Sentiment-Aware Personality Prediction: A Hybrid RCNN-RoBERTa Approach for MBTI Classification from Social Media Text

Xinni Song
Department of Computer Science
The Australian National University
u7439250@anu.edu.au

Abstract

Personality prediction from social media text presents significant challenges in capturing sentiment-driven linguistic patterns that characterize different Myers-Briggs Type Indicator (MBTI) personality types. This paper introduces PRCNN-RoBERTa, a sentiment-aware hybrid architecture that combines pre-trained RoBERTa with recurrent convolutional neural networks for enhanced personality classification. Our approach integrates a frozen RoBERTa encoder with bidirectional LSTM and convolutional operations to capture both global semantic context and local sequential dependencies. We evaluate our method on the MBTI dataset containing 8,675 PersonalityCafe forum users, comparing against traditional machine learning (TF-IDF+SVM, KNN+RF), deep learning (CNN+Embeddings), and transformer baselines (BERT, pure RoBERTa). Our PRCNN-RoBERTa stands out by delivering the highest exact-match accuracy overall and, in particular, achieving the best-in-class scores on the nuanced T/F and E/I dimensions while matching or surpassing transformer baselines on S/N. The model achieves 58.99% exact-match accuracy and 77.74% macro F1-score, outperforming pure RoBERTa by 1.06% in accuracy and 1.63% in F1-score. These results demonstrate that explicitly modeling sentiment-aware patterns through hybrid architectures effectively enhances personality prediction while maintaining computational efficiency. The code is publicly available at `https://github.com/echosongg/sentiment-detection-MBTI-classification`.

1 Introduction

Myers–Briggs Type Indicator (MBTI), based on Jung's theory of psychological types, is a widely used personality assessment model applied in career guidance, team building, and interpersonal research [1]. MBTI categorizes individual preferences along four dichotomies: Introversion (I) vs. Extraversion (E), Intuition (N) vs. Sensing (S), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P), yielding 16 distinct personality types that describe differences in information processing, decision-making, and behavior [2].

Early efforts in automated MBTI prediction employed shallow machine-learning techniques on bag-of-words or TF-IDF representations. For instance, Mushtaq *et al.* combined K-means clustering with XGBoost to achieve 85%–90% accuracy across the four dichotomies [3], while Ryan *et al.* balanced classes using SMOTE and compared logistic regression, linear SVC, SGD, Random Forest, XGBoost, and CatBoost, finding logistic regression yielded the highest F1-scores on the Kaggle MBTI dataset [4].

The advent of deep learning brought further improvements. Ahmad *et al.* fused Word2Vec embeddings with a CNN+LSTM architecture on forum posts to reach about 91% precision [5], and Alsini

et al. showed that combining TF-IDF, Word2Vec/GloVe, and sentence-level Bi-LSTM features attains 91.57% accuracy for the Agreeableness trait [6]. More recently, transformer-based models such as BERT and RoBERTa have set new state-of-the-art results by capturing richer contextual semantics [7].

Nevertheless, two key challenges remain. First, social media users often employ figurative language—especially sarcasm and irony—whose surface positivity masks critical intent, leading literal classifiers to misinterpret sentiment. Second, each MBTI dimension correlates with distinct linguistic styles, emotional tones, and topical interests, demanding explicit modeling of rhetorical and affective cues to learn truly robust representations.

In our MBTI forum dataset [8], we observe many sarcastic expressions that illustrate these difficulties, for example:

- "Tik Tok is a really great song. As long as you can mental block out the singer."
- "Internet IQ tests are funny. I score 140s or higher. Now, like the former responses of this thread I will mention that I don't believe in the IQ test."

To tackle such non-literal cues, Potamias *et al.* proposed a compact RCNN-RoBERTa: RoBERTa embeddings are refined by a Bi-LSTM, pooled through a wide receptive-field layer, and finally classified [9]. Their hybrid encoder outperformed pure Transformers on benchmark sarcasm datasets, suggesting that a light recurrent–convolutional stage can sharpen figurative-language detection.

Building on this insight, we adapt their RCNN-RoBERTa pipeline to develop PRCNN-RoBERTa (Personality-oriented Recurrent CNN RoBERTa) for multi-label MBTI typing and integrate a sentiment-aware objective; details follow in Section 3.

2 Problem Definition and Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the MBTI corpus, where each x_i is a forum post and $y_i = [y_i^{\text{EI}}, y_i^{\text{SN}}, y_i^{\text{TF}}, y_i^{\text{JP}}] \in \{0,1\}^4$ encodes the four MBTI dichotomies (1 for *E,N,T,J*, 0 for *I,S,F,P*). After byte-pair tokenisation with the RoBERTa vocabulary, every post is mapped to a sequence $\mathbf{w}_i = (w_{i,1}, \ldots, w_{i,L})$ with $L \leq 512$ tokens.

Formally, our goal is to learn a parametrised mapping $f_\theta : \mathcal{V}^{\leq L} \to [0,1]^4$ that outputs $\hat{y}_i = f_\theta(x_i; \theta)$ such that $\hat{y}_i$ approximates y_i in all four coordinates. Training minimises the mean *binary cross-entropy* (BCE)

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{4} \Big[y_i^{(k)} \log \hat{y}_i^{(k)} + \big(1 - y_i^{(k)}\big) \log\big(1 - \hat{y}_i^{(k)}\big) \Big]. \tag{1}$$

3 Method

Our proposed PRCNN-RoBERTa model combines RoBERTa's contextual understanding with recurrent convolutional neural networks' sequential pattern recognition to capture personality-indicative linguistic patterns in social media text. The architecture uses a frozen pre-trained RoBERTa encoder as feature extractor, followed by bidirectional LSTM and convolutional-like operations to model local and global text dependencies for personality trait detection (See Fig 1).

3.1 Model Architecture

PRCNN-RoBERTa processes input through a multi-stage pipeline. A frozen RoBERTa-base encoder converts byte-pair encoded sequences into contextual representations $\mathbf{H}^{(0)} \in R^{L \times 768}$, where L is sequence length. These embeddings are processed by bidirectional LSTM (hidden size 128 per direction), producing enhanced sequential representations $\mathbf{H}^{(1)} \in R^{L \times 256}$. The key innovation is the fusion mechanism concatenating RoBERTa and LSTM outputs at each time step, followed by global max pooling:

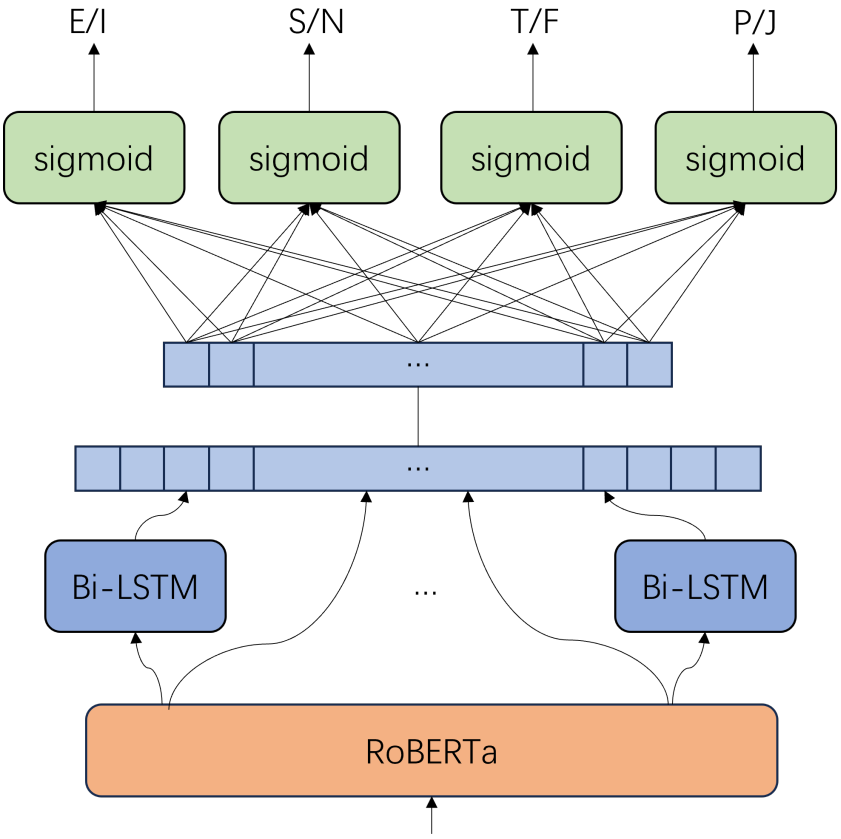

Figure 1: Architecture of our PRCNN-RoBERTa model for MBTI classification.

$$\mathbf{h}_t^{fused} = [\mathbf{h}_t^{roberta}; \mathbf{h}_t^{lstm}], \quad \mathbf{p} = \max_{t=1}^{L} \mathbf{h}_t^{fused} \tag{2}$$

where $\mathbf{p} \in R^{1024}$ is the pooled feature vector capturing both semantic understanding from RoBERTa and sequential dependencies from LSTM. Pooled features are processed through a fully connected layer with ReLU activation, simulating large-kernel 1D convolution:

$$\mathbf{z} = \text{ReLU}(\mathbf{W}_f \mathbf{p} + \mathbf{b}_f) \tag{3}$$

where $\mathbf{W}_f \in R^{100\times1024}$ and $\mathbf{b}_f \in R^{100}$. Four parallel sigmoid units generate independent predictions for each MBTI dimension:

$$\hat{\mathbf{y}} = \sigma(\mathbf{W}_o \mathbf{z} + \mathbf{b}_o) \tag{4}$$

where $\mathbf{W}_o \in R^{4\times100}$, $\mathbf{b}_o \in R^4$, and σ is sigmoid activation. This multi-output architecture enables simultaneous prediction of all four MBTI dimensions efficiently.

Computational complexity is dominated by RoBERTa encoder at $\mathcal{O}(L^2 d)$ for self-attention, where $d = 768$. Since RoBERTa weights remain frozen, only 1.4M parameters (LSTM, fully connected, and output layers) require gradient updates, reducing memory requirements for training on standard 8GB GPUs. The LSTM adds $\mathcal{O}(Ld^2)$ complexity, linear in sequence length beyond the encoder, enabling scalability to longer sequences.

4 Experiments

We conduct comprehensive experiments evaluating our PRCNN-RoBERTa model against established baselines across different text classification paradigms. Our experimental design encompasses traditional machine learning, modern deep learning, and state-of-the-art transformer models for thorough performance comparison.

4.1 Dataset and Preprocessing

Experiments utilize the Myers-Briggs Type Indicator (MBTI) dataset from PersonalityCafe forum posts, containing 8,675 users with self-reported personality types and associated text posts. Each sample includes concatenated forum posts per user. The dataset exhibits natural class imbalance reflecting real-world distributions: Intuitive (N) types comprise 86% of samples and Thinking (T) types represent 77% of users.

We apply minimal preprocessing to preserve authentic linguistic patterns: (1) lowercasing, (2) URL removal using regex `http\S+`, (3) whitespace consolidation, and (4) forum markup removal (`|||`) while preserving punctuation and emoticons. For transformer models, we use respective tokenizers (BERT WordPiece, RoBERTa BPE) with consistent maximum sequence length truncation.

4.2 Baseline Models and Configurations

We benchmark against five representative systems spanning three modeling paradigms.

4.2.1 Traditional Machine Learning Methods

TF-IDF + SVM: Employs term frequency-inverse document frequency with unigram/bigram features, limited to 5,000 features:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right) \tag{5}$$

where t is a term, d a document, N total documents, and D the document collection. MultiOutputClassifier wraps LinearSVC with decision function for MBTI dimension c:

$$f_c(\mathbf{v}) = \sigma(\mathbf{w}_c^T \mathbf{v} + b_c) \tag{6}$$

where $\mathbf{v} \in R^{5000}$ is TF-IDF features, $\mathbf{w}_c$ and b_c are learned parameters, and σ is sigmoid activation.

KNN + Random-Forest Ensemble. All documents are first vectorised in the identical TF–IDF space. For any test sample, a *K*-Nearest-Neighbours component ($k = 5$, cosine distance) retrieves the five closest training points and outputs, for each MBTI bit, the empirical neighbour proportion $p_c^{\text{KNN}} = \frac{1}{k}\sum_{j=1}^{k} 1[y_j^{(c)} = 1]$. In parallel, a 100-tree Random Forest produces probability estimates p_c^{RF} based on global decision-tree voting. The final posterior for class c is obtained by soft voting, $p_c = \frac{1}{2}\big(p_c^{\text{KNN}} + p_c^{\text{RF}}\big)$, thereby blending KNN's local pattern sensitivity with the feature-aggregation strength of the Random Forest.

4.2.2 Deep Learning Sequence Models

CNN: Uses 100-dimensional embeddings with filter sizes $\{3, 4, 5\}$ and 100 filters per size. For input $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$:

$$c_i = \text{ReLU}(\mathbf{W}_h \cdot \mathbf{x}_{i:i+h-1} + b_h) \tag{7}$$

where $\mathbf{W}_h \in R^{d \times (h \cdot d)}$ is filter matrix for window size h. Global max pooling extracts important features:

$$\hat{c}_h = \max(c_1, c_2, \ldots, c_{T-h+1}) \tag{8}$$

4.2.3 Transformer-Based Baselines

BERT-base: Fine-tunes `bert-base-uncased` (110M parameters). [CLS] token's final hidden state serves as sentence representation:

$$\mathbf{h}_{\text{[CLS]}} = \text{BERT}(\text{[CLS]}, x_1, x_2, \ldots, x_T, \text{[SEP]})_{[0]} \tag{9}$$

This 768-dimensional representation passes through 100-dimensional FC layer with ReLU, then four sigmoid units.

Pure RoBERTa: Fine-tunes `roberta-base` (125M parameters) following identical protocol. Uses first token's final hidden state as sequence representation through same FC+sigmoid architecture.

4.3 Evaluation Metrics

We employ comprehensive metrics for multi-label classification. For $C = 4$ MBTI dimensions and N test instances, with TP_c, FP_c, FN_c representing true/false positives and false negatives for dimension c:

Macro-averaged Precision: Unweighted average of per-class precision:

$$\text{Precision}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c} \tag{10}$$

Macro-averaged Recall: Unweighted average of per-class recall:

$$\text{Recall}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FN_c} \tag{11}$$

Macro-averaged F1-Score: Harmonic combination of macro precision and recall:

$$F1_{\text{macro}} = \frac{2 \cdot \text{Precision}_{\text{macro}} \cdot \text{Recall}_{\text{macro}}}{\text{Precision}_{\text{macro}} + \text{Recall}_{\text{macro}}} \tag{12}$$

Exact-Match Accuracy: Requires correct prediction across all four dimensions:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\prod_{c=1}^{C}\mathbf{1}[\hat{y}_{ic} = y_{ic}] \tag{13}$$

where $\mathbf{1}[\cdot]$ is indicator function, $\hat{y}_{ic}$ and y_{ic} are predicted and true labels.

Per-Dimension Metrics: Report precision, recall, and F1-score for each MBTI dimension (E/I, S/N, T/F, J/P).

4.4 Training Configuration

For fair comparison, we maintain consistent configurations: 512-token maximum sequence length, batch size 8, stratified 70%/15%/15% train/validation/test splits. Neural models use AdamW optimization with weight decay (1×10^{-5}), linear learning rate scheduling, gradient clipping (max norm: 1.0), and early stopping (patience: 3 epochs). Multi-label classification uses Binary Cross-Entropy with Logits loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}[y_{ic}\log(\sigma(z_{ic})) + (1 - y_{ic})\log(1 - \sigma(z_{ic}))] \tag{14}$$

where z_{ic} is logit output for instance i and class c.

All experiments use PyTorch with reproducible seeds (42) on NVIDIA RTX 4060 GPUs. We optimize hyperparameters for PRCNN-RoBERTa: batch size $\{8, 16\}$, sequence length $\{256, 384\}$, learning rate $\{1\times10^{-5}, 2\times10^{-5}\}$, LSTM hidden size $\{64, 128\}$, FC layer size $\{100, 200\}$, dropout rate $\{0.1, 0.2\}$ using validation performance. Final configuration uses batch size 16, sequence length 512, learning rate 2×10^{-5}, LSTM hidden size 128, FC layer size 100, and dropout rate 0.1, balancing performance with 8GB GPU memory constraints.

4.5 Experimental Results and Analysis

Our experimental evaluation demonstrates that PRCNN-RoBERTa achieves superior performance across multiple metrics on the MBTI personality prediction task. As shown in Table 1, our model attains the highest overall accuracy (0.5899), F1-score (0.7774), and recall (0.7483), while maintaining competitive precision and training efficiency. Compared to pure RoBERTa, our hybrid architecture achieves a 1.06% improvement in accuracy and 1.63% gain in F1-score, demonstrating the effectiveness of incorporating recurrent convolutional components for capturing personality-indicative linguistic patterns.

The training efficiency analysis reveals that PRCNN-RoBERTa maintains computational practicality with 3,615 seconds training time, comparable to BERT-base and more efficient than pure RoBERTa. This efficiency stems from freezing RoBERTa encoder weights and training only the additional RCNN components (approximately 1.4M parameters). Traditional approaches like TF-IDF+SVM, while computationally efficient (77 seconds), achieve substantially lower exact-match accuracy (0.5269), highlighting the importance of deep contextual understanding for personality prediction.

The dimension-wise analysis (Tables 2–5) reveals that PRCNN-RoBERTa consistently achieves the highest accuracy in three of four MBTI dimensions: E/I (0.8556), S/N (0.8978), and T/F (0.8295). The model demonstrates particularly strong performance on the challenging E/I dimension with the highest recall (0.565) and F1-score (0.643), effectively identifying minority class patterns. For the S/N dimension, exceptional performance across all models reflects the natural class imbalance, while the T/F dichotomy shows our model's strength in capturing emotional versus analytical language patterns.

Transformer-based models consistently outperform traditional approaches, confirming the importance of pre-trained contextual representations. However, the KNN+RF ensemble exhibits interesting complementary strengths, achieving the highest precision in several dimensions while our unified PRCNN-RoBERTa architecture delivers more consistent overall performance. These results collectively demonstrate that PRCNN-RoBERTa successfully leverages complementary strengths of transformer-based contextual understanding and recurrent convolutional pattern recognition for the challenging multi-label MBTI prediction task.

Table 1: Overall performance of all models on the MBTI test split. Bold: best, *: second-best.

Model	Train (s)	Accuracy	Precision	Recall	F1
BERT-base	3 611	0.5730	0.8038	0.7226	0.7508
Pure RoBERTa	3 792	0.5837*	0.8249*	0.7223	0.7649*
PRCNN-RoBERTa (ours)	3 615	**0.5899**	0.8159	**0.7483**	**0.7774**
TF–IDF + SVM	77*	0.5269	0.8000	0.7416*	0.7645
CNN + Embeddings	**20**	0.4416	0.7557	0.5986	0.6425
KNN+RF Ensemble	146	0.5661	**0.8399**	0.7002	0.7511

5 Conclusion

This work proposed PRCNN-RoBERTa, a hybrid architecture combining pre-trained transformers with recurrent convolutional neural networks for MBTI personality prediction from social media text. Our approach was motivated by the need to capture both global semantic understanding and local sequential patterns that characterize personality-indicative linguistic features.

Experimental results demonstrate the effectiveness of our method, achieving 58.99% exact-match accuracy and 77.74% macro F1-score, outperforming both traditional machine learning approaches and modern transformer baselines. The consistent performance across three MBTI dimensions (E/I, S/N, T/F) validates our hypothesis that integrating RCNN components with pre-trained language models better captures personality-relevant patterns.

These results make sense as personality expression involves both explicit content and implicit structural patterns, which our hybrid architecture addresses through complementary pathways. The frozen RoBERTa encoder provides rich contextual representations while trainable RCNN compo-

Table 2: E/I dichotomy. Bold: best, *: second-best.

Model	Acc	P	R	F1
BERT-base	0.8372	0.764	0.421	0.543
Pure RoBERTa	0.8533*	0.776*	0.508	0.614
PRCNN-R.	**0.8556**	0.745	**0.565**	**0.643**
TF–IDF+SVM	0.8525	0.756	0.528*	0.622*
CNN+Embed	0.8026	0.691	0.254	0.372
KNN+RF	0.8402	**0.794**	0.411	0.542

Table 3: S/N dichotomy. Bold: best, *: second-best.

Model	Acc	P	R	F1
BERT-base	0.8925	**0.932**	0.944	0.938
Pure RoBERTa	0.8963*	0.919	0.964	0.941*
PRCNN-R.	**0.8978**	0.929*	0.954	**0.941**
TF–IDF+SVM	0.8932	0.904	0.979	0.940
CNN+Embed	0.8671	0.874	0.988*	0.928
KNN+RF	0.8932	0.895	**0.993**	0.941*

Table 4: T/F dichotomy. Bold: best, *: second-best.

Model	Acc	P	R	F1
BERT-base	0.8264	0.813	0.810	0.812
Pure RoBERTa	0.8180	0.831*	0.760	0.794
PRCNN-R.	**0.8295**	0.809	**0.825**	**0.817**
TF–IDF+SVM	0.8272*	0.806	0.824*	0.815*
CNN+Embed	0.7488	0.736	0.711	0.723
KNN+RF	0.8134	**0.864**	0.707	0.778

Table 5: J/P dichotomy. Bold: best, *: second-best.

Model	Acc	P	R	F1
BERT-base	0.7711	0.707	0.715*	0.711*
Pure RoBERTa	0.7896	0.774	0.656	0.710
PRCNN-R.	0.7903*	0.781*	0.648	0.709
TF–IDF+SVM	0.7657	0.734	0.635	0.681
CNN+Embed	0.7135	0.722	0.441	0.548
KNN+RF	**0.8134**	**0.808**	**0.689**	**0.744**

nents enable task-specific adaptation. However, limitations on the J/P dimension suggest that ensemble approaches may be more effective for certain personality traits, and the modest overall improvements indicate that personality prediction remains inherently challenging.

The practical implications extend to computational psychology, personalized recommendation systems, and human-computer interaction. Our findings suggest that personality prediction benefits from architectures modeling both global context and local patterns, though the task's complexity calls for continued research into more sophisticated approaches and richer behavioral data sources.

References

[1] Isabel Briggs Myers. *A guide to the development and use of the Myers-Briggs type indicator: Manual*. Consulting Psychologists Press, 1985.

[2] Isabel Briggs Myers. *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Cpp, 2003.

[3] Zeeshan Mushtaq, Sagar Ashraf, and Nosheen Sabahat. Predicting mbti personality type with k-means clustering and gradient boosting. In *2020 IEEE 23rd International Multitopic Conference (INMIC)*, pages 1–5, 2020.

[4] Gregorius Ryan, Pricillia Katarina, and Derwin Suhartono. Mbti personality prediction using machine learning and smote for balancing data based on statement sentences. *Information*, 14(4):217, 2023.

[5] Hussain Ahmad, Muhammad Usama Asghar, Muhammad Zubair Asghar, Aurangzeb Khan, and Amir H Mosavi. A hybrid deep learning technique for personality trait classification from text. *IEEE Access*, 9:146214–146232, 2021.

[6] Raed Alsini, Anam Naz, Hikmat Ullah Khan, Amal Bukhari, Ali Daud, and Muhammad Ramzan. Using deep learning and word embeddings for predicting human agreeableness behavior. *Scientific Reports*, 14(1):29875, 2024.

[7] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40, 2021.

[8] Kaggle. Myers–Briggs Personality Type Dataset, 2017. Accessed: 1 May 2025.

[9] Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320, 2020.