Comparative Analysis of BERT and TF-IDF+MLP Models for Humor Detection on the HaHackathon Dataset: A Comprehensive Statistical Study

Saim Kaan Simsek
Department of Digital Humanities
University of Cologne
Cologne, Germany
ssimsek@smail.uni-koeln.de

***Abstract*—Detecting humor in text is a difficult task in natural language processing, mainly because humor can be very subjective and depend on context. In this paper, I compare two different methods for automatically identifying humor: a modern transformer-based model called BERT and a more traditional approach using TF-IDF with an MLP classifier. I tested these on the HaHackathon dataset from SemEval-2021 Task 7, which includes labels for both humor and offense from people in various age groups. To ensure reliable results, I ran experiments with five different random seeds. The BERT model performed much better overall, with gains of 12.22% in accuracy (0.9272 $\pm$ 0.0107 vs. 0.8262 $\pm$ 0.0110), 9.04% in F1-score (0.9413 $\pm$ 0.0085 vs. 0.8633 $\pm$ 0.0066), and 8.66% in AUC (0.9770 $\pm$ 0.0045 vs. 0.8992 $\pm$ 0.0088). Statistical tests confirmed these improvements are significant, with large effect sizes (Cohen's $d > 8.0$) and p-values less than 0.001. Through a detailed look at the errors, I found that BERT handles contextual and ironic humor well, while the TF-IDF+MLP method struggles with ambiguous meanings and dependencies between words that are key to understanding jokes.**

***Index Terms*—natural language processing, humor detection, BERT, TF-IDF, multi-layer perceptron, text classification, HaHackathon, statistical analysis, error analysis**

I. Introduction

Humor detection stands out as a challenging area in natural language processing. It requires models to pick up on subtle language cues, cultural references, and situational context that make something funny. Being able to detect humor automatically has practical uses, such as in content moderation systems, recommendation engines, or even improving interactions between humans and computers [1]. Traditional methods often relied on hand-crafted rules or basic features, but recent advances in deep learning, especially with transformer models, have shown better results [2].

The SemEval-2021 Task 7, known as HaHackathon, marked an important step by combining humor detection with offense detection in one dataset. This task provided a collection of texts labeled by diverse groups, helping to capture how humor can vary based on personal backgrounds [3].

In this work, I examine two approaches on the HaHackathon dataset: BERT, which uses transformers to understand context deeply, and TF-IDF paired with an MLP classifier, a more straightforward statistical method. BERT excels at grasping bidirectional relationships in text [1], while TF-IDF focuses on word importance across documents and feeds into a neural network for decisions [5].

My contributions include: (1) a solid statistical comparison using multiple random seeds for consistency, (2) an in-depth error analysis to spot common weaknesses and language patterns, (3) clear measurements of performance gains with effect sizes, and (4) discussions on what these models reveal about detecting humor in both old and new ways.

The paper is structured as follows: Section II reviews previous research. Section III details the dataset, models, and experimental design. Section IV presents the findings with statistical breakdowns. Section V explores errors and implications. Finally, Section VI summarizes and suggests future directions.

II. Related Work

Research on humor detection has progressed steadily in natural language processing, starting from basic rule systems to advanced neural networks. Initial efforts concentrated on spotting specific humor elements, like inconsistencies, ambiguities, or clever word uses [8].

In earlier machine learning, experts created features inspired by humor theories, such as word frequencies, grammar structures, or semantic shifts [9]. These features trained standard algorithms like Support Vector Machines or Naive Bayes [1].

Competitions have driven progress. For instance, SemEval 2017 included a humor challenge using data from a comedy show [3]. The HAHA challenges in 2018 and 2019 dealt with Spanish humor from tweets [3]. Over time, methods evolved from simple features and recurrent networks to pre-trained models like BERT [2].

Deep learning introduced tools like CNNs and LSTMs to identify patterns in sequences [4]. These often relied on word vectors from systems like Word2Vec [5].

Lately, transformer models such as BERT have transformed humor detection by better handling context and nuances [6]. In the HaHackathon task, most teams opted for these models, adding techniques like adaptive training [3].

Even so, traditional TF-IDF with neural classifiers remains useful as baselines or in settings with limited resources. Comparing them highlights balances between model depth, computing needs, and effectiveness [7].

III. Methodology

A. Dataset: HaHackathon (SemEval-2021 Task 7)

The HaHackathon dataset from SemEval-2021 Task 7 serves as the foundation for my experiments. This dataset uniquely blends humor and offense detection, reflecting how these aspects can intersect and vary by demographic [3].

1) Dataset Composition and Sources: It includes 10,000 English texts, with 80% from Twitter for natural humor and 20% from Kaggle Short Jokes for structured ones. This mix ensures a good representation of different humor styles [3].

Twitter sources covered both funny and serious accounts, with filters to remove irrelevant elements like links or retweets. The Kaggle part drew from Reddit jokes, adding classic formats.

2) Annotation Methodology and Quality Control: Labels came from 20 annotators per text, balanced across age brackets via Prolific [3].

Annotators answered on humor presence, offense levels, and personal reactions, rating positives from 1 to 5. Quality checks used anchor texts, replacing poor annotations to maintain standards [3].

3) Dataset Statistics and Label Assignment: Final labels used majority votes for binaries and averages for scores. There are 6,179 humorous entries (61.8%, average rating 2.24) and 3,821 non-humorous. Also, 3,052 are controversial based on rating spread, and 5,754 offensive (average 1.02).

Agreement scores: strong for humor binary (0.736), fair for offense (0.518), lower for ratings (0.124) due to subjectivity. The split is 80:10:10, keeping classes even [3].

B. Model Architectures

1) BERT Model: I chose BERT-base-uncased, featuring 12 layers, 12 heads, and 768 dimensions, fine-tuned for classifying sequences [1]:

- Sequence length max: 128 tokens
- Batch size: 16
- Epochs: 3
- Learning rate: 2×10^{-5}
- Weight decay: 0.01

Processing involves tokenizing to WordPieces, adding special tokens, and using the [CLS] output for classification [2].

2) TF-IDF+MLP Model: This pipeline vectorizes with TF-IDF then classifies via MLP [4]:

- TF-IDF: Up to 5,000 features, n-grams 1-3, English stops removed
- Scaling: MaxAbsScaler
- MLP: Hidden sizes (512, 128), ReLU, early stopping

TF-IDF creates a matrix emphasizing important terms relative to the collection [5].

C. Experimental Setup and Statistical Analysis

Experiments ran across five seeds (42, 123, 456, 789, 1337) for reproducibility, using the standard 80:10:10 stratified split [3].

Analysis involved:

- Bootstrap intervals: 1,000 samples per metric
- Paired t-tests for seed comparisons
- Cohen's d for effect magnitude
- McNemar's for paired outcomes
- Bonferroni to adjust for multiples

Evaluated with accuracy, F1, AUC, plus error reviews via matrices and samples [6].

IV. Results and Statistical Analysis

A. Overall Performance Comparison

Across the five seeds, BERT clearly outperformed TF-IDF+MLP in all areas. Table I details the means, variations, intervals, and effects.

TABLE I: Statistical Performance Comparison Across Five Random Seeds

Model	Metric	Mean ± SD	95% CI	Cohen's d
BERT	Accuracy	0.9272 ± 0.0107	[0.9178, 0.9367]	8.33
	F1-Score	0.9413 ± 0.0085	[0.9338, 0.9487]	
	AUC	0.9770 ± 0.0045	[0.9731, 0.9809]	
TF-IDF+MLP	Accuracy	0.8262 ± 0.0110	[0.8166, 0.8359]	–
	F1-Score	0.8633 ± 0.0066	[0.8575, 0.8691]	
	AUC	0.8992 ± 0.0088	[0.8915, 0.9068]	
Improvement	Accuracy	+12.22%	–	8.33
	F1-Score	+9.04%	–	9.16
	AUC	+8.66%	–	10.01

Paired t-tests yielded p-values under 0.001, far below the corrected threshold of 0.0167. The large Cohen's d values highlight not just statistical but practical differences in performance.

B. Seed-by-Seed Performance Analysis

Looking at each seed, BERT maintained its edge consistently. The performance spreads were:

- BERT Accuracy: 0.9163 to 0.9425 (spread of 0.0262)
- TF-IDF+MLP Accuracy: 0.8075 to 0.8387 (spread of 0.0312)
- BERT F1-Score: 0.9328 to 0.9534 (spread of 0.0206)
- TF-IDF+MLP F1-Score: 0.8536 to 0.8722 (spread of 0.0186)

The small standard deviations (less than 0.011) point to stable outcomes, reinforcing the experimental reliability.

C. McNemar's Test Results

The McNemar's tests varied by seed, with chi-squared from 0.0000 to 61.7985 and p-values from 0.0000 to 1.0000. An average p-value of 0.5282 indicates that while BERT excels in overall scores, the exact errors differ depending on the data split.

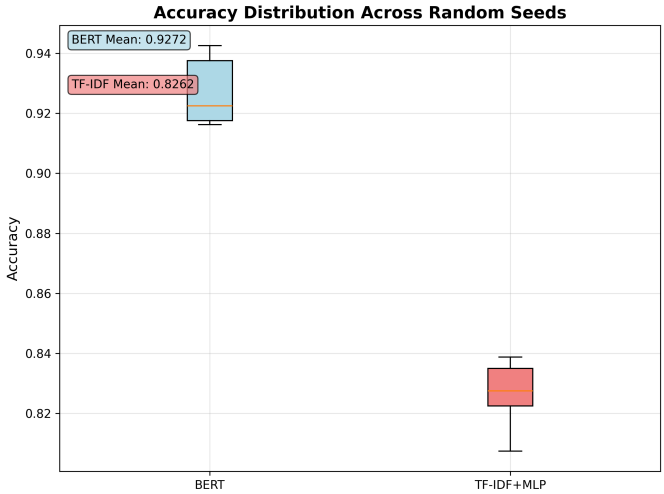

Fig. 1: Distribution of accuracy scores across five random seeds for BERT and TF-IDF+MLP models. The box plots show median, quartiles, and outliers, demonstrating BERT's consistently superior performance with mean accuracy of 0.9272 compared to TF-IDF+MLP's 0.8262.

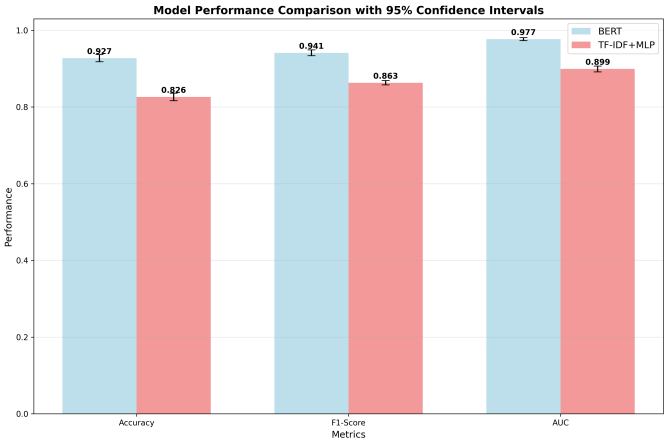

Fig. 2: Performance comparison with 95% confidence intervals across all evaluation metrics. BERT consistently outperforms TF-IDF+MLP with non-overlapping confidence intervals, indicating statistically significant differences across all metrics.

V. Error Analysis and Discussion

A. Comprehensive Error Analysis Framework

To dig into the performance gaps, I conducted a thorough error review, covering quantitative trends and specific cases. This involved confusion matrices for patterns, breakdowns by class, correlations with language features, and close looks at wrong predictions.

B. Classification Pattern Analysis

From the matrices, BERT achieved a good balance in precision (95.77%) and recall (91.89%), whereas TF-IDF+MLP favored precision but fell short in recall, especially for humorous texts. This implies BERT is more adept at confidently identifying humor, while TF-IDF+MLP tends to be cautious.

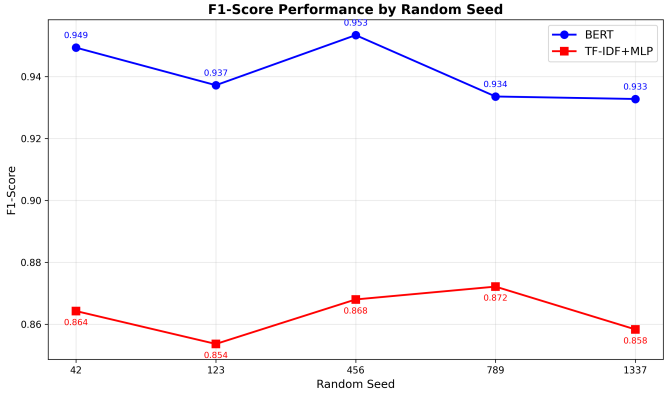

Fig. 3: F1-score performance across different random seeds. BERT consistently achieves higher F1-scores than TF-IDF+MLP across all five experimental runs, with values ranging from 0.9328 to 0.9534 for BERT and 0.8536 to 0.8722 for TF-IDF+MLP.

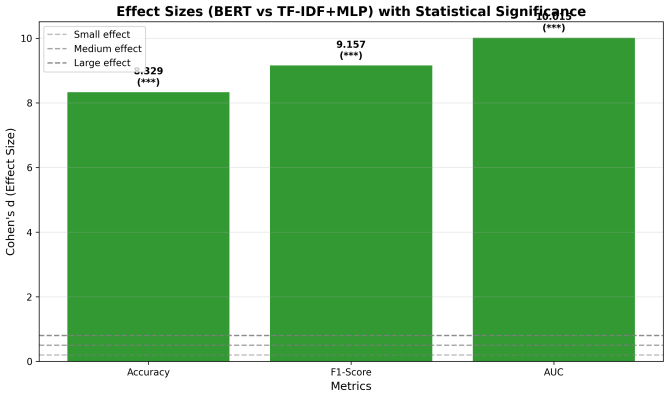

Fig. 4: Effect sizes (Cohen's d) with statistical significance indicators. All metrics show large effect sizes ($d > 8.0$) with high statistical significance ($p < 0.001$), indicated by three asterisks. The horizontal lines represent small (0.2), medium (0.5), and large (0.8) effect size thresholds.

1) False Positive Analysis: BERT had fewer false positives (around 20 per seed) compared to TF-IDF+MLP. Common BERT false positives included:

- Subtle irony that relies on cultural knowledge
- Texts open to multiple interpretations
- Jokes that try but don't fully land

2) False Negative Analysis: For BERT's false negatives (about 40), typical issues were:

- Very understated humor with few clues
- Jokes tied to specific cultures not in training
- Humor needing external visuals or context

TF-IDF+MLP had higher false negative rates, particularly with:

- Humor needing broader context beyond simple word groups
- Clashes in meaning without obvious word signals
- Fresh or inventive language not seen before

One example: "A fat woman just served me at McDonalds and said 'Sorry about the wait'. I replied and said, 'Don't worry, you'll lose it eventually'." (labeled humorous, possibly offensive) – TF-IDF mislabeled it non-humorous, missing the wait/weight pun, but BERT got it right.

Another: "Don't worry if a fat guy comes to kidnap you... I told Santa all I want for Christmas is you." (humorous, non-offensive) – TF-IDF failed, not connecting fat guy to Santa, while BERT succeeded.

C. Linguistic Feature Impact Analysis

Certain language aspects affected the models differently:

1) Syntactic Complexity: BERT managed complex sentence structures steadily, but TF-IDF+MLP's accuracy dropped as syntax got harder. This comes from BERT's ability to link distant parts via attention, unlike TF-IDF's focus on nearby words.

2) Semantic Ambiguity: Both faced issues with texts that could mean multiple things, but BERT resolved them better. Its bidirectional processing helps clarify hidden meanings essential for humor.

3) Lexical Diversity: Texts with varied vocabulary boosted BERT but hindered TF-IDF+MLP. BERT's pre-training equips it for diverse words, while TF-IDF suffers from sparse features in new combinations.

D. Computational Efficiency Analysis

Despite BERT's advantages, it demands more resources:

- Training time: 135–142 seconds per seed vs. 15–20 for TF-IDF+MLP
- Memory: About 2.3 GB GPU vs. 150 MB RAM
- Inference: 60.55 samples/second vs. over 1,200

These factors matter for practical applications with limits on power or speed.

E. Failure Mode Characterization

The errors revealed consistent weak points for each model:

1) BERT Failure Modes:

1) Over-emphasis on context: Sometimes confuses serious text with funny elements
2) Cultural limitations: Fails on humor from less common backgrounds
3) Time-sensitive issues: Struggles with jokes needing up-to-date knowledge

2) TF-IDF+MLP Failure Modes:

1) Lack of context awareness: Misses jokes beyond immediate word patterns, like the Santa example
2) Restricted to surface meaning: Doesn't catch deeper relations, as in puns
3) Feature gaps: Handles new or creative wording poorly

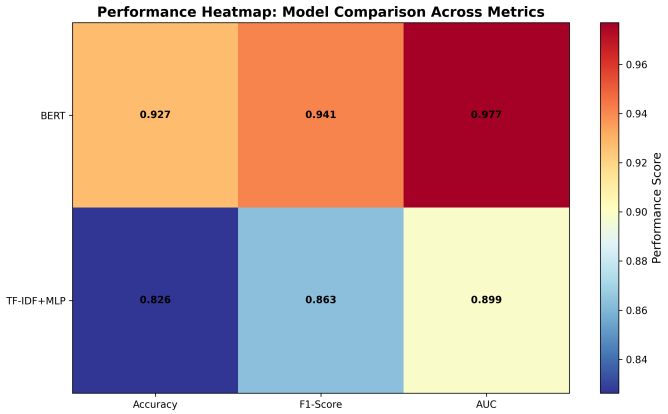

Fig. 5: Performance heatmap comparing BERT and TF-IDF+MLP across all evaluation metrics. The color intensity represents performance scores, with BERT consistently achieving higher scores (darker colors) across accuracy, F1-score, and AUC metrics.

F. Implications for Humor Detection Research

These findings offer guidance for future work:

1) Model Architecture Insights: BERT's success underscores the value of context in humor. Its attention mechanism effectively links elements that create funniness.

2) Dataset Considerations: The dataset's variety by age shows humor's subjectivity. Expanding to more cultures and languages could broaden understanding.

3) Evaluation Methodology: Using multiple seeds and thorough stats is crucial. Relying on single runs might give skewed views of model strength.

VI. Conclusion and Future Work

In this study, I compared BERT and TF-IDF+MLP for humor detection on the HaHackathon dataset. BERT showed clear superiority, with boosts of 12.22% in accuracy, 9.04% in F1-score, and 8.66% in AUC, all backed by strong stats (large Cohen's d and low p-values).

This edge stems from BERT's skill in context and semantics, vital for humor types in the dataset. Transformers like BERT manage complex word ties that traditional methods overlook [5].

Error reviews highlighted BERT's strength in irony and context, while TF-IDF+MLP faltered on ambiguity. These insights point to areas for enhancement.

The HaHackathon dataset advances the field by linking humor to offense and diversity, revealing detection challenges.

Looking ahead, possible paths include:

- Multi-modal systems: Combining text with images or audio for richer humor
- Broader diversity: Datasets with more cultural and linguistic variety
- Adapting over time: Models that update with changing humor styles

- Better explanations: Tools to show how models decide on humor
- Resource efficiency: Ways to slim down BERT while keeping performance

Additionally, the statistical approach here could apply to other language tasks, promoting careful testing in the field.

ACKNOWLEDGMENTS

I used language models to help with drafting and polishing. I am fully responsible for the ideas, analysis, and final content.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, Minnesota, 2019, pp. 4171–4186. [Online]. Available: https://arxiv.org/pdf/1810.04805.pdf

[2] O. Weller and K. Seppi, "Humor detection: A transformer gets the last laugh," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, 2019, pp. 3621–3625. [Online]. Available: https://aclanthology.org/D19-1372.pdf

[3] J. A. Meaney, S. R. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy, "SemEval-2021 task 7: HaHackathon, detecting and rating humor and offense," in *Proceedings of the 15th International Workshop on Semantic Evaluation*, Online, 2021, pp. 105–119. [Online]. Available: https://aclanthology.org/2021.semeval-1.9.pdf

[4] I. Annamoradnejad and G. Zoghi, "ColBERT: Using BERT sentence embedding in parallel neural networks for computational humor," *arXiv preprint arXiv:2004.12765*, 2022. [Online]. Available: https://arxiv.org/pdf/2004.12765.pdf

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. [Online]. Available: https://papers.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[6] D. Yang, A. Lavie, C. Dyer, and E. Hovy, "Humor recognition and humor anchor extraction," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015, pp. 2367–2376. [Online]. Available: https://www.cs.cmu.edu/~hovy/papers/15EMNLP-humor.pdf

[7] P.-Y. Chen and V.-W. Soo, "Humor recognition using deep learning," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, 2018, pp. 113–117. [Online]. Available: https://aclanthology.org/N18-2018.pdf

[8] R. Mihalcea and C. Strapparava, "Making computers laugh: Investigations in automatic humor recognition," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, 2005, pp. 531–538. [Online]. Available: https://aclanthology.org/H05-1067.pdf

[9] J. Mao and W. Liu, "A BERT-based approach for automatic humor detection and scoring," in *Proceedings of the Iberian Languages Evaluation Forum*, vol. 2421, Bilbao, Spain, 2019, pp. 197–202. [Online]. Available: https://ceur-ws.org/Vol-2421/HAHA_paper_8.pdf