# Text Classification of Milwaukee Bucks Fan Reactions on Reddit

**Anonymous ACL submission**

## 1 Problem

The goal of this project is to evaluate the performances of different machine learning and deep learning approaches to text classification on domain-specific social media data. The domain in this case will be the reaction from fans of the Milwaukee Bucks following games from the 2023-2024 NBA season. The text classification problem is as follows: given a fan comment, how well can one predict whether that comment followed a Milwaukee Bucks win or a loss? The difficulty in this problem lies in the nature of the data. Social media comments oftentimes contain informal language, spelling errors, slang, and oftentimes do not offer context in the way that other texts might offer. The problems of informal language and slang are amplified in domain specific data. Thus, by evaluating the performaces of different models, insights might be gained regarding model selection on this type of task.

## 2 Data

The data for this project has been collected from the social media site Reddit using Reddit's PRAW API. The Bucks subreddit, r/MkeBucks, labels all postgame threads (the space where fans comment after every game), and so comments in these postgame threads were scraped. Comments from games outside of the 2023-2024 NBA seasons were removed, and the desired data has been collected into a data frame that contains 12,266 comments from 62 games. I am choosing to use text data from a single NBA season so as to best capture fan reactions to a unique team, although the Milwaukee Bucks have undergone significant changes this season, most notably a coaching change. Such changes might effect the way in which fans react following games, however, the data chosen for modeling best reflects how fans view this specific Milwaukee Bucks team.

In order to prepare the data for modeling, I used Python's NLTK library to tokenize and perform stemming on the comments.

## 3 Method

The baseline methods used in this project were established using traditional machine learning algorithms Multinomial Naïve Bayes and Logistic Regression. These algorithms were chosen due to their relative rudimentary approach to text classification problems, and thus comparing their respective performances to more complicated approaches offer insight to the efficacy of different types of approaches to this type of text classification problems. Python's Scikit-learn library has been used to run the algorithms on the data. In order to establish the baseline scores, which for this project will be F1 scores, I did not introduce any variations to these methods.

Going forward, I will design a Convolutional Neural Network using the Python Library TensorFlow that classifies the text data. Upon completing this task, I will fine-tune a BERT model from Python's Hugging Face library

## 4 Preliminary Results

The baseline F1 scores established using Naïve Bayes and Logistic Regression are 0.62 and 0.64, respectively. These scores are lower than expected. I expected F1-scores to be around 0.7, with the hopes of better performances from the CNN and BERT models. That being said, these scores likely reflect the difficulty of classifying social media data, particularly data that comes from a very specific domain. It is possible that social media users in this domain use somewhat similar language following wins and losses, and perhaps these baseline models do not capture context as well as more

intricate models. Thus, I am optimistic that the following models will perform better on this data, but it could be the case that more rudimentary approaches perform better on this type of data.

## 5 Related Work

### 5.1 Paper 1

Kanish Shah and other contributors in their paper, "A comparative analysis of logistic regression, random forest and KNN models for the text classification." analyze the performances of different machine learning approaches to text classification problems. The data used in their project is BBC news articles. This project also compares different approaches to text classification, but specifically in regards to social media data.

### 5.2 Paper 2

Santiago González-Carvajal and Eduardo C. Garrido-Merchán in their paper, "Comparing BERT against traditional machine learning text classification." explore the efficacy of BERT models in text classification. They argue that BERT offers a flexibility that other models cannot offer in this sort of problem. The authors used an IMBD movie review dataset for their study. While this project will explore the performances of BERT on text classification, the project will evaluate the performance of BERT on a very different type of text data, and thus could yield drastically different results.

### 5.3 Paper 3

Shaomin Zheng and Meng Yang in their paper "A new method of improving bert for text classification." argue for the superiority for BERT models in text classification. That being said, they claim that shortcomings exists in BERT models, particularly in their inability to recognize context in longer text data. It is to be seen if this shortcoming exists when classifying social media data.

### 5.4 Paper 4

David Rogers and contributors in their paper, "Real-time text classification of user-generated content on social media: Systematic review." speak to the superior performance of neural networks over traditional machine learning techniques in text classification. In their reasearch, they studied social media data from a multitude of sites. It is in this respect that my project differs from theirs, as I analyze social media data from a specific site. It is to be seen whether neural networks outperform traditional machine learning techniques in my project.

### 5.5 Paper 5

Hemant Purohit and contributors in their paper, "Intent classification of short-text on social media." discuss the difficulties discerning intent in social media text data. This task will be a challenge in my analysis, as understanding intent will allow for more accurate classifications.

## 6 Division of Labor

I am doing the project alone and am hence responsible for all work.

## 7 Timeline

Over spring break, I am planning on finishing building the CNN, thereby establishing the baseline metrics to which the BERT's performance will be compared. I will also begin fine tuning the BERT model during the week off with the goal of finishing this task about two weeks from the submission of this report. Then, I will begin typing my final report and preparing my final presentation.

## 8 References

1. Shah, Kanish, et al. "A comparative analysis of logistic regression, random forest and KNN models for the text classification." Augmented Human Research 5 (2020): 1-16.

2. González-Carvajal, Santiago, and Eduardo C. Garrido-Merchán. "Comparing BERT against traditional machine learning text classification." arXiv preprint arXiv:2005.13012(2020).

3. Zheng, Shaomin, and Meng Yang. "A new method of improving bert for text classification." Intelligence Science and Big Data Engineering. Big Data and Machine Learning: 9th International Conference, IScIDE 2019, Nanjing, China, October 17–20, 2019, Proceedings, Part II 9. Springer International Publishing, 2019.

4. Rogers, David, et al. "Real-time text classification of user-generated content on social media: Systematic review." IEEE Transactions on Computational Social Systems 9.4 (2021): 1154-1166.

5. Purohit, Hemant, et al. "Intent classification of short-text on social media." 2015 ieee international conference on smart

city/socialcom/sustaincom (smartcity). IEEE, 2015.