Myers-Brigg Personality Classification with Twitter Feed

Wesley Kwong

Abstract

The Myers-Brigg Type Indicator (MBTI) is a popular personality classification test but challenging to model. We applied monolingual and multilingual BERT models to classify German, Spanish, Italian, and Dutch tweets from the TwiSty dataset to predict the author's MBTI personality. We have observed that BERT outperform CNN models. Our results suggest that the multilingual BERT model is preferred for MBTI classification compared to multiple monolingual BERT models due to the ability to access to a larger and smoother dataset. We also discussed custom accuracy metrics to measure MBTI classification performance, the challenges presented from imbalanced data, and the difficulty of MBTI classification.

1 Introduction

The Myers-Brigg Type Indicator (MBTI), based on the conceptual theory proposed by psychologist Carl Jung, is a popular personality metric that classifies personality across four dichotomies. Introversion vs extroversion (I vs E) describes how people respond and interact with the world around them. Sensing vs intuition (S vs N) is a measure of how people gather information through their own senses or through pattern recognition. Thinking vs. feeling (T vs F) is a scale that reflects how people make decisions based on the information they collected. Finally, perception vs judging (P vs J) involves how people tend to deal with the outside world. The MBTI combines all four dichotomies into an overall personality type. For example, INTP individuals are described as innovative inventors with an unquenchable thirst for knowledge. These dichotomies lead to a total of 16 distinct MBTI personality classifications. Personality detection also has potential in marketing, politics, and psychological assessment. However, this test scales poorly towards a larger population since the test requires the individual to answer 93 questions which takes on average 25 minutes to complete.

This project aims to explore methods to classify people's MBTI based on their Twitter tweet through a Bidirectional Encoder Representations from Transformers (BERT) model, with a convolutional neural network (CNN) natural language processing model serving as a baseline [1]. However, the tweets are in Spanish, German, Dutch, and Italian. Furthermore, studies have shown that the distribution in MBTI classes vary widely across ethic groups with certain classes being rarer in certain populations compared to other populations leading to a highly unbalanced dataset [2]. Therefore, this project aims to explore multiple questions. First, how does monolingual (one language) BERT models perform? Second, how does the multilingual (all four languages) BERT model compare to average result of multiple monolingual BERT models? Finally, does the increased size of the training set, due to access to wider corpus of multilingual data, have a significant effect on the prediction accuracy of multilingual models?

2 Background

Historically, automated personality classification systems have been difficult to create due to the limited availability of labeled data. Given the wide array of possible personality classifications that can have a varying amount of overlap between classes, a significantly large data set is needed for classification accuracy. Furthermore, many labeled datasets are comprised of large written essays which can be computational expensive to model and generalize poorly outside the essay's subject matter [3]. Therefore, more recent work focused on the use of social media for personality classification due to its abundance of training data.

The earliest machine learning models for personality classification involved SVM with an accuracy of 37% on a reddit MBTI dataset containing 22.9 million comments [4]. More recently, a BERT model has been applied on a 68k dataset from PersonalityCafe's MBTI forums discussing their

classification with an accuracy of 47.9%. While BERT models have recently shown promise, more work is needed to improve their accuracy. Furthermore, PersonalityCafe's MBTI forums generalize poorly to other social media platforms that include other content [5]. Not only has classifying individuals based on text into one of 16 different categories, that somewhat overlap, is challenging, but studies have also shown that not all 16 personality types are uniformly distributed in the population. This can lead to potential overfitting due to the unbalanced dataset.

3 Methods

3.1 Dataset

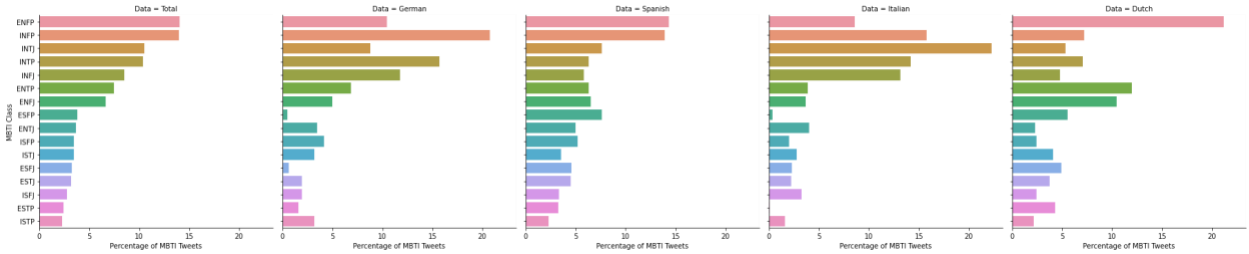

Figure 1: Distribution of MBTI classes in the multilingual dataset (left) and in monolingual datasets

The dataset used in this project was the TwiSty corpus generated. This dataset contains the tweet IDs of 34 million tweets by 18,168 authors in Dutch, German, French, Italian, Portuguese and Spanish [6]. Due to computational limitations, we sampled 250,000 Dutch, German, Italian and Spanish tweets for a total of 1 million tweets. With 80% of the sampled tweets for training, this leads to 200,000 tweets for training each monolingual model and 800,000 tweets for training the multilingual models. We used the tweepy Python package to scrape the corresponding tweet.

Figure 1 shows that distribution of MBTI classes greatly vary within and between languages. While this will present a challenge in modeling, we deemed the imbalanced nature of the data to be a feature to better reflect the varying personality distribution across culture. Furthermore, we can also determine whether the multilingual BERT model benefits from the "smoothened" dataset. To prevent the models from simply predicting the most frequent classes, we weighted each class based on their relative frequency using sklearn's class weight [7].

3.3 Baseline CNN Model

Before building a CNN model, the tweets must be first converted into tokens which is then converted into word vectors. Since Twitter tweets have user handles and often contain sentences that have repeated punctuation, the ntlk's tweet tokenizer was used to remove this and tokenize the tweet. The tokens are then converted into fastText's aligned word vectors that occupy what the authors call a "common space". This means that the word "chair" in German and French will have the same word vector. These vectors allow us to generate a multilingual CNN model without having to convert the tweets into one language [8, 9]. The CNN model has a convolutional layer, followed by a max polling layer, then a hidden layer of 100 neurons, and finally a sigmoid activation layer [10, 11]. Through hyperparameter optimization of all training tweets, it was determined that 30 epochs and a batch size of 256 works best for the CNN models. Categorical cross entropy was used as the loss function.

3.4 BERT Model

The following pretrained cased BERT models were downloaded from huggingface.co: Multilingual (bert-base-multilingual-cased), German (bert-base-german-cased), Spanish (dccuchile/bert-

base-spanish-wwm-cased), Italian (dbmdz/bert-base-italian-cased), and Dutch (GroNLP/bert-base-dutch-cased). The simple transformer Python package was used to create the data pipeline to input training and testing data into the various BERT models. Through hyperparameter optimization of all training tweets, it was determined that an epoch of 5 and a learning rate of 0.0001 worked best for BERT models. Just like in the CNN models, categorical cross entropy was used as the loss function.

3.5 Accuracy Metrics

Due to the nature of the MBTI classes and how classes can have varying dichotomic overlap, it was deemed that traditional model performance metrics would be misleading. Instead, we generated custom accuracy metrics, that to this authors knowledge, have not been used in measuring MBTI classification performance. There are five accuracy metrics used in this report: at least 1 match, at least 2 matches, at least 3 matches, perfect match, and average match. In the at least 1 match metric, at least 2 matches, and at least 3 matches metrics, this is the percentages of predictions that have at least 1, 2, and 3 dichotomies match respectively. A perfect match is the percentage of predictions that exactly predict all four dichotomies correctly. The average match is the percentage of accurately predicted dichotomies over the total number of possible dichotomies. If the true class was INTP and the predicted class was INTJ (one dichotomy was incorrectly predicted), then the metrics at least 1 match, at least 2 matches, at least 3 matches will all be 100%. Perfect match will be 0% and average match will be 75%. While the perfect match metric is more useful for applications purposes, the average match is used to summarize the overall classification accuracy of various models when accounting for the dichotomic nature of MBTI personality classes.

4 Results and Discussion

4.1 Monolingual CNN vs Monolingual BERT

Figure 2: Catplot presenting average monolingual BERT improvement across accuracy metrics

Generally, both model types tended to perform worse as the accuracy metric became more stringent on the number of correct dichotomies. The average match metric had a higher accuracy compared to the at least 3 matches and perfect match metrics. Overall, all four monolingual BERT models consistently achieved a higher accuracy on all accuracy metrics compared to all four CNN models as shown in Figure 2. To summarize the results, the accuracy metrics were averaged from all four monolingual CNN models and all four monolingual BERT models. The at least 3 matches metric shows the largest improvement compared to the CNN model with a 32% improvement in accuracy on the training set and a 17.6% improvement on the test set. The second largest improvement is the perfect match metric with a 32% improvement for training and a 10.5% for test set. The only area where CNN performed on par with BERT is on the at least 1 match metric where all monolingual models got 100%

accuracy. The CNN models also performed well on the at least 2 matches metric and averaging around 74.4% in training and 72.2% in testing. BERT improved on this metric by about 10% on average.

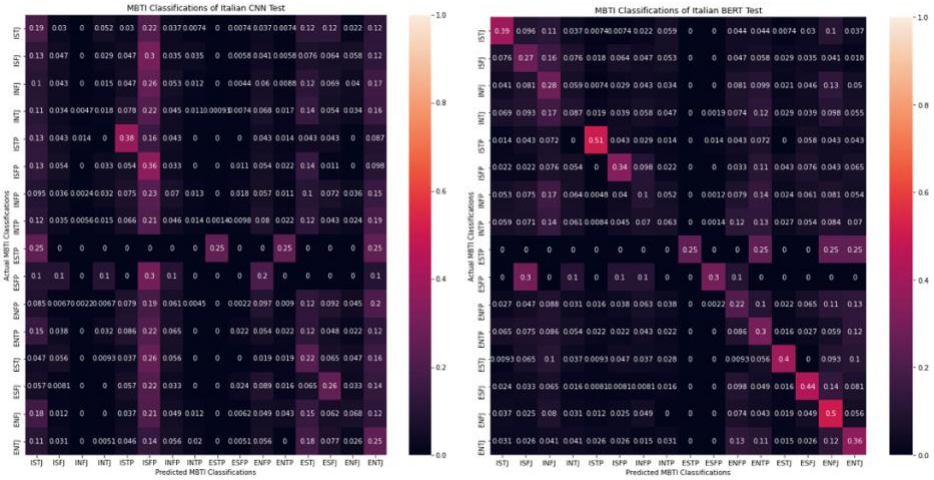

Figure 3: Confusion matrices of Italian test results from CNN (left) and BERT (right) models

There are a couple of reasons why the CNN models performed poorly compared to the BERT models. First, CNN models are highly sensitive to the MBTI class imbalance present within the dataset. While weighting by class distribution prevented both CNN and BERT models from predicting based on the most frequent classes, the CNN models as shown in Figure 3 tend to predict frequently on the rarest class. The strongest vertical column ISFP made up 5% of the Italian data. But there are barely any predictions for INFJ which make up 15% of the classes. Meanwhile, we do not observe this in BERT models, likely due to higher perfect match accuracy. Second, the CNN models do not have a strong enough resolution to detect the contextual signal present within the tweets for personality classification. Despite increasing the number of epochs to over 1,000, the CNN model could not overfit the training set. In contrast, the BERT model only needs a few epochs to overfit the training set.

4.2 Monolingual BERT vs Multilingual BERT

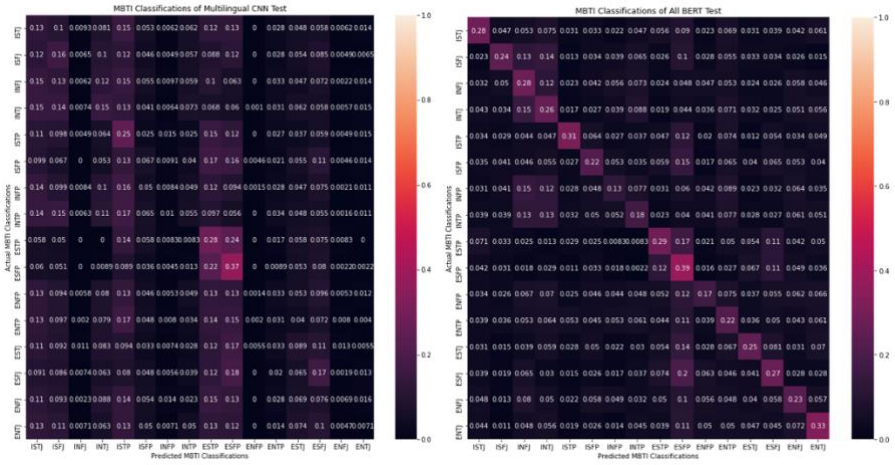

Figure 4: Confusion matrices of test results from CNN (left) and BERT (right) multilingual model

The multilingual models also showed similar tends seen in the monolingual models. BERT models consistently improved every metric except for the at least 1 metric which both models got 100% accuracy. At least 3 matches saw the largest improvement with 29% increase in training and 17.6% in test. The second largest improvement is the perfect metric with 27.7% increase in training and 12.2% in test. As previously stated, the theorized benefit of the multilingual model over monolingual models is that class imbalance within the dataset is smoothened. When comparing the multilingual models to the averaging of the monolingual models across accuracy metrics, there is no significant difference in results. We observe the multilingual CNN also predicts on least frequent class like the monolingual CNN models. But more interestingly, the right confusion matrix of Figure 4 reveals that the multilingual BERT model predicts all classes accurately 20% of the time. In comparison, the monolingual BERT models, as shown in Figure 3, have varying prediction accuracy across the personality classes. This suggest that while the smoothened dataset may not lead to significant improvement across accuracy metrics, it does allow BERT to better predict equally both common and rare personality classes. We recommend in future studies to ensure the dataset is balanced for improved MBTI classification performance for BERT models.

4.3 Training Size Variation

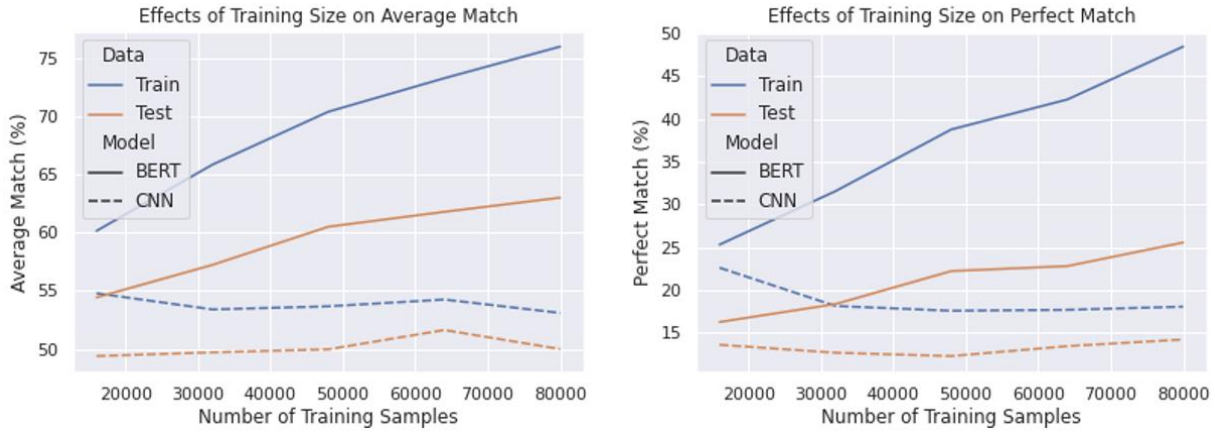

Figure 5: Effect of increasing number of training samples on multilingual model performance

Since the multilingual dataset was four times as large as the monolingual training set, we investigated the effects of increasing the number of training samples. The multilingual CNN and BERT models were trained on increments of 20,000 samples and evaluated on the entire multilingual training and test set. The training samples were also randomly sorted so tweets are not in order by language. Figure 5 shows that BERT consistently has a higher average and perfect match metric, which has been previously discussed. More interestingly, the multilingual BERT model shows a significant improvement across both metrics as the training size increases while the multilingual CNN model stays relatively constant. This trend is likely due to the BERT architecture better able to capture the context present within the tweet for personality classification. As stated before, CNN models have a finite amount of signal that they can capture from tweets and thus the relative amount of information gained from increasing the training size is minimal. The graphs also suggest that increasing the number of training samples further will likely continue to yield improved accuracy for the multilingual BERT model. Given that the average accuracy and perfect match accuracy metrics improved on the test data by 8% and 9% respectively when given a four-fold increase in the training size, an exponentially larger dataset is needed. This should not be a major concern since a larger pool of tweet data would be available for multilingual models.

4.4 Misclassified Tweets Analysis

Language	Translated Tweet	Predicted MBTI	Actual MBTI
German	*yes? sure? I had the feeling 1x that the packaging looked different. and then I was a little unsettled*	INFJ	INFP
Italian	*So it seems that only Catania voted Musumeci. #elesicilia*	INFJ	INTP
Spanish	*Nothing, I'm back and I love you because you didn't unfollow me ...*	INTJ	INTP

Table 1: Examples of misclassified tweets by the multilingual BERT model

Despite the improvement of multilingual BERT, it has been difficult to perfectly classify tweets. Some of the misclassifications are easy to identify like the German tweet in Table 1. INFJ individuals are described to be serious people who don't like surprises so it makes sense that the model would misclassify the German tweet. But the other tweets are examples of personality classes where it is difficult to find a strong difference between them. The Italian tweet is misclassified by two dichotomies. However, there is still a significant amount of personality overlap between INFJ and INTP. INFJ is predicted by the model, possibly due to the tweet's serious tone. However, the author is actually INTP. These individuals are described to be quiet and analytical, which also fits the tweet. The Spanish tweet is also difficult to parse out despite the difference in one dichotomy of P vs J. The tweet can be interpreted with a (J) judging tone accusing someone on Twitter or the author's analytical (P) perceptiveness of the situation. Overall, this table is a small example that illustrates the wider problem behind the difficulty of creating highly accurate MBTI models. MBTI cannot separate people into clear, definitive personality groups since people react to different situations differently based on a combination of factors. Just because someone was classified in the P vs J dichotomy as judging, it does not mean that people can be also more perceptive in certain situations.

5 Conclusion

In this study, we applied various monolingual and multilingual models to classify German, Spanish, Italian, and Dutch tweets to predict the author's MBTI personality type. We also presented novel MBTI accuracy metrics to better interpret MBTI model performance. Overall, BERT models performed better over their baseline CNN counterparts across the custom accuracy metrics. We observed there was not a significant difference between the multilingual BERT and the averaged results of the monolingual BERT. However, we determined that the unbalanced nature of our monolingual dataset, despite applying a class frequency weighting scheme, contributed towards a significant imbalance in prediction especially in the CNN models. This imbalanced is addressed by the smoothing of MBTI classes in the multilingual dataset which led to the multilingual BERT model able to predict rare and common MBTI classes. We also examined multilingual model performance based on training set size and determined that multilingual BERT model performance would continue to improve when provided more data. Finally, we also examined some misclassified tweets and discuss the difficulty of modelling the MBTI classification system. For all these reasons, we conclude that the multilingual BERT model should be used for future MBTI classification research.

For future studies, we recommend implementing stratified sampling and a significantly larger tweet dataset for further model prediction improvement. We suggest investigating multilingual BERT on each language individually to examine model performance. Finally, we also wanted to investigate customized loss functions that takes advantage of the varying dichotomous nature of MBTI classes instead of utilizing cross entropy loss function for improved model fitting.

Reference

1. J. Devlin, M. Chang, K. Lee and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
2. A. Hammer and W. Mitchell. 1996. The Distribution of MBTI Types In the US by Gender and Ethnic Group
3. S. Nowson and A. Gill. 2014. Look! Who's talking?: Projection of extraversion across different social contexts.
4. M. Gjurkovic and J. Šnajder. 2018. Reddit: A gold mine for personality prediction.
5. S. Keh and I. Cheng. 2019. Myers-Briggs Personality Classification and Personality-Specific Language Generation Using Pre-trained Language Models
6. B. Verhoeven, W. Daelemans and B. Plank. 2016. TwiSty: a Multilingual Twitter Stylometry Corpus for Gender and Personality Profiling
7. G. King and L. Zeng. 2001. Logistic Regression in Rare Events Data
8. A. Joulin, P. Bojanowski, T. Mikolov, H. Jegou, and E. Grave. 2018. Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion
9. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching Word Vectors with Subword Information
10. Y. Kim. 2014. Convolutional Neural Networks for Sentence Classification
11. CNN code: https://github.com/datasci-w266/2021-spring-main/blob/master/materials/lesson_notebooks/lesson_5_cnn.ipynb

Appendix

Trial Number	Epoch	Batch	Average Accuracy Train	Average Accuracy Test	Training Loss	Test Loss
0	30	128	51.9	49.4	2.60	2.83
1	**30**	**256**	**54.0**	**51.3**	**2.55**	**2.75**
2	30	512	50.6	48.6	2.66	2.81
3	40	128	53.3	50.3	2.54	2.84
4	40	256	51.9	49.2	2.56	2.83
5	40	512	53.8	51.0	2.55	2.76
6	50	128	53.8	50.2	2.52	2.88
7	50	256	52.9	50.0	2.53	2.83
8	50	512	50.8	48.3	2.62	2.86

Table A1: Hyperparameter Tuning of CNN Models with Multilingual Data

Trial Number	Learning Rate	Epochs	Training Loss	Test Loss
1	0.001	5	2.77	2.77
2	**0.0001**	**5**	**2.70**	**2.70**
3	0.001	4	2.78	2.77
4	0.0001	4	2.77	2.77
5	0.001	3	2.77	2.77
6	0.0001	3	2.76	2.76

Table A2: Hyperparameter Tuning of BERT Models with Multilingual Data

	Data	Language	Model	Number of Samples	At Least 1 Match	At Least 2 Matches	At Least 3 Matches	Perfect Match	Average Match
1	Train	Dutch	CNN	20000	100	73.56	33.07	14	49.06
2	Test	Dutch	CNN	5000	100	71.72	30.46	11.02	46.96
3	Train	German	CNN	20000	100	71.04	35.76	18	47.56
4	Test	German	CNN	5000	100	69.76	32.02	13.86	45.07
5	Train	Italian	CNN	20000	100	75.72	38.2	15.83	52.6
6	Test	Italian	CNN	5000	100	74.52	33.88	11.42	50.42
7	Train	Spanish	CNN	20000	100	77.44	43.06	19.58	55.05
8	Test	Spanish	CNN	5000	100	72.68	34.74	12.36	48.48

Table B1: Results of monolingual CNN models

	Data	Language	Model	Number of Samples	At Least 1 Match	At Least 2 Matches	At Least 3 Matches	Perfect Match	Average Match
1	Train	Dutch	BERT-Base-Dutch-Cased	20000	100	93.08	69.11	47.91	77
2	Test	Dutch	BERT-Base-Dutch-Cased	5000	100	85.14	50.54	22.52	62.17
3	Train	German	BERT-Base-German-Cased	20000	100	97.71	85.61	71.66	88.68
4	Test	German	BERT-Base-German-Cased	5000	100	87.7	58.98	32.66	68.34
5	Train	Italian	BERT-Base-Italian-Cased	20000	100	91.7	65.14	40.63	73.31
6	Test	Italian	BERT-Base-Italian-Cased	5000	100	85.1	51.28	22.2	62.12
7	Train	Spanish	BERT-Base-Spanish-Cased	20000	100	86	59.04	35.25	67.4
8	Test	Spanish	BERT-Base-Spanish-Cased	5000	100	77.48	40.64	13.22	52.72

Table B2: Results of monolingual BERT models

	Data	Language	Model	Number of Samples	At Least 1 Match	At Least 2 Matches	At Least 3 Matches	Perfect Match	Average Match
1	Train	All	CNN	80000	100	76.19	40.12	17.99	52.52
2	Test	All	CNN	20000	100	74.63	36.33	13.3	49.65
3	Train	All	BERT-Base-Multilingual-Cased	80000	100	91.25	69.07	45.66	74.87
4	Test	All	BERT-Base-Multilingual-Cased	20000	100	85.26	53.89	25.5	63.08

Table C1: Results of multilingual CNN and BERT models

	Data	Model	At Least 1 Match	At Least 2 Matches	At Least 3 Matches	Perfect Match	Average Match
1	Train	Averaged Monolingal BERT	100	92.1	69.7	48.9	76.6
2	Train	Multilingual BERT	100	91.3	69.1	45.7	74.9
3	Test	Averaged Monolingal BERT	100	83.9	50.4	22.7	61.3
4	Test	Multilingual BERT	100	85.3	53.9	25.5	63.1

Table C2: Results comparing averaged monolingual BERT models to multilingual BERT model