NLP Project Proposal

Team Name: Mongoose (Ryan Connolly, Clare Treutel, Duncan Collins)

Data: Bragging Dataset - https://archive.org/details/bragging_data

Task:

We will be attempting to replicate the paper **Automatic Identification and Classification of Bragging in Social Media** by Jin et al. This paper was presented at the 2022 Annual Meeting of the Association for Computational Linguistics (ACL).

In this paper, authors employed NLP to detect bragging within the text of social media posts. Here, bragging is specifically defined as "a speech act employed with the goal of constructing a favorable self-image through positive statements about oneself." The authors created a new publicly available dataset annotated with bragging type (including tagged instances of non-bragging tweets), then performed experiments with linguistically informed transformer-based models for **1**. bragging identification (binary classification) and **2.** bragging type classification (multiway classification).

The dataset produced by this analysis consists of 6,696 English-language tweets sampled from Twitter, along with annotations for each noting the presence and type of bragging (Achievement, Action, Feeling, Trait, Possession, Affiliation, or Not [Bragging]). This tag can be used for both the binary classification task and 7-way classification task.

The authors note that the dataset was produced using a mixture of random collection and keyword-based sampling. The sampling method of each tweet is recorded as an annotation in the dataset. The authors used the linguistic concept of "self-disclosure" and stylistic indicators to identify pronouns and other features likely to indicate bragging. All data was filtered for duplicates and non-useful tweets, including retweets, advertising, and tweets with little substantive text (i.e., picture only tweets).

Finally, tweets in the dataset were pre-processed by lower-casing all text, replacing user mentions with @USER, removing emojis, and removing hashtags which were used as known bragging terms and used to collect these tweets.

The authors of this paper used pre-trained transformer-based large language models, augmented with further external linguistic information.

1. **BERT, RoBERTa and BERTweet**

   The authors experimented with Bidirectional Encoder Representations from Transformers (BERT) and its derivatives, RoBERTa and BERTweet. They found that RoBERTA was more robust than BERT and performed better on a wide range of tasks; and BERTweet, which was specifically pre-trained on English tweets, achieves even better performance than BERT and RoBERTa on Twitter-based tasks. The authors fine-tuned BERT, RoBERTa and BERTweet for the binary and multiclass tasks by adding a classification layer.

2. **BERTweet with Linguistic Features**

   The authors "injected" linguistic knowledge into their BERTweet model. This linguistic knowledge includes emotions, sentiment, topic, psychology and more, which the authors

believed would contribute to the linguistic identification of bragging-type statements. This information was vectorized, concatenated and fed to the pre-trained BERTweet encoder for fine-tuning. Three linguistic approaches were explored:

NRC: a word-emotion lexicon mapping English words to ten categories of emotions and sentiment. Each tweet is represented as a ten-dimensional vector where each element is the proportion of tokens belonging to each category.

LIWC: Linguistic Inquiry and Word Count, a dictionary-based approach to counting words in linguistic, psychological, and topical categories. Each tweet is represented as a 93-dimensional vector according to LIWC 2015.

Clusters: Word2Vec clusters were used to represent each tweet as a 200-dimensional vector over thematic subjects.

The authors of this paper compared the performance of six of the models mentioned above, plus three baselines: Majority Class, LR-BOW (a bag-of-words logistic regression), and BiGRU-Att (a Bidirectional Gated Recurrent Unit with Self-Attention). Performance was compared across all models in both binary and multiclass classification using **Precision, Recall** and **Macro-F1. Confusion Matrices** were also created to understand model performance on the multiclass task, examining which types of bragging were likely to be mislabeled or confused by the models.

The authors conducted linguistic **feature analysis** by analyzing unigrams, LIWC and part-of-speech (POS) tags associated with bragging in this dataset. These features were correlated and ranked using univariate Pearson correlations to create a table of linguistic features found to be significant in the identification of bragging. The authors also analyzed post popularity as a factor, which falls outside the scope of this analysis due to its absence from the public dataset.

Our Evaluation:

Using the dataset created by the authors of this paper, we will perform our analysis with the same **data split**: using the keyword-sampled data for training, and the random-sampled data for testing. (In the dataset, each tweet is tagged with the type of sampling which procured it). We will train each model three times, as the authors did, and use the same parameters and hyperparameters recorded in the paper. We will analyze the **Precision, Recall, and Macro-F1** and compare our models' performance to that of the authors' models. We will also create **Confusion Matrices** and **Feature Correlations** to qualitatively analyze and compare the performance of our models.

We will expand our analysis through the introduction of our own ideas, such as:

- Experimenting with **additional training methods**; for instance, instead of splitting the training/testing data by sampling type, training and testing on tweets of all sampling types together using k-fold cross-validation.
- Exploring the potential for other **parameter/hyperparameter settings**.
- Hypothesizing **new linguistic augmentations** for fine-tuning models; for example, creating a new variation on the BERTweet model with a different linguistic feature as additional input.
- Exploring the possibility of a **different data source**.