Reproducibility Study: Automatic Identification and Classification of Bragging in Social Media

Fazle Mohammed Tawsif and Syeda Tasnim Fabiha
University of Southern California
{tawsif, fabiha}@usc.edu

1 Introduction

Human conduct is largely influenced by the urge to be perceived favorably, and self-presentation is a way of creating a positive social image (Bak et al., 2014). Bragging is one of the most known tactics of self-presentation which comprises promoting a desirable quality about the narrator (Dayter, 2014). Social media is the most common platform where bragging is often acceptable or sometimes it is even desired (Dayter, 2018). However, bragging can have unintended consequences such as dislike or lowered self-efficacy (Matley, 2018). Therefore, automatic bragging detection is necessary which can facilitate many peer groups such as online users to help them to enhance their self-presentation tactics, or the NLP applications that can be enhanced in terms of intent identification (Jin et al., 2022).

Considering the importance of automatic bragging identification, Jin et al. (2022) have conducted a brief study of bragging that aims at connecting the gaps between previous research on linguistics and pragmatics. They have contributed a new public dataset, that consists of 6,696 English tweets labeled with bragging along with their classifications (seven types of bragging). Additionally, they have conducted experiments with transformer-based models for bragging identification and bragging classification. Finally, a linguistic analysis of markers of bragging in twitter data along with the models' behavior in detecting bragging have been made in their research.

In this reproduction study, we aim to evaluate Jin et al. (2022)'s analysis on bragging . By using the given dataset, we will implement three transformer based models, a) **BERT** (Devlin et al., 2018), b) **RoBERTa** (Liu et al., 2019), and c) **BERTweet** (Nguyen et al., 2020) for the binary and multi-class classification of bragging. The six different bragging classifications observed from the dataset with which our model will be trained on include: **Achievement**, **Affiliation**, **Action**, **Feeling**, **Possession**, and **Trait**. We will compare the outcome with RoBERTa model combined with linguistic features. Finally, we will include additional data to the test dataset for cross-validating our results with Jin et al. (2022).

2 Scope of reproducibility

The presence of bragging is prevalent in social media, and it seriously impacts online communication. Although online communication plays a vital role in our day-to-day lives, such linguistic analysis on bragging has hardly been conducted on a larger computational scale. Furthermore, automatic bragging detection can influence various aspects of natural language processing such as intent detection (Wen et al., 2017) or conversation modeling (Hori and Hori, 2017). Therefore, a re-evaluation of the study by Jin et al. (2022) might enable us to work on such a topic more extensively.

The core contribution of the original study is the publicly available twitter dataset annotated with bragging and their types. They experimented with three transformer based models and attempted to improve them by including linguistic features like LIWC, NRC or Word2Vec clusters. They fine-tuned BERT, RoBERTa and BERTweet for bragging detection task by adding a classification layer to the pre-trained models. Comparison with baseline classifiers such as Linear Regression with Bag of Words (**LR-BOW**), Bidirectional Gated Recurrent Unit network with self-attention (**BiGRU-Att**) was also performed. The two downstream tasks performed over the new dataset are:

1. Binary Bragging Classification
2. Multi-class Bragging Classification

The study shows that the transformer based models outperform all baseline classifiers. Among them, BERTweet model performs better than BERT

and RoBERTa. Additionally, for binary classification task, injecting LIWC feature alongside BERTweet improves the performance of the model even better. However, the case is not the same for multi-class bragging classification. Injecting Word2Vec Cluster improves BERTweet model's performance instead of LIWC in multi-class classification task. They have also claimed that BERTweet-LIWC model does not get affected by any subset of test data.

2.1 Addressed claims from the original paper

For this study, we will implement three models (BERT, RoBERTa, and BERTweet) using the tweeter dataset provided by the primary study. To validate their claim, we will additionally collect more recent English tweets which will be annotated following the original approach and be added to the test dataset of this study. This will help us identify whether the change in dataset makes any difference to the models' performance. We will further re-evaluate their model combined with linguistic feature LIWC (BERTweet-LIWC) and finally compare the results based on Macro-F1 value.

We will evaluate following claims made on the original work:

- BERTweet model performs the best over BERT and RoBERTa in bragging classification.
- BERTweet-LIWC improves the result in Binary Bragging Classification task and has no impact on Multi-class Bragging Classification task.
- In binary classification the best performing model (BERTweet-LIWC) is not influenced by any change in the test dataset.

3 Methodology

3.1 Model descriptions

In this study we will experiment with the vanilla transformer based models Vaswani et al. (2017) and attempt to improve them by injecting external linguistic feature.

BERT: Bidirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) is intended to jointly condition on both left and right context in all layers in order to pretrain deep bidirectional representations from unlabeled textual data (Devlin et al., 2018). Therefore, a pre-trained BERT model can be fine-tuned by adding only one output layer to perform several downstream tasks. The architecture of BERT is similar to a multi-layer bidirectional transformer encoder based on the original implementation reported by Vaswani et al. (2017). In this study we will fine-tune BERT by adding a extra layer for classification task which will receive the [CLS] token as input.

RoBERTa: RoBERTa is an improved and robust version of BERT model that can match or even outperform the performance of all post-BERT techniques (Liu et al., 2019). RoBERTa uses a transformer architecture with a certain number of layers where self-attention is considered in each block with a specific hidden dimension.

BERTweet: BERTweet is a widely accessible language model that is pre-trained on English Tweets Nguyen et al. (2020). The model is trained following the RoBERTa pre-training technique and depends on the original architecture of BERT.

BERTweet combined with Linguistic Features: Prior study shows injecting linguistic feature to the model might be effective in terms of model's performance (Jin and Aletras, 2021). The approach is developed from Rahman et al. (2020), which uses a fusion mechanism known as the Multimodal Adaption Gate to combine multimodal information (such as auditory and visual) in transformers (MAG). Jin et al. (2022) describe how the injection is done by enlarging the linguistic data vectors to a size that is comparable to the embeddings provided to the pre-trained transformer, then employing MAG to combine contextual and language representations, and finally sending the output to a pre-trained BERTweet encoder for fine-tuning.

BERTweet-LIWC: The scope of this study includes the experiment with the linguistic feature named Linguistic Inquiry and Word Count (**LIWC**) (Pennebaker et al., 2001) injected to the BERTweet model. LIWC-2015 was used in the original study to represent each tweet as a 93-dimensional vector. We will re-use that to generate BERTweet-LIWC model for result analysis.

3.2 Data descriptions

In this study, we will use the English Tweet dataset provided by our base study conducted by Jin et al. (2022)[1].

As per our knowledge, there was no prior dataset for bragging detection. The authors of our primary study have manually collected the data used from *Twitter* using the premium Twitter Search API for academic research as these process is widely used in detection e.g. (Mohammad et al., 2018), (Mohammad et al., 2016), (Rosenthal et al., 2019). Two ways of data collection were followed: (1) Random sampling, (2) keyword based sampling. For (1), 10k data was collected per day for one year. They preserved 1% of daily twitter data and for (2) they used keywords to search data that increases the hit rate for bragging. Two different indicators of positive self-disclosure (e.g. I, just)and stylistic indicators, e.g. positive emotion words are applied as search keyword. After filtering and random sampling, 6696 tweets were finally preserved for the final experiment. These tweets contain bragging information from one of the sample classes. Data statistics are shown in Table 1[2]. The majority of the data set contains non-bragging data (88.34%) as tweets are not always intended to provide bragging information. Authors used the keyword based sampling data for training and random sampled data for test purposes which is nearly 50% of data for each part.

We will collect 200 more recent English tweets and annotate them according to the annotation process followed on the primary study. Our plan is to gather 40 tweets per day for 5 days and include them the test data set. This will enable us to conduct the ablation study and validate the model's consistency in detecting bragging with the recent data.

3.3 Hyperparameters

TODO

3.4 Implementation

The source code is not publicly available For the original study. We will implement the three transformer based models (BERT, RoBERTa, BERTweet) from scratch. However, we have received the code for BERTweet-LIWC code along with the LIWC linguistic feature file from the authors. This will help us compare the previous three models. We will generate BERTweet-LIWC model using that code and re-evaluate their claims.

Label	Training Set Keyword sampling	Dev/Test set Random sampling	All
Binary			
Bragging	544 (16.09%)	237 (7.15%)	781 (11.66%)
Not Bragging	2838 (83.91%)	3077 (92.85%)	5915 (88.34%)
Multi-class			
Achievement	166 (4.91%)	71 (2.14%)	237 (3.54%)
Action	127 (3.76%)	58 (1.72%)	185 (2.76%)
Feeling	39 (1.15%)	27 (0.82%)	66 (0.99%)
Trait	91 (2.69%)	48 (1.45%)	139 (2.08%)
Possession	58 (1.72%)	28 (0.84%)	86 (1.28%)
Affiliation	63 (1.86%)	5 (0.15%)	68 (1.01%)
Not Bragging	2838 (83.91%)	3077 (92.85%)	5915 (88.34%)
Total	3382	3314	6696

Table 1: Bragging Data Statistics

3.5 Experimental setup

TODO

3.6 Computational requirements

Pre-trained models like BERT, RoBERTa, BERTweet requires high amount of GPU resources. It can be speed up by using TPUs (Tensor Processing Units - Google's custom circuit built specifically for large ML models). Unfortunately we can not use TPUs as these resources are expensive to obtain. For our experiment, we are using GPU servers from *Google Cloud* with a base server of 8-core CPU and 24GB of memory. According to original BERT paper (Devlin et al., 2018), fine tuning a training dataset on GPU will require few hours. We are planning to use the resource about 10 hours for 7 days. For the reproduction study, our base paper did not provide any estimation about the computational requirements for generating their models. We are estimating the runtime according to the original BERT models.

4 Results

TODO

4.1 Result 1

TODO

4.2 Result 2

TODO

[1]Dataset link provided by the base study: https://archive.org/details/bragging_data

[2]The statistics are captured from Jin et al. (2022)'s dataset analysis

4.3 Additional results not present in the original paper

TODO

5 Discussion

TODO

5.1 What was easy

TODO

5.2 What was difficult

TODO

5.3 Recommendations for reproducibility

TODO

6 Communication with original authors

TODO

References

JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014. Self-disclosure topic model for classifying and analyzing Twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1986–1996, Doha, Qatar. Association for Computational Linguistics.

Daria Dayter. 2014. Self-praise in microblogging. *Journal of Pragmatics*, 61:91–102.

Daria Dayter. 2018. Self-praise online and offline. *Internet Pragmatics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Chiori Hori and Takaaki Hori. 2017. End-to-end conversation modeling track in dstc6.

Mali Jin and Nikolaos Aletras. 2021. Modeling the severity of complaints in social media. *arXiv preprint arXiv:2103.12428*.

Mali Jin, Daniel Preoţiuc-Pietro, A Seza Doğruöz, and Nikolaos Aletras. 2022. Automatic identification and classification of bragging in social media. *arXiv preprint arXiv:2203.05840*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

David Matley. 2018. "this is not a# humblebrag, this is just a# brag": The pragmatics of self-praise, hashtags and politeness in instagram posts. *Discourse, context & media*, 22:30–38.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Liyun Wen, Xiaojie Wang, Zhenjiang Dong, and Hong Chen. 2017. Jointly modeling intent identification and slot filling with contextual and hierarchical information. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 3–15. Springer.