Amsterdam University College

Text Mining

Jelke Bloem

Sarcasm Detection in YouTube Comments Using Transformer-Based Models

Project Report

Tom Unger
tom.unger@student.auc.nl
Julia Olifiers
julia.olifiers@student.auc.nl
Sterre Rosdorff
sterre.rosdorff@student.auc.nl
Raluca Dinu
raluca.dinu@student.auc.nl

23.05.2025

1 Abstract

This paper explores sarcasm detection in YouTube comments using transformer language models. We fine-tune BERTweet and RoBERTa and evaluate their performance with and without additional context features. The results show that incorporating context improves model accuracy and F1-score, particularly for the sarcastic class. BERTweet with context overall performed the best, showing the benefits of domain-specific pretraining and contextual cues. Our findings underline the challenges posed by sarcastic language in sentiment analysis and indicate the need for tailored, platform-specific approaches to improving the robustness of natural language processing systems.

2 Introduction

As our world becomes increasingly digital, it is of growing concern to ensure safety on online platforms. For instance, the YouTube platform receives millions of comments from users on a daily basis, a good number of which are checked for hate speech and toxic phrases (YouTube, n.d.). These moderation systems rely heavily on accurate sentiment classification. However, detecting sentiment is not always straightforward, especially when sarcasm is involved. Sarcasm often disguises an underlying negative sentiment as a positive one. This sentiment poses challenges for machines to interpret thoughts accurately and automatically (Maynard, 2014). Improved detection of sarcasm is crucial to creating safer, more inclusive online environments, particularly as NLP technologies assume an increasing role in sustaining civil dialogue within large online communities (Anup et al. 2020).

Even the most advanced models can be misled by sarcasm (Rathod et al. 2023). Successful detection depends not only of linguistic features, but it also requires understanding of the context, world knowledge, and even the broader discourse. The fact that comments are inherently perplexing renders sarcastic structures challenging to annotate correctly even by human annotators (González et al. 2011).

A majority of the available literature utilizes social media data, resting primarily on Twitter where sarcasm is typically self-annotated with hashtags like #sarcasm, or selected by scholars with inter-annotator agreement on a sarcastic sentence. Although abundant, the datasets are usually short, noisy, and informal, filled with emojis, hashtags, and slang (Rathod et al. 2023). This then poses a unique challenge: distinguishing signal from noise. Moreover, there is a lack of domain-specific datasets for websites such as YouTube, where sarcasm is expressed differently and is situated within vast multi-faceted comment sections.

To address these gaps, our research focuses on sarcasm detection in YouTube comments by fine-tuning pretrained transformer models on a hand-labelled dataset. We evaluate their ability to detect sarcasm and how they add to the overall sentiment classification accuracy. The goal of this study is to aid in the development of NLP technologies that are not only advanced in their engineering, but are also attuned to the complexities of human expression on digital forums.

Related Work

Purigilla et al. (2030) compare traditional ML models and advanced transformer-based models (including RoBERTa and BERTweet) when exploring the automatic classification of conversational humor. They found that fine-tuned BERT models significantly outperform traditional approaches in both binary and multi-label tasks. Satire, sarcasm, and schadenfreude emerged as the most common humor types. As an extension of their work, we make use of RoBERTa and BERTweet to test the performance of sarcasm detection on YouTube comments.

3 Methodology

Overview

To investigate the effectiveness of transformer-based language models in detecting sarcasm in YouTube comments, we focus on two pre-trained models: RoBERTa and BERTweet. Both models are evaluated with and without added contextual information. The goal is to assess how these models perform in sarcasm detection and whether incorporating contextual information improves classification accuracy.

Data Collection

Our dataset was built by selecting four YouTube videos from the TLC show *My Strange Addiction.* The series features people with unusual habits or lifestyles that often cause strong reactions from viewers. We purposefully selected those videos because their content is controversial, which tends to provoke a wide range of comments, including sarcasm, mockery, disbelief, and sometimes empathy.

To collect the data, we developed a Python scraper that gathers user comments from the selected videos using the YouTube Data API.

Dataset Description with Statistics

We created a CSV file with all the raw YouTube comments, organized in four columns: video_id, comment_number, sarcastic, and comment. The sarcastic field is a Boolean label that we manually assigned and each row represents a single comment. We were required to manually label the data, due to the lack of publicly available datasets on sarcastic YouTube comments. Figure 1 illustrates a summary of the data, including the distribution of sarcastic remarks for the chosen videos.

Video ID	Total Comments	Sarcastic Comments	Sarcasm Percentage (%)
4VGd-pvSc0w	300	79	26.33
Ryq4ILnTmog	285	94	32.98
a1rpr0Afhfg	300	61	20.33
eBfw5NMgizU	300	83	27.67

Figure 1. *Summary of labeled sarcasm in comments across four YouTube videos from TLC series My Strange Addiction.*

Preprocessing

As part of our preprocessing pipeline, we added a filtering condition to our scraper to keep only comments that were longer than two words. This step was designed to reduce the noise in the data, because very brief comments such as "lol" or "nice" are often unclear and offer little to no linguistic context, which can affect the model's performance.

One of the models we decided to use could handle informal language, while the other required more normalized data. We fixed common misspellings, cleaned up the text overall, and changed abbreviations (e.g. "omg" to "oh my god", "ik" to "I know") to accommodate this. Preprocessing was necessary in order to reduce variability in data and make sure that the differences in performance were due to the models' architecture and not due to inconsistencies in the input.

Model Selection

Among the evaluated models, BERTweet and RoBERTa-base emerged as the most promising candidates. As transformer-based architectures, both leverage self-attention mechanisms to discern complex word relationships, an essential capability for detecting sarcasm. Although comparable in size, the two models diverge in their training domains. BERTweet is pretrained on a dataset of 850 million English-language tweets.

This makes it particularly adept at processing the informal, slang-heavy, and often sarcastic language common in YouTube comments, which closely resemble tweets in tone and structure. Additionally, its tokenizer is tailored for social media content, effectively handling elements like emojis and hashtags. On the other hand, RoBERTa-base is trained on 160GB of diverse web text. While it possesses a broad linguistic knowledge, it is less effective at capturing the informal and nuanced language typical of social media.

Context Integration

To test model performance with added context, we concatenate a short description of the video content with the corresponding user comment. The following format was used as model input: [Video Context] [SEP] [User Comment]. The [SEP] token acts as a segment delimiter, allowing the models to distinguish between the contextual information and the target comment. The video context is a simple, descriptive sentence summarizing the topic and main content of the video. For example, for a video titled “Woman Eats More Than 100 Bars Of Soap Every Year”, the video context was “A Woman is addicted to eating soap”.

Model	**Accuracy**	**Sarcastic Precision/ Recall / F1**	**Not Sarcastic Precision / Recall / F1**
RoBERTa	0.85	0.77 / 0.53 / 0.63	0.86 / 0.95 / 0.90
RoBERTa + Context	0.87	0.79 / 0.63 / 0.70	0.89 / 0.95 / 0.92
BERTweet	0.81	0.60 / 0.63 / 0.61	0.88 / 0.87 / 0.87
BERTweet + Context	0.85	0.67 / 0.79 / 0.72	0.93 / 0.87 / 0.90

Figure 2: *Overview of Model performance without context integration*

4 Results

Figure 2 presents the classification metrics for each model, including precision, recall, and F1-score, reported separately for sarcastic and non-sarcastic instances, as well as overall accuracy. Precision measures how many of the comments predicted as sarcastic were actually sarcastic, while recall indicates how many of the actual sarcastic comments the model successfully identified. The F1-score balances these two by combining precision and recall into a single value, giving a more general overview of the model's performance.

RoBERTa achieved an overall accuracy of 0.85. Adding contextual information improved its performance, reaching an accuracy of 0.87. For the *sarcastic* class, recall increased from 0.53 to 0.63 and the F1-score improved from 0.63 to 0.70. RoBERTa’s performance on the *not sarcastic* class remained high and largely stable after context integration, with slight gains in both precision and F1-score. These results suggest that the additional context helped the model better classify sarcastic content.

BERTweet initially performed less accurately overall (0.81), but benefited significantly from the inclusion of context, reaching an accuracy score of 0.85. For the *sarcastic* class, recall significantly improved from 0.63 to 0.79 and the F1-score increased from 0.61 to 0.72-the highest recall and F1-score on the *sarcastic* class across all model configurations. The added contextual information improved the model’s performance on the *not sarcastic* class as well, with precision rising to 0.93 from 0.88.

Overall, RoBERTa remained the most accurate model in general, while BERTweet with context achieved the best balance for sarcasm detection, demonstrating the highest F1-score and the most improvement from integrating context.

5 Discussion

A key challenge we encountered during this project was the inherent subjectivity of sarcasm. Unlike more straightforward forms of sentiment, sarcasm often depends on cultural cues, context, or even the tone of voice, none of which are fully captured in plain text. As a result, even human annotators may disagree on whether a comment is sarcastic or not. In González et al. (2011), a cross-evaluation was conducted comparing the performance of machine learning techniques and human judges. Neither group performed well, highlighting the inherent difficulty of the task. This subjectivity not only affects dataset reliability but also underscores the broader limitations of sarcasm detection in text-based settings. In manual annotation, beyond the issue of subjectivity, there is also the challenge of labeling ambiguous cases consistently, which contributes to noisy training data. Additionally, error analysis revealed that most misclassifications involved ambiguous or context-dependent sarcasm, often lacking clear linguistic cues. This suggests the need for models and datasets that better account for uncertainty in sarcastic expression. One potential improvement would be to include a measure of annotator certainty. By indicating how confident annotators are in their labels, models could learn to differentiate between clear and ambiguous training examples.

To further improve sarcasm detection in our corpus, one potential approach would be to pretrain our model on existing sarcasm datasets, such as SARC, a Reddit-based dataset, or the Twitter Sarcasm Dataset. However, applying these resources to YouTube comments poses several challenges. Unlike Twitter, YouTube comments are typically longer, lack the hashtag-based cues common in tweets, and often reference video content rather than text. Additionally, they do not offer the hierarchical thread structure available in Reddit data, which helps provide contextual cues for sarcasm. These differences in linguistic style and context make it difficult to directly apply existing datasets to our task, underscoring the need for platform-specific approaches.

6 Conclusion

Our results demonstrate that transformer-based models, particularly BERTweet with added context, are effective for detecting sarcasm in YouTube comments. Contextual information improved classification performance, especially for informal and ambiguous expressions. However, our error analysis highlighted ongoing challenges, with many misclassifications involving subtle, context-dependent sarcasm. These findings mirror difficulties in manual annotation, where subjectivity and inconsistency introduce noise into the data. To address this, future work could incorporate annotator certainty scores and explore domain-specific pretraining. While existing resources like SARC or the Twitter Sarcasm Dataset offer valuable training data, their differences in structure and style limit direct applicability to YouTube. Our study underscores the importance of platform-specific sarcasm detection approaches for improving sentiment analysis in real-world applications.

Author Contributions

Tom: Youtube Scraper, Comment Labeling, Model setup and tuning, Presentation, Writing Report

Sterre: Model setup and tuning, Comment Labeling, Gathered results, Presentation, Writing Report,

Raluca: Comment Labeling, Dataset Creation, Dataset Preprocessing, Presentation, Writing Report

Julia: Comment Labeling, Dataset, Preprocessing, Literature Research, Presentation, Writing Report

References

Anup, A., G., S., H. R., S., Upadhyaya, M., Ray, A. P., & Manjunath, T. C. (2020). Sarcasm detection in natural language processing. *Materials Today: Proceedings, 37*, 3324–3331. https://doi.org/10.1016/j.matpr.2020.09.124

Datasaur. (2023). Using NLP to keep YouTube comments safe. https://datasaur.ai/blog-posts/nlp-keep-youtube-comments-safe

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers* (pp. 581–586). Association for Computational Linguistics. https://www.researchgate.net/publication/220874376

Jain, D., Kumar, A., & Garg, G. (2020). Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN. *Applied Soft Computing, 91*, 106198. https://doi.org/10.1016/j.asoc.2020.106198

Kataria, A., & Rathod, S. (2023). Sarcasm detection using natural language processing. *SSRN Electronic Journal.* https://ssrn.com/abstract=4451909

Maynard, D. G., & Greenwood, M. A. (2014). Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. European Language Resources Association (ELRA). [Insert page numbers if available]

Purigilla, G. (2023). *Automatic classification of conversational humor with a focus on COVID-19 tweets* (Doctoral dissertation, International Institute of Information Technology Hyderabad).

TLC. (2013, March 25). *This woman can't stop drinking paint* [Video]. YouTube. https://www.youtube.com/watch?v=K3TqdG4_7oc

TLC. (2013, April 3). *Woman eats more than 100 bars of soap every year* [Video]. YouTube. https://www.youtube.com/watch?v=TWkH1xZp3Gs

TLC. (2014, August 20). *Addicted to bathing in bleach* [Video]. YouTube. https://www.youtube.com/watch?v=E_5kX-5I3M8

TLC. (2014, August 27). *I'm addicted to eating beds* [Video]. YouTube. https://www.youtube.com/watch?v=OqsBuxCk4L8

YouTube. (n.d.). *Hate speech policy*. YouTube Help. https://support.google.com/youtube/answer/2801939?hl=en