# Detecting Language Differences between Online Communities: A Binary Classification Problem

George Tao | Jeremiah Sahabu

## I. Abstract

Social media has become a global phenomenon that generates large amounts of data every moment on various different platforms. Each platform has its own purpose, community, and culture. On LinkedIn, users try to portray themselves as professionals, while on Tinder users try to portray themselves as attractive. Our goal was to use NLP to classify the language that each social media community uses in daily interactions. We use LSTMs and RoBERTa to perform binary text classification on posts from Facebook and Twitter to see how well a model can differentiate the two platforms. We show how these different models perform against a naive baseline model and analyze ways in which we could have improved our research.

## II. Introduction

Social media has grown to cover almost every aspect of life, with its influence covering many fields like sports, politics, and product advertising. More importantly, social media allows for us to connect with our friends through posted public messages. Social media has been introduced in many different contexts, with each platform differing in how formal or informal its users express themselves.

Our goal is to fine-tune models to effectively recognize the difference between posts from Facebook and posts from Twitter. By fine-tuning models in a simple task like social media classification, we can eventually tackle more difficult tasks like detecting “fake news” or sarcasm. As social media platforms become more and more diverse, the need for differentiating between these different platforms increases.

We present two different natural language processing models to classify text and we compare this against a naive baseline model. The NLP models used in this research paper are both motivated by existing research in the field.

## III. Background

**Improving Classification of Twitter Behavior During Hurricane Events - Stowe et al. (2018)**
In this paper, the topic of text classification is explored in depth. The researchers try to create an NLP model using convolutional neural networks that could classify the genre of a Twitter post, focusing primarily on posts during a hurricane event. They chose to test various different models, but ultimately chose to go with a CNN due to its superior results. We noticed that they chose to forego the use of Long Short Term Memory (LSTM) networks and RoBERTa, two models we wanted to focus on. However, because RoBERTa was not released at the time, CNNs were the best language model at the time.

**RoBERTa: A Robustly Optimized BERT Pretraining Approach - Liu et al. (2019)**
Enter RoBERTa, a new approach to language modeling. RoBERTa is a language model built upon the previously state-of-the-art BERT model that optimizes existing hyperparameters. There are 4 main optimizations: larger batches trained on

more data, removing the next sentence prediction objective, training on longer sequences, and an adjustment to the masking pattern. This paper highlights the importance of changing parameters within existing models to explore potentially overlooked design choices.

The best RoBERTa models achieved state-of-the-art results on multiple evaluation metrics: GLUE, RACE, and SQuAD.

### Using Functional Schemas to Understand Social Media Narratives - Yan et al. (2018)

In the following paper, the authors explored differences in schematic structures in different subreddit threads. The article comments that different subreddits are made of different communities which utilize different language structures. An analogy would be varying dialects of the same language depending on region: they are similar as they speak the same language but the language style varies. The authors utilize various deep learning models on different environmental subreddits to see whether these models can pick up the subtle differences in word embeddings. The models trained in the article had good outcomes which is a good indication of performance for our task.

## IV. Methods

Originally, we wanted to create three NLP models in addition to our baseline model - LSTM, CNN, and RoBERTa. However, we decided to focus on only two models and chose LSTM and RoBERTa. Using these two NLP models, we can better capture the difference between the state-of-the-art language models now and over 20 years ago, when LSTMs was first introduced.

### Data

We collected raw unprocessed data from both Facebook and Twitter. For Facebook, we utilized a scraper to pull posts from public verified pages. We took data from five categories: celebrities, politicians, sports organizations, businesses, and news stations. For Twitter, we used an API to collect the most recent 3000 tweets from all accounts followed by George Tao, which included both famous celebrities as well as friends. In total, we collected 160K text examples (20K from Facebook, 140K from Twitter). When making our train and test data, we trimmed the Twitter Data to match the number of text posts we received on Facebook. We believe that our data is representative of the vernacular of each platform, though we also recognize there may be some bias as the methods of collection for each platform was not uniform.

### Baseline Naive Model

The baseline model was used as a benchmark for the deep learning models later on. Knowing that Twitter as a platform has a 280 character limit, we leveraged this knowledge to predict texts naively. If texts had over 280 characters, it was predicted to be a Facebook Post. Otherwise, it would be a Tweet. Without looking at the context or word embeddings, we are confident that this is a strong way to differentiate between posts. The runtime of our Baseline Model was almost instantaneous.

### LSTM (Long Short Term Memory)

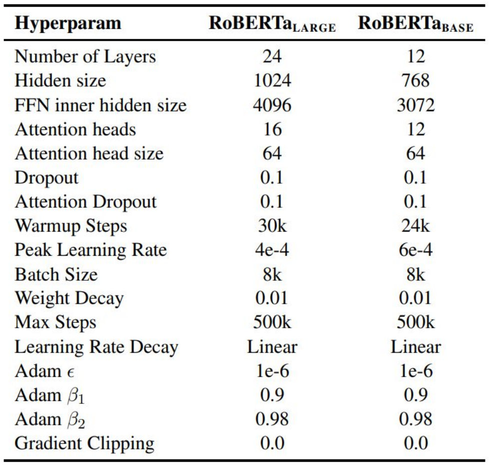
To increase accuracy from the baseline model, we implemented an LSTM model. The LSTM model takes in the text, converts all posts into a sequence of integers before predicting a label for each post. The neural net runs over 3 epochs, with a large batch size of 64 as to avoid overfitting and allowing there to be space before the weights are updated. By using an LSTM, we found that our model was successful in identifying subtle differences in schematic structures between Facebook and Twitter posts. The runtime for the LSTM model was around 15 minutes, which is substantially longer than our Naive model.

### RoBERTa

Our final model, RoBERTa, is the current state-of-the-art model, and because of this, we had the highest expectation for this model. We chose to use RoBERTa$_{\text{BASE}}$, the smaller variant of RoBERTa with less heads, layers, and hidden layers. We primarily chose to use this smaller model due to computing limitations. Additionally, because of the triviality of our task and the large size of our dataset, we believed that the difference in accuracy would be negligible compared to the difference in performance.

The RoBERTA$_{\text{BASE}}$ model has 12 heads and 12 layers, allowing for 144 distinct attention mechanisms, similar to the design of BERT$_{\text{BASE}}$. One of the main differences as noted before is the drastic increase in batch size, with RoBERTA$_{\text{BASE}}$ having a batch size of 8000. Further specifications can be seen in Figure 1 shown on the right.

We preprocessed our dataset into tokenized features that RoBERTa could read and added a classification layer on top of the model for the purpose of fine-tuning with our data. Overall, the runtime for training this model was around 90 minutes, which was the longest out of all the models.



| Hyperparam | RoBERTa$_{\text{LARGE}}$ | RoBERTa$_{\text{BASE}}$ |
|---|---|---|
| Number of Layers | 24 | 12 |
| Hidden size | 1024 | 768 |
| FFN inner hidden size | 4096 | 3072 |
| Attention heads | 16 | 12 |
| Attention head size | 64 | 64 |
| Dropout | 0.1 | 0.1 |
| Attention Dropout | 0.1 | 0.1 |
| Warmup Steps | 30k | 24k |
| Peak Learning Rate | 4e-4 | 6e-4 |
| Batch Size | 8k | 8k |
| Weight Decay | 0.01 | 0.01 |
| Max Steps | 500k | 500k |
| Learning Rate Decay | Linear | Linear |
| Adam $\epsilon$ | 1e-6 | 1e-6 |
| Adam $\beta_1$ | 0.9 | 0.9 |
| Adam $\beta_2$ | 0.98 | 0.98 |
| Gradient Clipping | 0.0 | 0.0 |

**Figure 1**: Hyperparameters for RoBERTa

## V. Results

### Baseline Naive Model

Our Naive Model reached an accuracy of 52%. This is reasonable as the model only predicts based on character length, and our train data had an equal number of Twitter and Facebook posts.

### LSTM

The LSTM model performed better than the Naive model with an accuracy of 87%. This 35% increase in accuracy can be attributed to the model looking at the context and spatial structure of texts rather than only character lengths.

### RoBERTa

As expected RoBERTa performed the best of all the models. However, with an accuracy of 97.1%, the model outperformed our expectations which may be partially due to the flaws in the data collection.

One thing we wanted to see was how well the fine-tuned RoBERTa model performed on George’s Twitter account since all of the Twitter training data was gathered from the account that he follows. When running our fine-tuned model on a random subset of his tweets, our model evaluated the data to be Twitter data with an accuracy of 97.4%.

## VI. Conclusion

### Final Remarks

Overall, binary text classification is an important tool to be able to separate a group of texts by two distinct features using NLP. We found that RoBERTa outperformed the other language models and even our expectations, with the highest accuracy at 97%. This high accuracy proves its effectiveness in being able to detect differences in language vernacular between two online communities which can speak to better understanding how we portray ourselves online.

### Further exploration

To further explore the social media space, we can train a multilabel text classification model to compare multiple social media platforms at once. By doing so, we can detect which social media platforms have similar vernacular patterns. In addition, we can apply the binary text classification model to harder NLP problems, such as detecting sarcasm, or differences in language between different languages altogether.

## VII. Bibliography

Liu, Yinhan, et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” ArXiv:1907.11692 [Cs], July 2019. arXiv.org, http://arxiv.org/abs/1907.11692.

Stowe, Kevin, et al. “Improving Classification of Twitter Behavior During Hurricane Events.” Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, 2018, pp. 67–75. ACLWeb, doi:10.18653/v1/W18-3512.

Yan, Xinru, et al. “Using Functional Schemas to Understand Social Media Narratives.” *Proceedings of the Second Workshop on Storytelling*, Association for Computational Linguistics, 2019, pp. 22–33. *ACLWeb*, doi:10.18653/v1/W19-3403.