Tweet like (a) Trump

Paletto Andrea, Tuninetti André

September 5, 2023

Contents

1 Introduction **3**
1.1 Requirements 3

2 Methodology **3**
2.1 Data Preprocessing 3
2.2 Language Model Training 3
2.3 Synthetic Tweet Generation 3
2.4 Perplexity Evaluation 4

3 Conclusion **4**
3.1 Result and Critique 4
3.2 Critique 4

1 Introduction

In this task,we are required to create a language model based on n-grams from approximately 300 tweets extracted from Donald Trump's Twitter profile. We will then generate synthetic tweets that emulate the style of Trump's tweets using this language model.

1.1 Requirements

- Build an n-gram-based language model using the provided tweet dataset.
- Generate synthetic Trump-like tweets using the trained language model.

2 Methodology

The solution involves the following steps:

2.1 Data Preprocessing

1. Load the dataset containing Trump's tweets from a CSV file.
2. Clean and preprocess each tweet by removing special characters, URLs, and normalizing certain punctuation marks. Tokenize the cleaned tweets using the TweetTokenizer from NLTK.
3. Calculate basic statistics on the dataset, including the average length in characters and words of the tweets.
4. Split in train and test sets

2.2 Language Model Training

1. Choose the value of 'N' for the n-grams (e.g., bigram or trigram).
2. Create n-grams and a vocabulary from the preprocessed tweet corpus using the NLTK library.
3. Train a Laplace-smoothed n-gram language model on the n-grams and vocabulary.
4. Train a MLE n-gram language model on the n-grams and vocabulary.

2.3 Synthetic Tweet Generation

For each trained model:

1. Generate synthetic Trump-like tweets using the trained language model.
2. For each generated tweet, ensure that it meets a specified word count and does not contain duplicate sentences.

2.4 Perplexity Evaluation

1. Calculate the perplexity of the trained language model on the generated synthetic tweets. Perplexity measures how well the model predicts the generated text.

3 Conclusion

3.1 Result and Critique

The language model we developed successfully generates synthetic tweets that emulate the style of Donald Trump's tweets. We trained an n-gram-based language model, with $N = 3$, on a dataset . In the evaluation of our model, we calculated the perplexity, which for our trigram model was found to be 1968.00. Meanwhile for the MLE language model the perplexity was **inf**, this indicates that the model is assigning a probability of zero to at least one word in the test set. This could be due to the fact that the MLE smoothing technique does not handle unseen words (words that have not occurred in the training corpus) well.

3.2 Critique

- **Coherence and Context**: Although our model successfully mimics the style, it often falls short in terms of coherence and contextual relevance. Generated tweets may lack logical flow or relevance to current events, which is a critical aspect of Trump's Twitter presence.
- **Repetition**: The model occasionally generates repetitive content, mirroring some of the recurring phrases and themes in Trump's tweets. While this is characteristic of his style, it can lead to a lack of diversity in generated tweets.
- **Originality**: Our model heavily relies on the patterns present in the training data. As a result, it may not produce entirely original content but rather paraphrases or reproduces segments from Trump's actual tweets due to the low cardinality of the dataset.

In conclusion, our language model is a valuable tool for generating synthetic tweets that resemble Donald Trump's tweets, capturing the essence of his distinctive style. However, critical evaluation highlights areas for improvement in terms of coherence, contextuality, and originality.