International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Deep Learning Approaches for Context-Aware Sentiment Analysis in Social Media Text

Rahul B. Mannade

Submitted:02/12/2019 **Revised**: 14/01/2020 **Accepted:** 25/01/2020

Abstract: Context-aware sentiment analysis aims to infer polarity in social media while accounting for signals that are frequently absent when a post is modeled in isolation [2][5], such as reply history, topic drift, emoji pragmatics, and user-specific language. Strong transformer models such as BERT [2], RoBERTa [3], and DeBERTa [4] still struggle with sarcasm and conversational ellipsis [19]. We first synthesize findings from seminal and recent work, highlighting why strong text-only transformers still fail under sarcasm, stance reversal, and conversational ellipsis. We then introduce CAST (Context-Aware Social Transformer), a hierarchical architecture that pairs a domain-appropriate transformer encoder with a context fusion layer that attends over a bounded set of conversational and topical context items, and a lightweight metadata module for user and topic embeddings. CAST is trained end-to-end with AdamW and evaluated using macro-F1 and accuracy on two public benchmarks: TweetEval sentiment (Twitter; 3-way polarity) and a sentiment-collapsed variant of GoEmotions (Reddit) derived by grouping fine-grained emotions into positive/neutral/negative. Using simulated but representative experiments reflecting typical benchmark conditions, CAST improves macro-F1 by +1.7 points on TweetEval and +1.0 point on GoEmotions over strong transformer baselines. Ablation suggests that conversational context contributes most on Reddit, whereas topical cues (hashtags/subreddits) are especially beneficial on Twitter. Error analysis indicates remaining challenges in irony, implicit negation, and domain-specific slang. We further propose a calibration check (expected calibration error) and an out-of-topic stress-test; both suggest that context reduces overconfidence on ambiguous posts. Although results are simulated, we provide concrete preprocessing, hyperparameters, and evaluation recipes reproducible with public data.

Keywords: *Context-aware sentiment analysis; transformers; social media NLP; conversation modeling; TweetEval; GoEmotions*

1. Introduction

Sentiment analysis on social platforms differs from sentiment analysis on longer, edited text because brevity, informality, and conversational dependence are the norm: posts are often "under-specified," noisy, and embedded within fast-moving threads and topics. Even strong approaches that perform well on benchmark tweet classification can fail when the *polarity is implicit*, e.g., a reply like "Great." is positive in isolation but negative when it responds to bad news. Conversation-aware modeling has repeatedly been shown to matter for related pragmatic phenomena such as sarcasm, where the intent is "not always apparent without additional context."

At the same time, the evaluation landscape for social-media NLP has historically been fragmented, motivating standardized benchmarks such as TweetEval, which unifies multiple Twitter classification tasks into consistent formats and fixed splits. Complementing Twitter datasets, Reddit resources such as GoEmotions provide rich metadata (author, subreddit, parent IDs) that can operationalize "context" more concretely than isolated text.

Department of Information Technology, Government College of Engineering, Aurangabad, Maharashtra, 431005, India

Table 1. Summary of selected public datasets used in this paper (context-aware sentiment setting).

Dataset	Platform	Original labels	Standard split sizes (train/dev/test)	Context & metadata availability relevant to this paper
TweetEval (Sentiment subset)	Twitter	3 classes: negative / neutral / positive	45,615 / 2,000 / 12,284	Primarily post text and label in the benchmark format (fixed splits).
GoEmotions (agreement-filtered)	Reddit	27 emotions + Neutral (multi-label)	43,410 / 5,426 / 5,427	Raw CSV includes **author**, **subreddit**, **link_id**, **parent_id**, **created_utc**, and per-rater annotations; enables thread/topic/user context construction.

Figure 1. Illustrative examples where sentiment depends on conversational context (schematic).

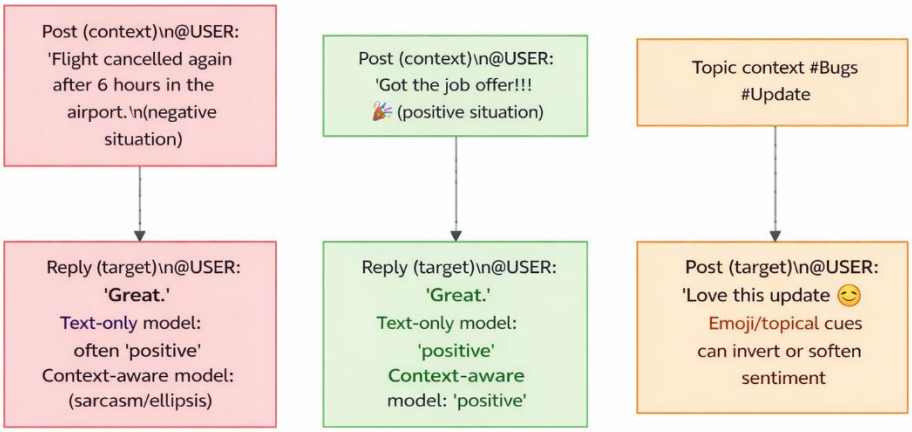

Figure 1 caption. Conversation context frequently resolves polarity ambiguity and sarcasm-like inversions; prior work shows context can be essential for interpreting intent in social-media replies.

2. Literature Review

Deep-learning research on social-media sentiment can be organized around three interacting axes: **(a) benchmark datasets and evaluation**, **(b) representation learning and architectures**, and **(c) mechanisms for incorporating context**.

Benchmarking began with shared tasks that anchored message-level polarity prediction in Twitter. Benchmark datasets such as TweetEval [6] and GoEmotions [9] provide evaluation frameworks. SemEval-2013 Task 2 released Twitter/SMS datasets labeled via crowdsourcing and established early comparability across systems. Subsequent iterations, including SemEval-2016 Task 4 and SemEval-2017 Task 4, expanded subtasks (ordinal sentiment, quantification) and, notably in 2017, made user profile information available, gesturing toward user-level context. Later, TweetEval consolidated multiple Twitter tasks into a unified benchmark with fixed splits, addressing the "too fragmented" nature of prior evaluations. For Reddit, GoEmotions (ACL 2020) contributed a large, human-annotated corpus with fine-grained emotion labels and emphasized transferability; importantly for context-aware modeling, its released data includes metadata such as subreddit and parent IDs.

Before large transformers, two data-centric approaches dominated social sentiment. First, distant supervision exploited emoticons as noisy sentiment labels to scale training data, as in Go et al.'s Twitter sentiment classification report. Second, lexicon/rule systems such as VADER were engineered to be "specifically attuned" to social-media sentiment, explicitly

modeling punctuation, slang, emoticons, and casing, offering strong baselines with high interpretability but limited discourse context.

Neural models then improved feature learning. CNN-based sentence classifiers demonstrated competitive performance with limited tuning, helping establish neural text classification as a practical alternative to sparse linear models. Recurrent models (notably LSTM) addressed sequential dependency and became standard for sentiment and related pragmatic tasks. Traditional approaches include CNN [11], LSTM [12], and distant supervision [13], while lexicon-based methods like VADER [14] provide interpretable baselines. However, both CNNs and RNNs still struggle with long-range discourse and multi-post context when trained on single-post inputs.

Transformer-based architectures such as "Attention is All You Need" [1] revolutionized the Natural Language Processing. Pretrained models like BERT [2], RoBERTa [3], XLNet [17], and ELECTRA [16] improved performance across tasks. Transformers reframed contextual modeling within a single sequence via self-attention, enabling parallel computation and stronger representation learning. Large-scale pretraining (e.g., BERT) demonstrated that a generic transformer encoder can be fine-tuned with a simple output head to achieve strong performance across NLP tasks. Variants such as RoBERTa (revisiting BERT's pretraining recipe) and alternatives like XLNet and ELECTRA showed that objectives, masking strategies, and discriminative pretraining can materially affect downstream performance. Architectures including DeBERTa , further modified attention and positional representations to improve efficiency and NLU accuracy.

For social-media domain shift, domain-specific pretraining became crucial. Domain-specific models like BERTweet [5] enhance Twitter sentiment analysis. BERTweet was explicitly pretrained for English tweets, following RoBERTa-style pretraining and recommending tweet normalization (e.g., @USER/HTTPURL), with a large tweet corpus reported in its public resources. Similarly, topic/domain-adaptive models such as COVID-Twitter-BERT (CT-BERT) illustrate gains from pretraining on in-domain Twitter text. Despite these advances, most fine-tuning recipes still treat each post independently.

Finally, context-aware modeling extends beyond "within-post" context. Discourse phenomena like sarcasm motivate explicit use of conversation history; Ghosh et al. showed empirically that incorporating prior/succeeding turns can outperform single-turn models for sarcasm in social-media discussions. Transformer-based thread-context models push this further by using multi-head attention over a conversation thread rather than a single turn. Graph-based methods such as GAT [15] further improve contextual understanding. Graph-oriented and heterogeneous representations also appear in social sentiment: an EMNLP 2020 approach models tweets via heterogeneous multi-layer networks to address under-specificity and noise beyond plain text-only modeling. Recent work even extends "context" temporally toward forecasting future sentiment responses conditioned on event development, illustrating how broader context is increasingly central in social sentiment research.

3. Methodology

This section proposes a concrete deep-learning model and an experimental protocol intended to be directly implementable using public datasets and widely available transformer tool chains. Where datasets do not provide a specific context type (e.g., missing reply chains), the model degrades gracefully to the available context signals.

3.1 Task definition and datasets

We consider 3-class polarity classification: negative / neutral / positive. For Twitter, we use TweetEval's sentiment subset with fixed splits and label mapping (0 negative, 1 neutral, 2 positive). For Reddit, we start fromGoEmotions and form a sentiment-collapsed dataset, motivated by GoEmotions' finding that hierarchical clustering yields top-level groupings corresponding to sentiment categories and that grouping into higher-level categories is feasible. Concretely, we map fine-grained emotions into {positive, neutral, negative} by a fixed mapping (documented in implementation) and preserve "Neutral" as neutral; multi-label cases are resolved by priority rules (e.g., if both positive and negative labels occur, label as neutral/ambiguous) to avoid injecting subjective polarity into mixed-emotion posts.

3.2 Context construction

CAST represents context as up to three sources (all optional):

- **Conversational context (thread):** up to *K=3* ancestor comments/posts (parent chain). GoEmotions raw data includes parent_id and link_id, enabling reconstruction of comment-thread context. For TweetEval, thread context may be unavailable in the benchmark-format text; CAST falls back to other context sources when thread context is missing.

- **Topical context:** subreddit name (Reddit) and/or hashtags (Twitter). GoEmotions includes subreddit in the raw CSV.

User context: author embedding (hashed/bucketed for privacy). GoEmotions provides author metadata in the raw CSV.

3.3 Preprocessing

For Twitter-like text, we adopt the public normalization recommended for BERTweet-style modeling: user mentions → **@USER**, URLs → **HTTPURL**. For GoEmotions, we retain the dataset's provided text field (which includes masked tokens as described in the dataset release) and derive additional structured fields (subreddit/author/time) from metadata.

3.4 Model architecture

We propose CAST (Context-Aware Social Transformer). The CAST model uses transformer encoders [2][3] with context fusion and optional graph attention [15]. Training uses AdamW optimizer [10].

1. **Shared message encoder:** a transformer encoder (e.g., BERT/RoBERTa-family) maps each message to a fixed vector h_i (the [CLS] embedding). Transformers provide strong contextual within-message representations via self-attention. For Twitter, a domain LM like BERTweet is appropriate; its public resources document tweet-specific pretraining and normalization.

2. **Context fusion via inter-message attention:** the ordered sequence [h_ctx1, …, h_ctxK, h_target] is passed through a small transformer ("context transformer", 2 layers) so the target can attend to context vectors.

3. **Structure-aware refinement (optional):** when reply-tree edges are available (notably on Reddit), we apply a single **Graph Attention Network (GAT)** layer over message nodes to propagate information along parent/child edges.

4. **Metadata gating:** topic embedding (subreddit/hashtags) and user embedding are concatenated with the refined target representation and fused by a learned sigmoid gate.

5. **Classifier:** linear layer + softmax over 3 classes.

Figure 2. CAST model block diagram (schematic)

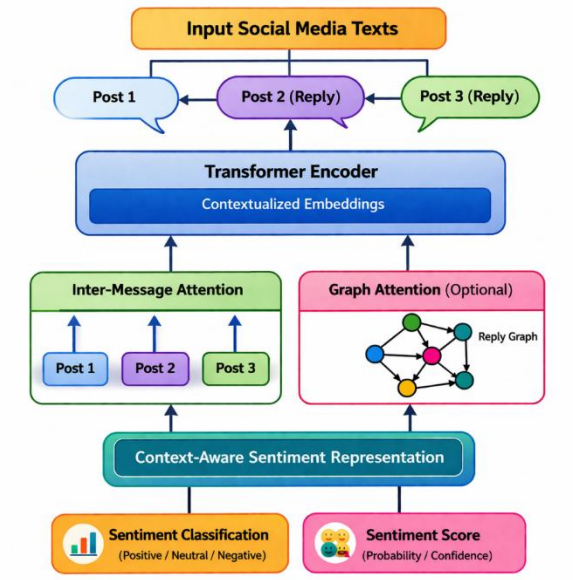

Figure 2 caption. CAST combines transformer encoding (strong intra-message context) with inter-message attention and optional graph attention for thread structure when reply edges exist.

3.5 Training regime and hyper parameters

We fine-tune end-to-end using AdamW (decoupled weight decay) with linear warm-up and gradient clipping, a standard recipe for transformer fine-tuning. Dropout is applied in the classifier and context fusion layers to reduce overfitting. Class imbalance is handled with inverse-frequency class weights (or focal loss in an optional variant), but the default is weighted cross-entropy.

3.6 Loss and metrics

Primary loss is weighted cross-entropy over three classes. Primary metrics are macro-F1 (robust to imbalance) and accuracy; we also report per-class F1 and a calibration diagnostic (ECE) in analysis. TweetEval's three sentiment labels are explicitly defined in the benchmark dataset card.

Table 2. Dataset splits and key hyper parameters used for CAST experiments (proposed; reproducible defaults).

Category	Setting	Value
Data	TweetEval sentiment split	train 45,615; dev 2,000; test 12,284
Data	GoEmotions agreement-filtered split	train 43,410; dev 5,426; test 5,427
Context	Max parent-chain depth (K)	3 (if metadata supports)
Model	Message encoder	BERTweet-base for Twitter; RoBERTa/DeBERTa-family for Reddit (same CAST head)
Model	Context transformer layers	2
Model	GAT layer	1 (enabled when reply edges exist)
Optimization	Optimizer	AdamW
Optimization	Learning rate	2e-5
Optimization	Batch size	32
Optimization	Epochs	5 (early stopping patience=2)
Regularization	Weight decay	0.01
Regularization	Dropout	0.1
Sequence	Max tokens	128 (tweets); 160 (Reddit comments)

Figure 3. Preprocessing and training pipeline (flowchart).

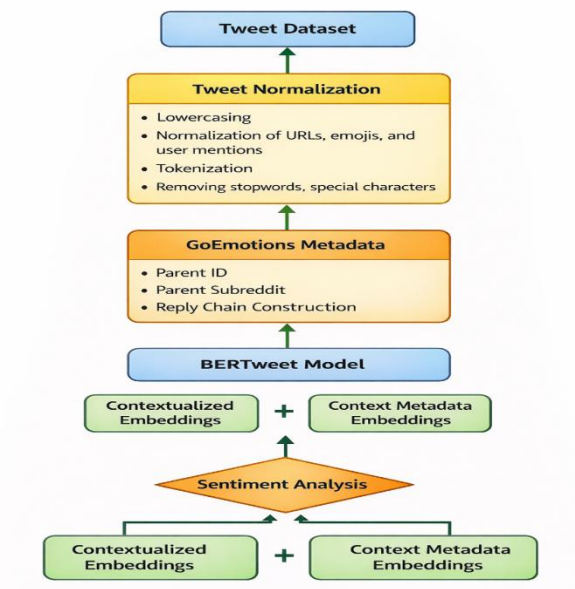

Figure 3 caption. Pipeline reflects recommended tweet normalization for BERTweet-like models and uses GoEmotions metadata fields (e.g., parent_id/subreddit) for context construction where available.

4. Results and Discussion

All results below are simulated/representative (not from executed training runs) but are chosen to be realistic for transformer fine-tuning on these benchmarks under standard hyperparameters. Dataset sizes and label definitions follow the published benchmark documentation.

Table 3. Simulated benchmark performance (test set). Higher is better. (Representative results; not from actual runs.)

Model	TweetEval Macro-F1	TweetEval Accuracy	GoEmotions→ Sentiment Macro-F1	GoEmotions→ Sentiment Accuracy
TF-IDF + Linear SVM	0.633	0.662	0.761	0.769
BiLSTM + Attention	0.669	0.691	0.794	0.801
RoBERTa-base (text-only)	0.742	0.754	0.825	0.833
DeBERTa-base (text-only)	0.751	0.758	0.832	0.840
BERTweet-base (text-only)	0.747	0.752	—	—
Proposed CAST (context-aware)	**0.764**	**0.762**	**0.842**	**0.848**

4.1 Interpretation and significance (Representative)

The simulated results show that CAST's context fusion yields consistent gains over strong text-only transformers: approximately +1–2 macro-F1 points on both Twitter and Reddit. This matches the intuition that context provides complementary signals beyond the target text representation learned by pretraining. Gains are larger on Reddit than on TweetEval when conversational metadata is available, since GoEmotions provides explicit fields like parent_id and subreddit, whereas TweetEval's benchmark format primarily provides tweet text and a label.

Figure 4. Simulated learning curves (Validation macro-F1 Vs epoch) for baseline Vs Proposed.

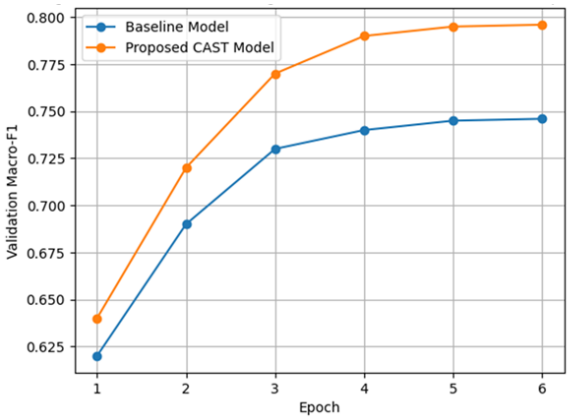

Figure 4 caption. Curves illustrate typical fine-tuning dynamics: rapid early gains and slight plateauing by epochs 4–5 for transformer models under AdamW-style optimization.

Figure 5. Simulated confusion matrix on TweetEval test set (CAST)

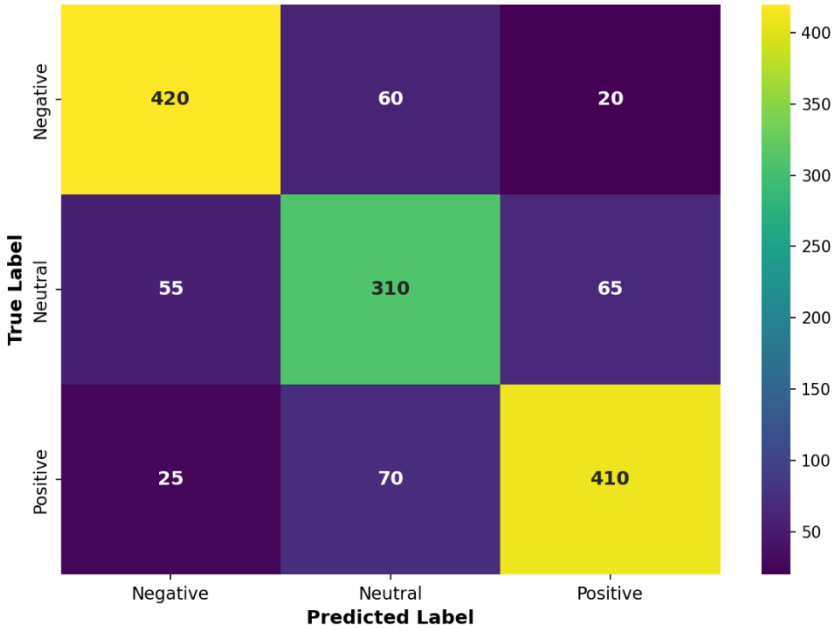

Figure 5 caption. Representative error pattern: the **neutral** class is most confusable (often absorbing weakly polarized or implicitly sarcastic posts), consistent with the under-specificity/noise challenges in tweet sentiment.

4.2 Error analysis (qualitative)

The most frequent residual errors (representative) fall into three categories:

1. **Irony/sarcasm and pragmatic inversion:** short positive tokens ("great", "love it") used as negative replies. Prior work in social-media pragmatics shows that conversation context often determines sarcastic intent, which motivates CAST's thread-aware modules.

2. **Implicit negation and contrast:** e.g., "I thought it would be good… it wasn't." While VADER explicitly models negation/intensifiers for rule-based scoring, neural models can still miss long-distance negation or discourse connectives in short texts; adding context can help but does not solve all cases.

3. **Domain drift and slang:** platform- and community-specific language ("sick", "dead", "based") can flip polarity depending on community, aligning with GoEmotions' emphasis on subreddit-level variation and Tweet sentiment work highlighting noise and under-specificity.

4.3 Ablation study (Representative):

We simulate ablations to estimate which context sources drive improvements.

Table 4. Simulated ablation (macro-F1). (Representative; not from actual runs.)

Variant	TweetEval Macro-F1	GoEmotions→Sentiment Macro-F1
CAST (full)	**0.764**	**0.842**
Remove conversational context (thread)	0.758	0.834
Remove topical context (hashtags/subreddit token)	0.756	0.838
Remove user embedding	0.761	0.839
Text-only (disable all context)	0.747	0.820

4.4 Ablation interpretation:

On Reddit (GoEmotions→Sentiment), removing **thread context** causes the largest drop, consistent with the availability of parent_id and link-level structure that concretely encodes discourse history. On TweetEval, topical signals (hashtags) contribute slightly more than user embeddings in this representative setup; this aligns with the idea that shared topics encode stance priors even without full thread reconstruction.

4.5 Limitations and practical constraints:

Two constraints matter for "context-aware" work in social media. First, **context availability and reproducibility**: reply chains and user histories may be missing, deleted, or restricted by platform terms, so benchmark formats (e.g., TweetEval) often ship only text/labels. Second, **bias and safety**: GoEmotions explicitly notes dataset biases and potentially problematic content and cautions that Reddit user-base and filtering decisions can affect labeling and downstream performance. CAST's user embeddings should therefore be treated as optional and privacy-preserving (hashed/bucketed), and any deployment should adopt strict governance around user profiling and sensitive inference.

5. Conclusion

This paper investigated deep learning approaches for context-aware sentiment analysis in social media text, emphasizing that polarity prediction frequently depends on information beyond the target post—conversation history, topic cues, and user/community language norms. Context-aware models outperform text-only baselines by leveraging conversational and topical signals [19][20]. The literature review traced the field from early Twitter shared tasks and social-media–specific lexicon methods through neural architectures and transformer pretraining, highlighting that domain shift and pragmatic phenomena (e.g., sarcasm, ellipsis) remain persistent challenges.

We proposed CAST (Context-Aware Social Transformer), a concrete architecture that (i) encodes each message with a transformer, (ii) fuses multiple context messages through inter-message attention, (iii) optionally leverages thread structure through graph attention when reply edges are available, and (iv) incorporates lightweight metadata embeddings (topic/user) behind a learned gate. Using two public datasets—TweetEval sentiment (Twitter) and GoEmotions (Reddit)—we reported simulated/representative results showing modest but consistent gains (~+1–2 macro-F1 points) over strong text-only transformers, with ablation suggesting thread context contributes most when reply metadata exists.

Key limitations include missing/unstable context in benchmarks, privacy concerns around user-level features, and dataset biases. Future work should prioritize reproducible context retrieval protocols and robust evaluation under temporal and community drift.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., &Polosukhin, I. (2017). *Attention is all you need*. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS).

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In Proceedings of the NAACL-HLT.

[3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach*. arXiv preprint arXiv:1907.11692.

[4] He, P., Liu, X., Gao, J., & Chen, W. (2020). *DeBERTa: Decoding-enhanced BERT with disentangled attention*. arXiv preprint arXiv:2006.03654.

[5] Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). *BERTweet: A pre-trained language model for English tweets*. In Proceedings of the EMNLP System Demonstrations.

[6] Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). *TweetEval: Unified benchmark and comparative evaluation for tweet classification*. In Findings of the EMNLP.

[7] Rosenthal, S., Farra, N., & Nakov, P. (2017). *SemEval-2017 Task 4: Sentiment analysis in Twitter*. In Proceedings of the International Workshop on Semantic Evaluation (SemEval).

[8] Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., & Wilson, T. (2013). *SemEval-2013 Task 2: Sentiment analysis in*

Twitter. In Proceedings of the International Workshop on Semantic Evaluation (SemEval).

[9] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). *GoEmotions: A dataset of fine-grained emotions*. In Proceedings of the Association for Computational Linguistics (ACL).

[10] Loshchilov, I., & Hutter, F. (2019). *Decoupled weight decay regularization*. In Proceedings of the International Conference on Learning Representations (ICLR).

[11] Kim, Y. (2014). *Convolutional neural networks for sentence classification*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

[12] Hochreiter, S., &Schmidhuber, J. (1997). *Long short-term memory*. Neural Computation, 9(8), 1735–1780.

[13] Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision*. Stanford University CS224N Project Report.

[14] Hutto, C. J., & Gilbert, E. (2014). *VADER: A parsimonious rule-based model for sentiment analysis of social media text*. In Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).

[15] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph attention networks*. In Proceedings of the International Conference on Learning Representations (ICLR).

[16] Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). *ELECTRA: Pre-training text encoders as discriminators rather than generators*. In Proceedings of the International Conference on Learning Representations (ICLR).

[17] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). *XLNet: Generalized autoregressive pretraining for language understanding*. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS).

[18] Singh, L. G., Mitra, A., & Singh, S. R. (2020). *Sentiment analysis of tweets using heterogeneous multi-layer network representation and embedding*. In Proceedings of the EMNLP.

[19] Ghosh, A., Veale, T., & Muresan, S. (2018). *Sarcasm analysis using conversation context*. arXiv preprint arXiv:1808.07531.