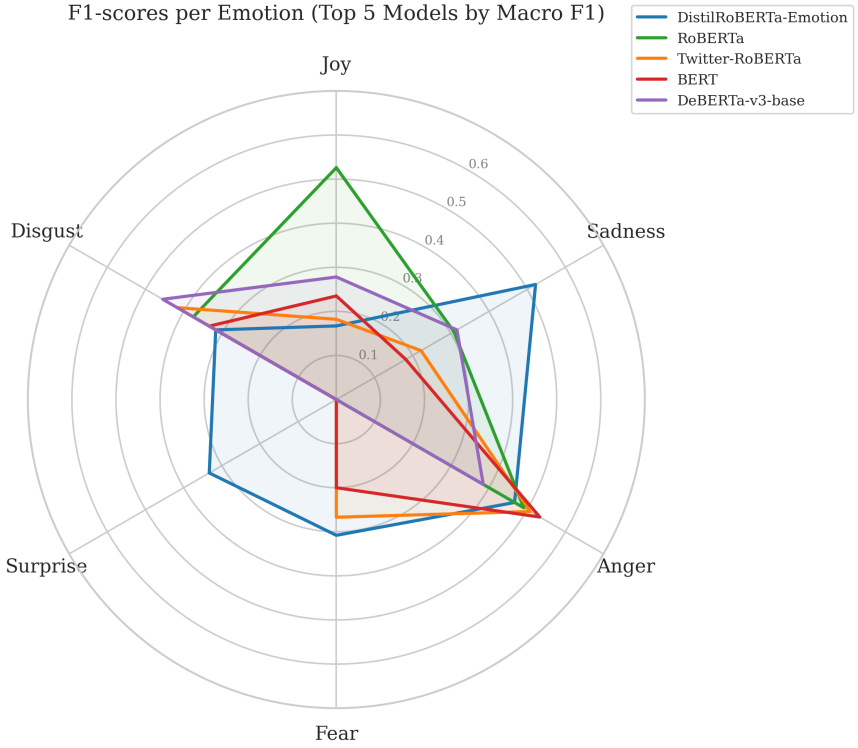

Figure 2. Per-Emotion F_1-Scores for the Top Five Models Ranked by Macro F_1 (DistilRoBERTa-Emotion, RoBERTa, Twitter-RoBERTa, BERT, and DeBERTa-v3-base). *Note.* The *Neutral* category was excluded because the top five models achieved similarly high F_1-scores for that class, ranging from 0.79 to 0.85. Including *Neutral* would compress the scale and obscure meaningful differences across the six discrete emotion categories.

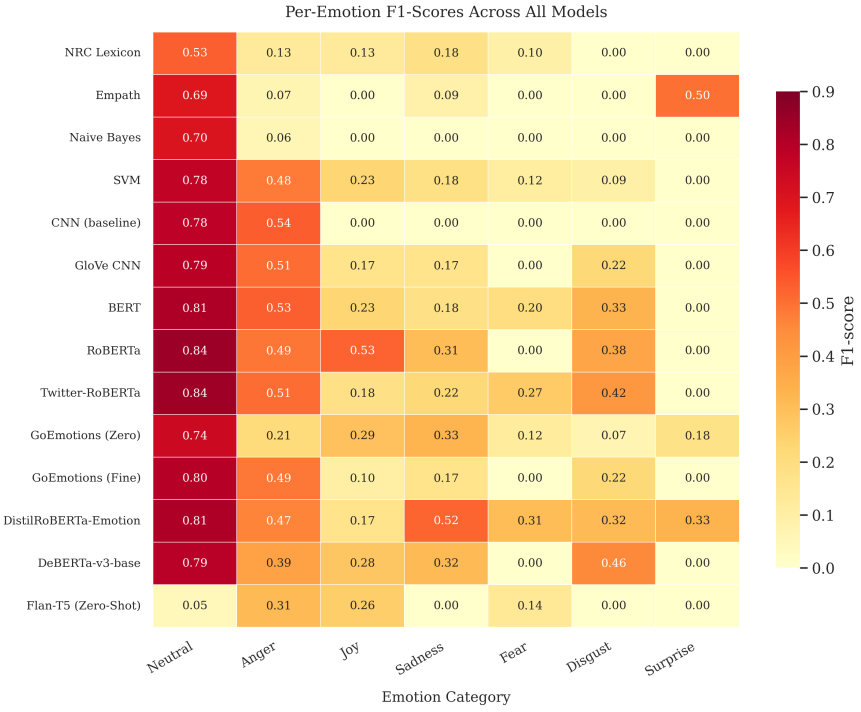

Figure 3. Heatmap of Per-Emotion F_1-Scores Across All 14 Evaluated Models. *Note.* Rows represent models and columns represent emotion categories, ordered from most to least frequent in the dataset. Darker cells indicate higher F_1-scores.

Fine-tuned BERT achieved 58.3% Accuracy and Macro F_1 of 0.328. Although its Accuracy was slightly lower than that of the GloVe CNN (0.616), BERT outperformed it on Macro F_1 and showed notably stronger agreement ($\kappa = 0.396$), indicating more balanced performance across emotion categories rather than reliance on the majority class. RoBERTa achieved the highest overall Accuracy (65.6%) and the strongest agreement with the gold standard ($\kappa = 0.461$) of all models evaluated, with a Macro F_1 of 0.364. Twitter-RoBERTa followed closely with 61.6% Accuracy and Macro F_1 of 0.348, reflecting the benefits of domain-specific pretraining on social media data. These transformer models capture deep contextual features that allow them to distinguish between emotion categories that often appear in similar lexical forms but differ semantically.

The zero-shot GoEmotions model achieved 57.5% Accuracy and Macro F_1 of 0.278. When fine-tuned on our dataset, the model improved in class agreement ($\kappa = 0.348$) but paradoxically saw a lower Macro F_1 (0.254). This result suggests that although the pre-trained model transfers some general emotion understanding, fine-tuning on a small, imbalanced dataset may reduce its sensitivity to minority classes. More precisely, replacing the original 28-class output head with a 7-class head trained on limited, skewed data collapses the model's rich emotional representations toward dominant categories, suppressing minority class sensitivity even without classical overfitting.

A detailed analysis of RoBERTa's classification behavior is provided in Figure A1 (see Appendix A), which presents the normalized confusion matrix for the best-performing model by overall Accuracy. The matrix reveals that the dominant failure mode was the tendency to misclassify minority emotions as *Neutral*: this pattern is most pronounced for *Fear* (0.67 misclassified as *Neutral*), but also affects *Sadness* (0.33) and *Joy*/*Happiness* (0.29). A secondary pattern is confusion between emotionally adjacent categories, particularly between *Anger* and *Disgust*. In other words, the model had difficulty detecting emotionally restrained or low-intensity expressions of minority emotions, which were frequently absorbed into the dominant *Neutral* category, and also struggled to distinguish between closely related negative emotions.

Among the additional transformer models evaluated, DistilRoBERTa-Emotion stood out, achieving the highest Macro F_1 of any model in the study (0.418), with an overall Accuracy of 60.9% and $\kappa = 0.420$. Notably, it detected *Sadness* (0.522) and *Surprise* (0.333) more effectively than any other fine-tuned model, including RoBERTa, and remained competitive on *Disgust* (0.316), although DeBERTa-v3-base achieved the highest score on this category (0.455). Taken together, these results show that DistilRoBERTa-Emotion had the strongest Macro F_1-score and performed comparatively well on several minority categories, particularly *Sadness* and *Surprise*, despite not achieving the highest overall Accuracy.

DeBERTa-v3-base achieved 57.0% Accuracy and a Macro F_1 of 0.318, performing below RoBERTa on all key metrics. The zero-shot Flan-T5-base model achieved an overall Accuracy of only 17.2% and a Macro F_1 of 0.108, performing substantially below all other approaches, including the lexicon-based methods. The model's F_1 on the dominant *Neutral* class was only 0.048, indicating that it almost never predicted *Neutral* despite this being the majority category in the dataset. Across all models, performance was consistently strongest on the *Neutral* category and declined for minority emotions, with several models producing F_1-scores of 0.000 on *Fear*, *Surprise*, or *Disgust* (see Table 2). Lexicon-based methods (NRC, Empath) failed to detect *Disgust* entirely, and NRC additionally produced an F_1 of 0.000 on *Surprise* (Figure 3).

4. Discussion

Communication scholars have long examined how emotional appeals influence public opinion and behavior. Extending this line of research, this study makes a methodological contribution to emotion analysis by systematically comparing emotion classification approaches applied to political social media discourse. In doing so, it builds on prior validation work comparing manual, machine-learning, and dictionary-based approaches to sentiment classification, which demonstrated that trained human coders remain an important benchmark for evaluating automated methods [33]. Our study extends that foundation by addressing the more demanding task of fine-grained emotion detection, which moves beyond polarity-based sentiment analysis and involves a multi-class classification problem with greater conceptual and methodological complexity.

Unlike sentiment analysis, which typically captures overall positive or negative valence, emotion analysis seeks to identify discrete emotional states (e.g., *anger*, *joy*, *fear*, *sadness*) (e.g., [35,84]). This task is substantially more challenging for several reasons. First, emotions are highly context-sensitive and often co-occur within the same message, particularly in ambiguous, ironic, or sarcastic expressions (e.g., [42,43]). Second, compared to broad sentiment cues, linguistic markers of specific emotions tend to be more subtle and diffuse (e.g., [37,64]). These challenges emphasize that emotion detection cannot be treated as a straightforward extension of sentiment analysis, either conceptually or computationally, and require careful evaluation of model performance under realistic conditions.

From a methodological perspective, our findings provide important insights into the validity and reliability of computational emotion analysis tools. Model performance varies considerably depending on context, dataset characteristics, and the domain to which tools are applied (e.g., [85]). In our case, the additional challenge of imbalanced emotional distributions further compounds these validity concerns. In politically charged discourse, such as conflict-related social media content, language is often metaphorical, emotionally intensified [86], or sarcastic [87], which increases the likelihood of misclassification. As a limitation of the present study, emotions expressed through irony, sarcasm, or complex mixtures were not explicitly modeled. This limitation likely affected classification accuracy, particularly for minority emotion categories. Indeed, prior research demonstrates that even state-of-the-art large language models continue to struggle with accurately detecting sarcasm and emotional intensity in text (e.g., [87]), raising important concerns for construct validity in automated emotion detection.

Empirically, our evaluation demonstrates that transformer-based models consistently outperform convolutional neural networks, traditional machine-learning algorithms, and dictionary-based methods in fine-grained emotion classification. Importantly, this performance advantage is not marginal; it is stable across evaluation metrics and particularly pronounced for infrequently expressed emotions. Cohen's κ scores further confirm this pattern by quantifying chance-corrected agreement between each model's predictions and the gold-standard labels. Given the imbalanced distribution of emotion categories, this convergence across Cohen's κ and Macro F_1 metrics strengthens confidence in the comparative ranking of models. Nevertheless, even the best-performing transformer models do not fully approximate the human-coded gold standard, emphasizing the continued importance of human annotation in emotion analysis (e.g., [33,88]). These findings extend the existing evidence base by validating these patterns on emotionally charged, real-world political social media data under authentic class imbalance conditions. Figure A2 (Appendix A) further illustrates these patterns at the model family level, confirming that emotion-specific transformers achieve the highest average Macro F_1, whereas lexicon-based methods and zero-shot approaches show the weakest overall performance.

Two model-specific findings deserve closer examination, as they illustrate that strong empirical performance does not follow automatically from architectural sophistication or model scale. Despite being a more recent and architecturally advanced model than BERT and RoBERTa, DeBERTa-v3-base did not outperform RoBERTa on our task. This result is consistent with known properties of the DeBERTa architecture: its disentangled attention mechanism provides the greatest benefit on longer, more syntactically complex text, where precise positional encoding plays a larger role. On short social media posts truncated to 50 tokens, the model does not have sufficient context to fully leverage these architectural advantages. This finding highlights an important practical consideration for communication researchers: architectural sophistication does not automatically translate to better performance on short social media text, and model selection should be guided by the nature of the target data rather than benchmark rankings alone.

The poor performance of Flan-T5-base in the zero-shot setting tells a related but distinct story. Its failure reflects a known tendency of instruction-tuned models to favor specific, meaningful answers when prompted, making them reluctant to assign a neutral or non-committal label. This explains the model's near-zero F_1-score on the *Neutral* class: when prompted to choose an emotion, the model almost always selected a non-neutral category, even when *Neutral* would have been the correct label. This result does not suggest that large language models are inherently unsuited for emotion detection, but it does demonstrate that zero-shot prompting without any task-specific adaptation is insufficient for reliable emotion classification on short, imbalanced social media text. Fine-tuning on labeled data remains essential for achieving meaningful performance on this type of task.

Our findings highlight the need for methodological transparency and cautious interpretation when adopting computational emotion analysis tools. While computational models offer clear advantages in scalability and efficiency (e.g., [33]), they should not be viewed as replacements for human judgment in contexts where interpretive precision is essential (e.g., [89]). Instead, our findings suggest that a complementary, multi-method approach offers the most reliable path forward, one in which computational emotion detection models are used to identify large-scale patterns, while key findings are validated through human annotation.

From a practical standpoint, our findings offer clear guidance for researchers conducting emotion analysis in social media contexts. Transformer-based models should be prioritized for large-scale or highly imbalanced datasets, where fine-grained emotional distinctions are critical. CNN-based approaches may offer a reasonable compromise for mid-sized projects with fewer computational resources, whereas traditional machine-learning and dictionary-based methods show limited effectiveness for complex emotion detection tasks. Across all approaches, manual annotation remains essential for validation, particularly when studies aim to draw substantive conclusions about emotional dynamics in public discourse.

Limitations and Future Directions

The scope of this study is limited to English-language tweets during a specific geopolitical event, which may affect generalizability. We treated emotion categories as mutually exclusive to support consistent and interpretable classification results. However, future work could explore multi-label and intensity-based emotion annotations. Additionally, while we tested a representative range of models, larger foundation models (e.g., GPT-4) are yet to be extensively evaluated. Our model selection focused on the range of state-of-the-art transformer architectures fine-tuned explicitly for emotion recognition tasks, such as Twitter-RoBERTa and GoEmotions, which have demonstrated competitive performance on relevant benchmarks. The selection and fine-tuning of the analyzed state-of-the-art models

were conducted before the emergence and widespread availability of several recent large language models (LLMs) optimized for emotion recognition. While establishing such scope boundaries is a standard practice in empirical research to maintain feasibility, focus, and consistency, we acknowledge the importance of these newer models and recommend that future studies systematically examine their capabilities on complex social media emotion detection tasks.

Three additional methodological observations merit discussion. First, the relatively small size of the held-out test partition ($n = 151$) means that per-class metrics for rare emotion categories such as *Surprise* ($n = 2$ test instances) and *Disgust* ($n = 8$) are subject to high variance, and single misclassifications can produce large swings in class-level F_1-scores. These figures should therefore be interpreted with appropriate caution. Second, the observed decline in macro F_1 for the fine-tuned GoEmotions model (0.254) relative to its zero-shot counterpart (0.278) is best understood not as classical overfitting but as a consequence of adapting a model pre-trained on 28 fine-grained emotion categories to a coarse seven-class, heavily imbalanced task with limited labeled data. Replacing the original classification head discards pretrained output representations, and the resulting fine-tuning pressure from dominant classes suppresses sensitivity to minority emotions. Future work should explore parameter-efficient fine-tuning strategies (e.g., LoRA, adapter layers) or multi-task objectives that preserve fine-grained emotional representations while adapting to target label spaces. Third, as discussed above, the poor performance of Flan-T5-base in the zero-shot prompting setting (Accuracy = 17.2%, Macro F_1 = 0.108) should not be interpreted as a definitive assessment of large language models for emotion detection. More capable instruction-tuned models such as GPT-4, combined with carefully designed prompts or few-shot examples, may yield substantially different results. Future research should systematically evaluate such models under controlled conditions comparable to those used in this study, including consistent evaluation sets and explicit handling of class imbalance.

Future work should also examine how classification accuracy affects substantive interpretations in communication studies, particularly when measuring emotional framing effects or emotional contagion in political discourse. The present study focuses on one conflict context and one platform, but different conflicts, larger samples, and alternative social media environments may yield different emotional patterns. Therefore, future studies should expand the scope by examining multiple conflict events across diverse contexts and incorporating platforms beyond Twitter/X to assess the generalizability of the current findings.

Author Contributions: Conceptualization, A.V. and V.P.; methodology, A.V. and V.P.; software, V.P.; validation, A.V. and J.S.; formal analysis, V.P.; investigation, A.V.; resources, A.V., V.P., and J.S.; data curation, A.V. and J.S.; writing—original draft preparation, A.V. and V.P.; writing—review and editing, A.V., J.S., and P.B.; visualization, V.P.; supervision, A.V.; project administration, A.V. and V.P. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by Texas Tech University.

Data Availability Statement: Data and materials related to this study are available from the corresponding authors upon reasonable request.

Acknowledgments: Claude Pro was used for refining visualization codes. The authors reviewed and verified all outputs and take full responsibility for the final content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Additional Figures

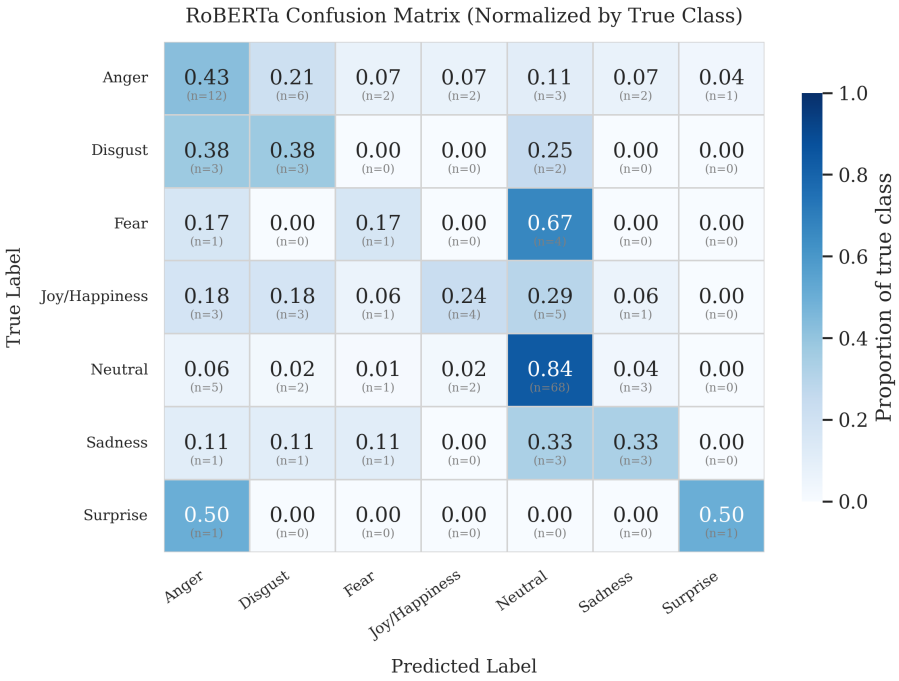

Figure A1. Normalized Confusion Matrix for RoBERTa, the Best-Performing Model by Overall Accuracy and Cohen's κ. *Note.* Each cell shows the proportion of true class instances predicted as each label, with raw counts in parentheses. The diagonal represents correct classifications. RoBERTa performs best on the *Neutral* class (0.84). Performance is weaker for several emotion categories, including *Fear* (0.17), *Joy/Happiness* (0.24), and *Sadness* (0.33). Although *Surprise* shows a diagonal value of 0.50, this result should be interpreted cautiously because the class includes only two cases. A notable pattern is the frequent misclassification of *Fear* as *Neutral* (0.67), indicating that the model often failed to distinguish fear-related tweets from neutral content in this dataset.

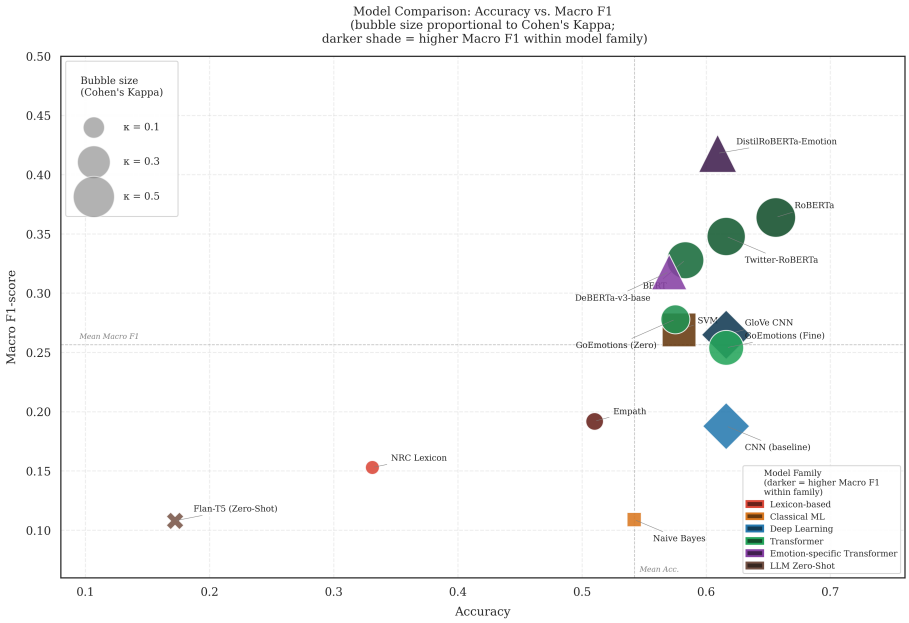

Figure A2. Model Performance Across Accuracy, Cohen's κ, and Macro F_1-Score, Grouped by Model Family. *Note.* Overall, the emotion-specific transformer *DistilRoBERTa-Emotion* achieves the highest Macro F_1-score, while general-purpose transformer models such as *RoBERTa* and *Twitter-RoBERTa* show strong performance in Accuracy and Cohen's κ. Lexicon-based methods and the zero-shot LLM model show weaker performance overall, suggesting that task-specific or supervised transformer-based approaches are more reliable for emotion classification in short social media texts.

References

1. Lamprianou, I.; Ellinas, A.A. Emotion, sophistication and political behavior: Evidence from a laboratory experiment. *Political Psychol.* **2019**, *40*, 859–876. [CrossRef]
2. Rico, G.; Guinjoan, M.; Anduiza, E. The emotional underpinnings of populism: How anger and fear affect populist attitudes. *Swiss Political Sci. Rev.* **2017**, *23*, 444–461. [CrossRef]
3. Jones, P.E.; Hoffman, L.H.; Young, D.G. Online emotional appeals and political participation: The effect of candidate affect on mass behavior. *New Media Soc.* **2013**, *15*, 1132–1150. [CrossRef]
4. Valentino, N.A.; Brader, T.; Groenendyk, E.W.; Gregorowicz, K.; Hutchings, V.L. Election night's alright for fighting: The role of emotions in political participation. *J. Politics* **2011**, *73*, 156–170. [CrossRef]
5. Vandenbroek, L.M. Disentangling aversion: Experimentally testing the impact of disgust and anger on political participation. In Proceedings of the APSA 2011 Annual Meeting Paper, Seattle, WA, USA, 1–4 September 2011.
6. Young, L.E. Mobilization under threat: Emotional appeals and pro-opposition political participation online. *Political Behav.* **2023**, *45*, 445–468. [CrossRef]
7. Mustonen, A.M.; Paakkonen, T.; Ryökäs, E.; Nieminen, P. Abortion debates in Finland and the Republic of Ireland: Textual analysis of experiential thinking and argumentation in parliamentary and layperson discussions. *Reprod. Health* **2017**, *14*, 163. [CrossRef] [PubMed]
8. Gerstlé, J.; Nai, A. Negativity, emotionality and populist rhetoric in election campaigns worldwide, and their effects on media attention and electoral success. *Eur. J. Commun.* **2019**, *34*, 410–444. [CrossRef]
9. Serazio, M. Branding politics: Emotion, authenticity, and the marketing culture of American political communication. *J. Consum. Cult.* **2017**, *17*, 225–241. [CrossRef]
10. Van Kleef, G.A.; Van den Berg, H.; Heerdink, M.W. The persuasive power of emotions: Effects of emotional expressions on attitude formation and change. *J. Appl. Psychol.* **2015**, *100*, 1124. [CrossRef]
11. Hasell, A.; Weeks, B.E. Partisan provocation: The role of partisan news use and emotional responses in political information sharing in social media. *Hum. Commun. Res.* **2016**, *42*, 641–661. [CrossRef]
12. Molina, M.D. Do people believe in misleading information disseminated via memes? The role of identity and anger. *New Media Soc.* **2025**, *27*, 847–870. [CrossRef]
13. Caruso, G. Emotional cycles and collective action: Global crises and the World Social Forum. *Psychoanal. Cult. Soc.* **2023**, *28*, 1–19. [CrossRef]
14. Fink, O.; Hasan Aslih, S.; Halperin, E. Two paths to violence: Individual versus group emotions during conflict escalation in the Occupied Palestinian Territories. *Group Process. Intergroup Relat.* **2025**, *28*, 355–376. [CrossRef]
15. Leonard, D.J.; Moons, W.G.; Mackie, D.M.; Smith, E.R. "We're mad as hell and we're not going to take it anymore": Anger self-stereotyping and collective action. *Group Process. Intergroup Relat.* **2011**, *14*, 99–111. [CrossRef]
16. Bestvater, S.; Shah, S.; River, G.; Smith, A. Politics on Twitter: One-Third of Tweets from Us Adults Are Political. Available online: https://www.pewresearch.org/politics/2022/06/16/politics-on-twitter-one-third-of-tweets-from-u-s-adults-are-political/ (accessed on 17 May 2025).
17. McClain, C.; Anderson, M.; Gelles-Watnick, R. How Americans Navigate Politics on TikTok, X, Facebook and Instagram. The Experiences and Views of Each Site's Users–From How Much Political Content They See to the Platforms' Impact on Democracy. Available online: https://www.pewresearch.org/internet/2024/06/12/how-americans-navigate-politics-on-tiktok-x-facebook-and-instagram/ (accessed on 17 May 2025).
18. Bil-Jaruzelska, A.; Monzer, C. All About Feelings? Emotional Appeals as Drivers of User Engagement with Facebook Posts. *Politics Gov.* **2022**, *10*, 172–184. [CrossRef]
19. Choi, J.; Lee, S.Y.; Ji, S.W. Engagement in emotional news on social media: Intensity and type of emotions. *Journal. Mass Commun. Q.* **2021**, *98*, 1017–1040. [CrossRef]
20. Himelboim, I.; Borah, P.; Lorenzano, K.; Lee, J.J.; Cao, X. If it Bleeds, it Doesn't Lead: Emotional Appeals and Engagement in Immigration and Election Conversations on Twitter. *J. Broadcast. Electron. Media* **2025**, *69*, 1–23. [CrossRef]
21. Rietveld, R.; Van Dolen, W.; Mazloom, M.; Worring, M. What you feel, is what you like influence of message appeals on customer engagement on Instagram. *J. Interact. Mark.* **2020**, *49*, 20–53. [CrossRef]
22. Stieglitz, S.; Dang-Xuan, L. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *J. Manag. Inf. Syst.* **2013**, *29*, 217–248. [CrossRef]
23. Ross, A.A. *Mixed Emotions: Beyond Fear and Hatred in International Conflict*; University of Chicago Press: Chicago, IL, USA, 2019.
24. Otto, L.P. Beyond simple valence: Discrete emotions as mediators of political communication effects on trust in politicians. *SCM Stud. Commun. Media* **2018**, *7*, 364–391. [CrossRef]
25. Crabtree, C.; Golder, M.; Gschwend, T.; Indriđason, I.H. It is not only what you say, it is also how you say it: The strategic use of campaign sentiment. *J. Politics* **2020**, *82*, 1044–1060. [CrossRef]

26. Haselmayer, M. Candidates rather than context shape campaign sentiment in French Presidential Elections (1965–2017). *Fr. Politics* **2021**, *19*, 394–420. [CrossRef]
27. Kosmidis, S.; Hobolt, S.; Molloy, A.; Whitefield, S. Party competition and emotive rhetoric. *Comp. Political Stud.* **2019**, *52*, 811–837. [CrossRef]
28. Schöne, J.P.; Parkinson, B.; Goldenberg, A. Negativity spreads more than positivity on Twitter after both positive and negative political situations. *Affect. Sci.* **2021**, *2*, 379–390. [CrossRef] [PubMed]
29. Shin, J.; Thorson, K. Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *J. Commun.* **2017**, *67*, 233–255. [CrossRef]
30. Blassnig, S.; Udris, L.; Staender, A.; Vogler, D. Popularity on Facebook during election campaigns: An analysis of issues and emotions in parties' online communication. *Int. J. Commun.* **2021**, *15*, 4399–4419.
31. Zerback, T.; Wirz, D.S. Appraisal patterns as predictors of emotional expressions and shares on political social networking sites. *Stud. Commun. Sci.* **2021**, *21*, 27–45. [CrossRef]
32. Young, L.; Soroka, S. Affective news: The automated coding of sentiment in political texts. *Political Commun.* **2012**, *29*, 205–231. [CrossRef]
33. Van Atteveldt, W.; Van der Velden, M.A.; Boukes, M. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Commun. Methods Meas.* **2021**, *15*, 121–140. [CrossRef]
34. Gross, T.W. Sentiment analysis and emotion recognition: Evolving the paradigm of communication within data classification. *Appl. Mark. Anal.* **2020**, *6*, 22–36. [CrossRef]
35. Vendrell Ferran, I. Emotions and sentiments: Two distinct forms of affective intentionality. *Phenomenol. Mind* **2022**, *23*, 20–34. [CrossRef]
36. Botha, E.; Reyneke, M. To share or not to share: The role of content and emotion in viral marketing. *J. Public Aff.* **2013**, *13*, 160–171. [CrossRef]
37. Schuff, H.; Barnes, J.; Mohme, J.; Padó, S.; Klinger, R. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Copenhagen, Denmark, 8 September 2017; pp. 13–23.
38. Ekman, P. Are there basic emotions? *Psychol. Rev.* **1992**, *99*, 550–553. [CrossRef] [PubMed]
39. Bestvater, S.E.; Monroe, B.L. Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Anal.* **2023**, *31*, 235–256. [CrossRef]
40. Suhaimin, M.S.M.; Hijazi, M.H.A.; Alfred, R.; Coenen, F. Modified framework for sarcasm detection and classification in sentiment analysis. *Indones. J. Electr. Eng. Comput. Sci.* **2019**, *13*, 1175–1183. [CrossRef]
41. Aymé, E.D.; Rivadeneira, R.; Fuertes, W. Emotional Tone Detection in Hate Speech Using Machine Learning and NLP: Methods, Challenges, and Future Directions—A Systematic Review. *Appl. Sci.* **2025**, *15*, 12686. [CrossRef]
42. Štajner, S. Exploring reliability of gold labels for emotion detection in Twitter. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Online, September 2021; pp. 1350–1359.
43. Acheampong, F.A.; Wenyu, C.; Nunoo-Mensah, H. Text-based emotion detection: Advances, challenges, and opportunities. *Eng. Rep.* **2020**, *2*, e12189. [CrossRef]
44. Zhen, Z.; Li, Y. C-STEER: A Dynamic Sentiment-Aware Framework for Fake News Detection with Lifecycle Emotional Evolution. *Informatics* **2026**, *13*, 4. [CrossRef]
45. Lim, D.; Lee, T. Enhanced Recommender System with Sentiment Analysis of Review Text and SBERT Embeddings of Item Descriptions. *Mathematics* **2026**, *14*, 184. [CrossRef]
46. Keinan, R.; Margalit, E.; Bouhnik, D. Emotional analysis in a morphologically rich language: Enhancing machine learning with psychological feature lexicons. *Electronics* **2025**, *14*, 3067. [CrossRef]
47. Anthony, P.; Zhou, J. Leveraging Transformer with Self-Attention for Multi-Label Emotion Classification in Crisis Tweets. *Informatics* **2025**, *12*, 114. [CrossRef]
48. Yixin, Z. Harnessing computational techniques to analyze and enhance fandom engagement through interaction ritual chains. *Appl. Comput. Eng.* **2024**, *82*, 100–105. [CrossRef]
49. Xiong, H.; Lv, S. Factors affecting social media users' emotions regarding food safety issues: Content analysis of a debate among chinese weibo users on genetically modified food security. *Healthcare* **2021**, *9*, 113. [CrossRef]
50. Yoo, W.; Namkoong, K.; Choi, M.; Shah, D.V.; Tsang, S.J.; Hong, Y.; Aguilar, M.; Gustafson, D.H. Giving and receiving emotional support online: Communication competence as a moderator of psychosocial benefits for women with breast cancer. *Comput. Hum. Behav.* **2014**, *30*, 13–22. [CrossRef]
51. Wang, Z.; Zhang, X.; Han, D.; Zhao, Y.; Ma, L.; Hao, F. How the use of an online healthcare community affects the doctor-patient relationship: An empirical study in china. *Front. Public Health* **2023**, *11*, 1145749. [CrossRef]
52. Peng, S.; Yang, Y.; Haselton, M.G. Menstrual Symptoms: Insights from Mobile Menstrual Tracking Applications for English and Chinese Teenagers. *Adolescents* **2023**, *3*, 394–403. [CrossRef]

53. Coviello, L.; Sohn, Y.; Kramer, A.; Marlow, C.; Franceschetti, M.; Christakis, N.; Fowler, J. detecting emotional contagion in massive social networks. *PLoS ONE* **2014**, *9*, e90315. [CrossRef] [PubMed]
54. Crocamo, C.; Viviani, M.; Famiglini, L.; Bartoli, F.; Pasi, G.; Carrà, G. Surveilling COVID-19 emotional contagion on twitter by sentiment analysis. *Eur. Psychiatry* **2021**, *64*, e17. [CrossRef] [PubMed]
55. Jazeera, A. Bucha: World Reacts to 'Unbearable' Civilian Killings in Ukraine. 2022. Available online: https://www.aljazeera.com/news/2022/4/4/unbearable-world-reacts-to-civilian-killings-in-bucha-ukraine (accessed on 18 May 2025).
56. Ramos, L.; Chang, O. Sentiment analysis of russia-ukraine conflict tweets using roberta. *Uniciencia* **2023**, *37*, 1–11. [CrossRef]
57. Gabel, S.; Reichert, L.; Reuter, C. Discussing conflict in social media: The use of twitter in the jammu and kashmir conflict. *Media War Confl.* **2020**, *15*, 504–529. [CrossRef]
58. Andrich, A. Beyond Neutrality: A Longitudinal Analysis of Gender, Party, and Media Visibility in the Tone of US Political Coverage. *Int. J. Press/Politics* **2025**. [CrossRef]
59. Boukes, M.; van de Velde, B.; Araujo, T.; Vliegenthart, R. What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools. *Commun. Methods Meas.* **2020**, *14*, 83–104. [CrossRef]
60. Birkenmaier, L.; Lechner, C.M.; Wagner, C. The Search for Solid Ground in Text as Data: A Systematic Review of Validation Practices and Practical Recommendations for Validation. *Commun. Methods Meas.* **2024**, *18*, 249–277. [CrossRef]
61. Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; Ravi, S. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5 July 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA; pp. 4040–4054. [CrossRef]
62. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
63. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
64. Nandwani, P.; Verma, R. A review on sentiment analysis and emotion detection from text. *Soc. Netw. Anal. Min.* **2021**, *11*, 81. [CrossRef] [PubMed]
65. Mohammad, S.M.; Turney, P.D. Crowdsourcing a word–emotion association lexicon. *Comput. Intell.* **2013**, *29*, 436–465. [CrossRef]
66. Fast, E.; Chen, B.; Bernstein, M.S. Empath: Understanding Topic Signals in Large-Scale Text. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 7–12 May 2016; CHI '16, pp. 4647–4657. [CrossRef]
67. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009.
68. McCallum, A.; Nigam, K. A Comparison of Event Models for Naive Bayes Text Classification. *AAAI Conference on Artificial Intelligence*, 1998. Available online: https://api.semanticscholar.org/CorpusID:7311285 (accessed on 17 May 2025).
69. Peng, F.; Schuurmans, D.; Wang, S. Augmenting Naive Bayes Classifiers with Statistical Language Models. *Inf. Retr.* **2004**, *7*, 317–345. [CrossRef]
70. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
71. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
72. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
73. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2015**, arXiv:1412.6980.
74. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543. [CrossRef]
75. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
76. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
77. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv* **2019**, arXiv:1910.03771.
78. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692. [CrossRef]
79. Barbieri, F.; Camacho-Collados, J.; Espinosa-Anke, L.; Neves, L. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 1644–1650. [CrossRef]

80. Cardiff NLP. Twitter-RoBERTa-Base for Sentiment Analysis (Latest). 2022. Available online: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest (accessed on 17 May 2025).
81. Hartmann, J. Emotion English DistilRoBERTa-Base. 2022. Available online: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base (accessed on 17 May 2025).
82. He, P.; Gao, J.; Chen, W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv* **2021**, arXiv:2111.09543.
83. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling Instruction-Finetuned Language Models. *arXiv* **2022**, arXiv:2210.11416. [CrossRef]
84. Sailunaz, K.; Alhajj, R. Emotion and sentiment analysis from Twitter text. *J. Comput. Sci.* **2019**, *36*, 101003. [CrossRef]
85. Lu, H.; Ehwerhemuepha, L.; Rakovski, C. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Med. Res. Methodol.* **2022**, *22*, 181. [CrossRef]
86. Huguet Cabot, P.L.; Dankers, V.; Abadi, D.; Fischer, A.; Shutova, E. The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November, 2020; pp. 4479–4488. [CrossRef]
87. Bojić, L.; Zagovora, O.; Zelenkauskaite, A.; Vuković, V.; Čabarkapa, M.; Veseljević Jerković, S.; Jovančević, A. Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Sci. Rep.* **2025**, *15*, 11477. [CrossRef]
88. Song, H.; Tolochko, P.; Eberl, J.M. In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis. *Political Commun.* **2020**, *37*, 550–572. [CrossRef]
89. van der Meer, T.G. Automated content analysis and crisis communication research. *Public Relat. Rev.* **2016**, *42*, 952–961. [CrossRef]