A Pharmacovigilance Application of Social Media Mining: An Ensemble Approach for Automated Classification and Extraction of Drug Mentions in Tweets

PRESENTER: **Luis Alberto Robles Hernandez**

INTRODUCTION:

Using a social media like Twitter to perform this task is challenging since the content of the tweets may have misspellings or ambiguities

To address this task, we proposed an ensemble model to classify these tweets, as well as two approaches (a dictionary-based approach and a Named Entity Recognition BERT-based model) for the extraction of drug mentions from tweets

METHODS:

1. Data was collected from Critical Assessment of Information Extraction Systems in Biology (6), and Social Media Mining For Health Applications (SMM4H'18)
2. For the classification dataset, an ensemble model was implemented, containing transformer models
3. For the extraction process, two approaches were performed. In the first one, a matching process from a dictionary of slang terms for drug names was carried out on tweets classified in the positive class. While in the second approach, an NER BERT model was trained in order to extract the possible drug mentions from tweets

RESULTS:

As shown in ***Figure 1***, we can observe the ensemble model performed better (in the validation dataset) than any fine-tuned BERT model

We demonstrated that the ensemble model performed better on imbalanced data than any individual fine-tuned model

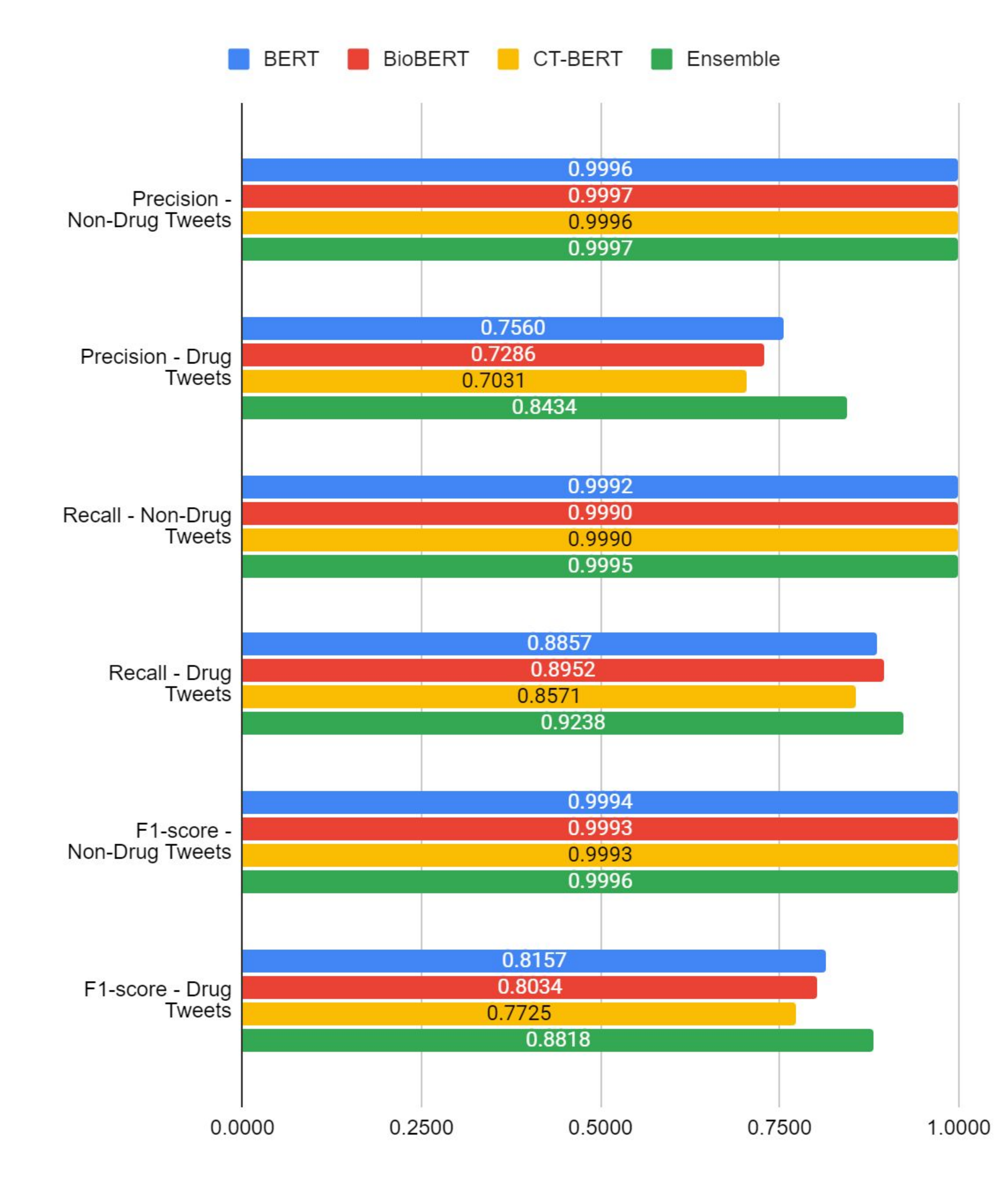

Figure 1*. Single model vs ensemble model performance*

Read the full paper by scanning the following QR Code

Additionally, from ***Figure 2*** we can see that using the first approach (with the help of the dictionary of slang terms for drug names) we were able to extract only 9 drug mentions. While in the second approach (by using a BERT model) we were able to extract 85 drug mentions. Therefore, we can clearly see the second approach performed a lot better

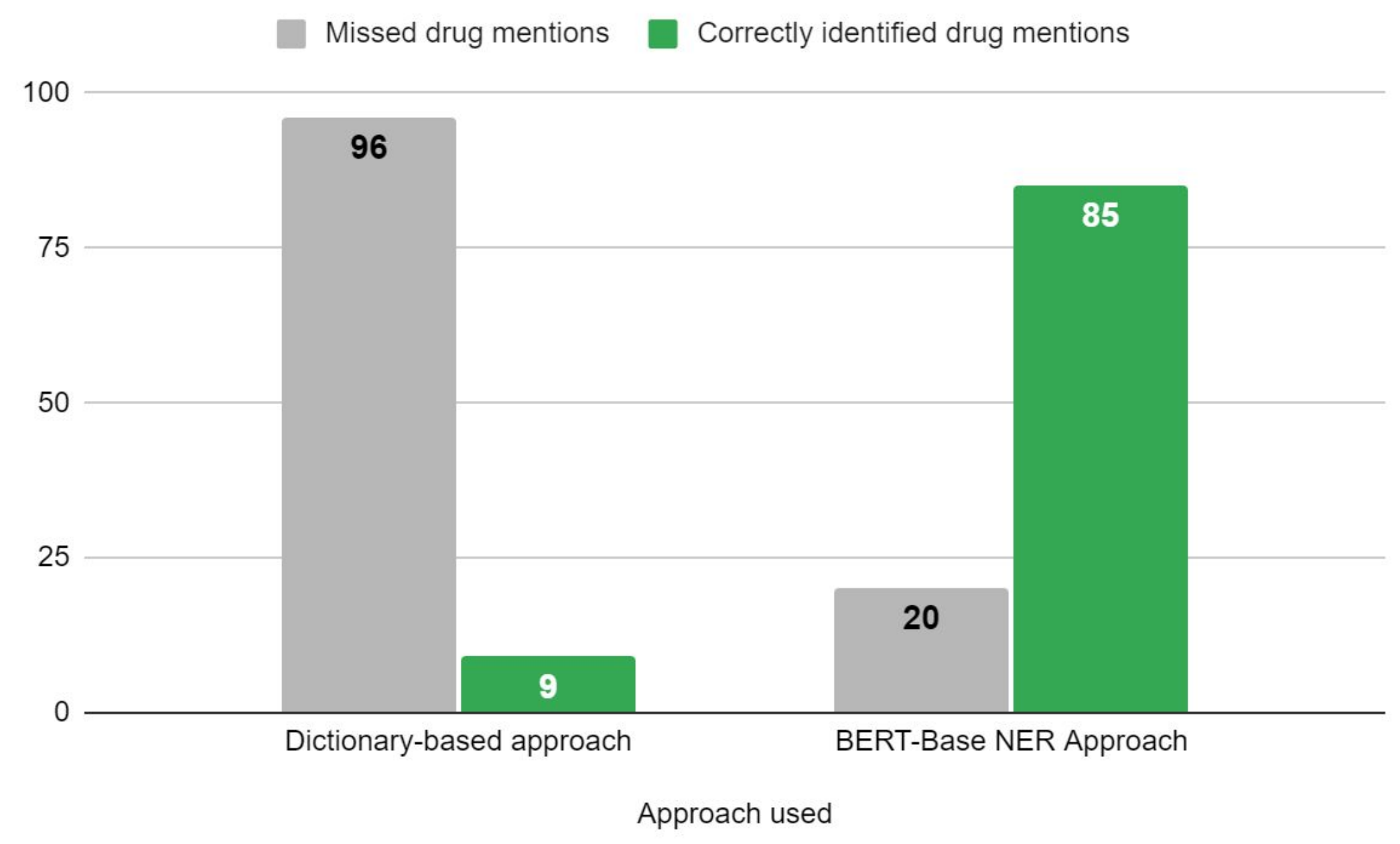

Figure 2*. Extraction performance comparison - Dictionary-based approach vs BERT-based approach*

CONCLUSION:

By creating automated approaches we can remove labor intensive manual curation, thus making the process faster. Furthermore, despite the very imbalanced classes presented in the training and validation dataset when implementing this automated approach, the ensemble model was able to perform better than any single fine-tuned model, specifically for the F1-score obtained

The second approach implemented by using a Named Entity Recognition BERT model, performed a lot better than using only a dictionary with drug slang terms. Also, additional steps can be done in order to extract even more drug mentions

REFERENCES:

For the references, visit the paper at: https://openreview.net/forum?id=0XuLGq933Y6

Luis Alberto Robles Hernandez, Rajath Chikkatur Srinivasa,
Juan M. Banda
Georgia State University, Atlanta, Georgia, USA

Georgia State University

NEURAL INFORMATION PROCESSING SYSTEMS