## Review #1

### Summary of Weaknesses\*

1. I am not sure if the paper as is really fits into the selected track ("Low-resourced and Less Studied Languages") which – in my understanding – focuses on languages that have a lack of available material and linguistic resources and solutions that address the resulting problems. This paper seems to focus strongly on "low-resourced" in the sense of low-resourced technical environments (which in itself is obviously very relevant). The sense of "not enough available data" is in the current state of the paper only mentioned in form of some side notes.
   * We make an artificial scarcity of data by randomly selecting 500 samples where 250 samples are positive (1) and 250 samples are negative (0). Then, we split these 500 samples into 80-20 ratios for train and testing. The reason for this method is to make an artificial environment where we do not have enough information or data, but with our models and procedures, it can work almost accurately.
   * From the technical environment aspect, we tried to keep track of the computational units, CO2 emission, and power consumption (some values are shown in Table 1). Physical tests on cutting-edge devices are part of our next work plan.

### 

## Review #2

### Summary of Weaknesses\*

1. Reproducibility Concerns: There are mentions of hardware switching during experiments (Google Colab T4/P100 GPU), which may introduce variability in energy and timing measurements, raising concerns about reproducibility.
   * Yes, we could not simulate them with only one GPU due to a huge number of simulations. However, according to the reproducibility of QQP and newly tested CoLA datasets, which were trained on only T4 GPU, we can observe that the dissimilarity is minimal compared to QNLI and SST2 datasets. Furthermore, most of our final simulations and results were taken using the T4 GPU.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Models | Using | Mean Accuracy | Variance | Standard Deviation |
| QQP | LSTM+LSTM | Fourier | 60.80 | 1.96 | 1.19 |
| Max Pooling PCA | 63.44 | 2.25 | 1.50 |
| CNN+LSTM | Fourier | 62.15 | 2.02 | 1.42 |
| Max Pooling PCA | 63.14 | 1.42 | 1.19 |
| BILSTM+LSTM | Fourier | 62.59 | 0.83 | 0.91 |
| Max Pooling PCA | 63.99 | 2.00 | 1.41 |
| DistilBERT |  | 58.20 | 22.76 | 4.77 |
| TinyBERT |  | 55.75 | 18.59 | 4.31 |
| CoLA | LSTM+LSTM | Fourier | 59.70 | 1.22 | 1.10 |
| Max Pooling PCA | 59.59 | 4.34 | 2.08 |
| CNN+LSTM | Fourier | 62.15 | 0.12 | 0.35 |
| Max Pooling PCA | 65.20 | 4.36 | 2.09 |
| BILSTM+LSTM | Fourier | 62.54 | 0.74 | 0.86 |
| Max Pooling PCA | 60.65 | 0.63 | 0.79 |
| DistilBERT |  | 57.65 | 19.83 | 4.45 |
| TinyBERT |  | 50.40 | 14.24 | 3.77 |
| QNLI | LSTM+LSTM | Fourier | 63.10 | 9.07 | 3.01 |
| Max Pooling PCA | 63.25 | 10.27 | 3.21 |
| CNN+LSTM | Fourier | 62.09 | 8.29 | 2.88 |
| Max Pooling PCA | 64.60 | 4.85 | 2.20 |
| BILSTM+LSTM | Fourier | 63.09 | 2.79 | 1.67 |
| Max Pooling PCA | 65.70 | 4.61 | 2.15 |
| DistilBERT |  | 52.80 | 20.26 | 4.50 |
| TinyBERT |  | 52.20 | 19.56 | 4.42 |
| SST2 | LSTM+LSTM | Fourier | 69.69 | 7.61 | 2.76 |
| Max Pooling PCA | 69.50 | 8.16 | 2.86 |
| CNN+LSTM | Fourier | 65.74 | 7.59 | 2.75 |
| Max Pooling PCA | 65.65 | 15.42 | 3.93 |
| BILSTM+LSTM | Fourier | 66.20 | 6.25 | 2.50 |
| Max Pooling PCA | 65.69 | 18.22 | 4.27 |
| DistilBERT |  | 69.95 | 34.65 | 5.89 |
| TinyBERT |  | 60.90 | 33.79 | 5.81 |

**Table- Reproducibility** **test for four different datasets**

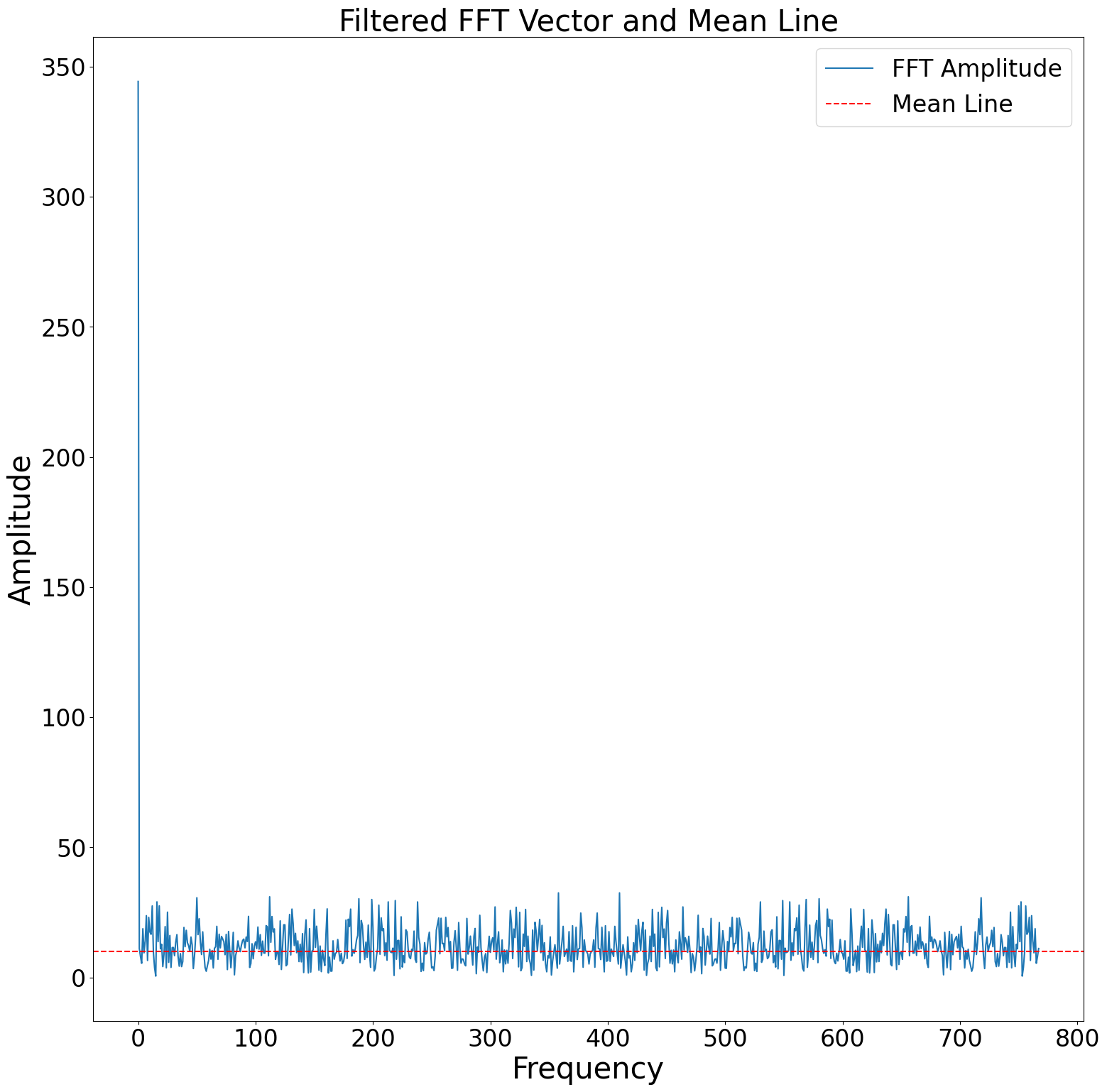
1. Lack of Dataset Diversity: The evaluation focuses only on a limited set of GLUE datasets, which may not provide a comprehensive view of the model's performance across different NLP tasks and datasets.
   * Though these are limited sets of GLUE datasets, the domains for each dataset are different: SST-2 → sentiment; QQP → paraphrase; QNLI → Question Answering.
   * Additionally, we check the reproducibility (shown in the above Table) and some of the other metrics for the CoLA dataset, which also aligns with our previous findings for other datasets.
   * Testing on other types of datasets is part of our next work plan.
2. Limited Explanation of Trade-offs: Although energy efficiency is discussed, the trade-offs between accuracy and model size are not sufficiently explored, especially in terms of how much accuracy is sacrificed for efficiency.
   * They are used for Figure 3(a). Here, we can see that the trained model size is already a few hundred KB before pruning, which gives good accuracies for the following merging functions with minimal power consumption.
   * The application of pruning the model is part of our next work plan.

|  |  |  |  |
| --- | --- | --- | --- |
| **Mean: CNN+LSTM** | | | |
| Merge Embeddings | Accuracy | Power (kWh) | **Size (MB)** |
| (x\*y) | 63.660 | 0.016 | 0.440 |
| (x/y) | 62.191 | 0.029 | 0.451 |
| (x+y) | 64.525 | 0.013 | 0.440 |
| log(x\*y) | 63.593 | 0.017 | 0.440 |
| log(x/y) | 63.193 | 0.020 | 0.451 |
| log(x+y) | 63.327 | 0.017 | 0.439 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Mean: LSTM+LSTM** | | | |
| Merge Embeddings | Accuracy | Power (kWh) | **Size (MB)** |
| (x\*y) | 62.957 | 0.026 | 0.656 |
| (x/y) | 65.591 | 0.035 | 0.673 |
| (x+y) | 64.327 | 0.033 | 0.657 |
| log(x\*y) | 63.459 | 0.027 | 0.657 |
| log(x/y) | 63.459 | 0.026 | 0.671 |
| log(x+y) | 63.991 | 0.031 | 0.656 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Mean: BILSTM+LSTM** | | | |
| Merge Embeddings | Accuracy | Power (kWh) | **Size (MB)** |
| (x\*y) | 62.327 | 0.035 | 0.828 |
| (x/y) | 64.791 | 0.044 | 0.834 |
| (x+y) | 64.725 | 0.033 | 0.828 |
| log(x\*y) | 62.257 | 0.038 | 0.828 |
| log(x/y) | 62.791 | 0.043 | 0.834 |
| log(x+y) | 64.059 | 0.038 | 0.828 |

1. Spectral Analysis Underdeveloped: The spectral analysis step, which reduces the dimensionality of contextual embeddings, is mentioned as crucial but lacks an in-depth explanation or visual representation of how it benefits the overall model performance.
   * In embeddings, each dimension can capture a different aspect of word meaning, similar to how each frequency in a Fourier series captures a different signal aspect.
   * According to this plotting, we are taking embedding for one sentence and applying the Fourier transformer. This method provides the frequency spectrum of the embedding file, focusing solely on the words with a frequency level exceeding 10%, which helps to reduce the dimensions. It determines the importance of the words for the context.

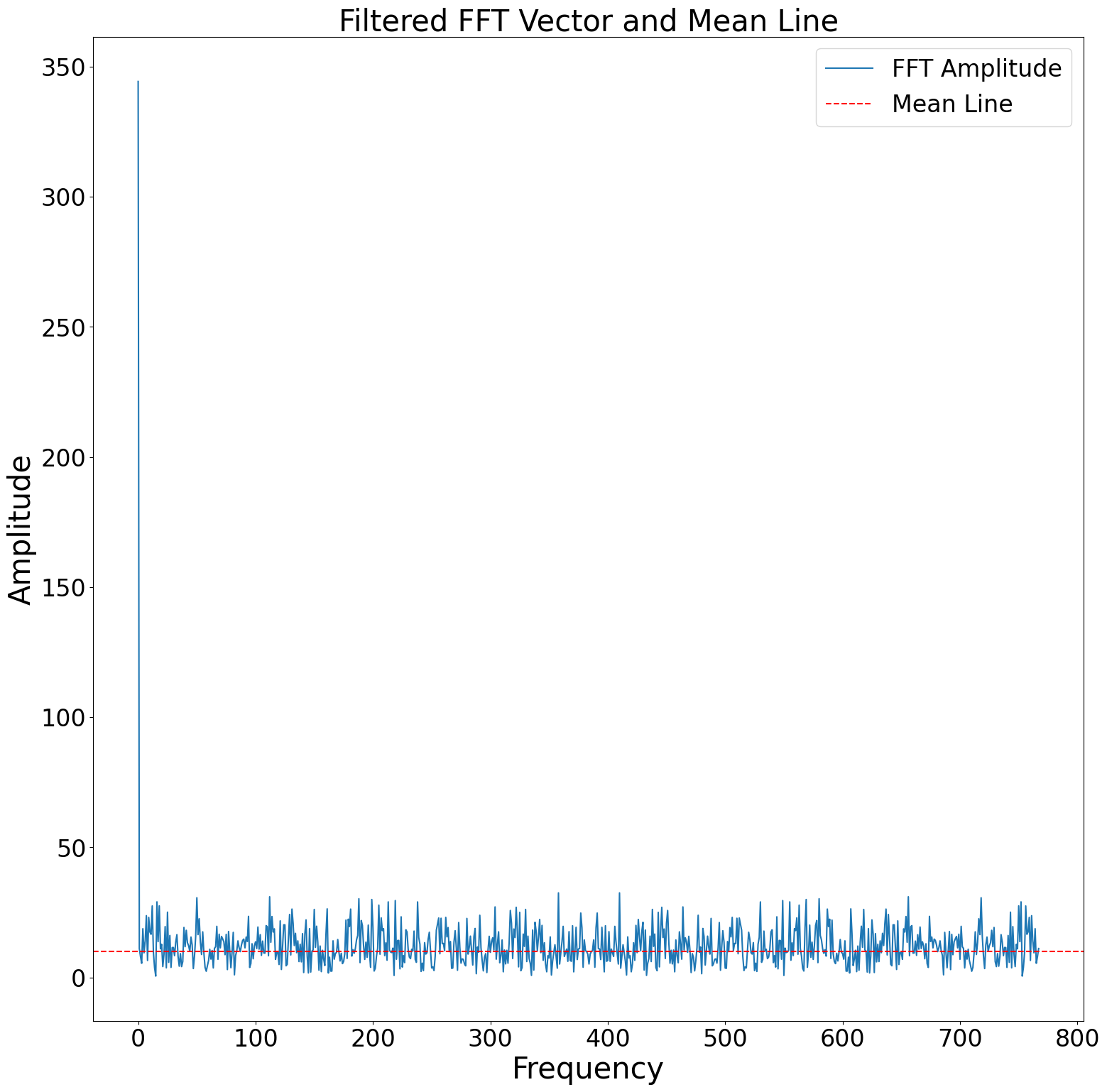


1. No Clear Baseline Comparisons: Although the paper compares blended embeddings with models like DistilBERT, it could benefit from a clearer comparison with purely static or purely contextual embeddings in similar experimental setups to quantify the exact gain.
   * These kinds of comparisons are shown in Table 6 (Blend PCA 75, Contextual PCA 75, Glove PCA 75, Purely Static Glove 300).
   * For the selected two models using Figure 3 (selection process explained in the appendix, Figure 4), we showed n = 75, 100, 150, 300 dimensions for log(x/y), log(x+y), purely contextual, and purely static in Table 9, Table 10, and Table 11.

## Review #3

### Summary of Weaknesses\*

1. While the paper is generally easy to read, it is sometimes difficult to see how everything hangs together. The GLUE dataset is mentioned, but it is unclear exactly how these datasets are tested, and how the models are trained.
   * We make an artificial scarcity of data by randomly selecting 500 samples where 250 samples are positive (1) and 250 samples are negative (0). Then, we split these 500 samples into 80-20 ratios for train and testing. The reason for this method is to make an artificial environment where we do not have enough information or data, but with our models and procedures, it can work almost accurately.
   * Then, we train the model with 50 trails where the base lr is 1e-3 and the max lr is 6e-3 and perform a random search from where we get the best accuracy for each model.
2. The role of the spectral analysis and the frequency component could be made more concrete. It is not tied to well to other parts of the paper. I am not so familiar with the technique, so for those who are more familiar, this might be clearer, but a few sentences making this more explicit would be nice for the general reader, I think. For example, it is not immideately clear to me how Fourier series and embeddings are related, but again, this might just be me.
   * In embeddings, each dimension can capture a different aspect of word meaning, similar to how each frequency in a Fourier series captures a different signal aspect.
   * According to this plotting, we are taking embedding for one sentence and applying the Fourier transformer. This method provides the frequency spectrum of the embedding file, focusing solely on the words with a frequency level exceeding 10%, which helps to reduce the dimensions. It determines the importance of the words for the context.

****

1. The training and experiments are a bit difficult to understand. I am not sure I understand how it was done exactly.
   * For this experiment,we make an artificial scarcity of data by randomly selecting 500 samples where 250 samples are positive (1) and 250 samples are negative (0). Then, we split these 500 samples into 80-20 ratios for train and testing. The reason for this method is to make an artificial environment where we do not have enough information or data, but with our models and procedures, it can work almost accurately.
   * Then, we train the model with 50 trails where the base lr is 1e-3 and the max lr is 6e-3 and perform a random search from where we get the best accuracy for each model.
   * Table 3 checks how the max pooling PCA and embedding sort works after applying FFT. The observation is that applying PCA is more helpful than sorting.
   * Table 4 shows the comparison of the blending embedding using several merging functions, and later on, the observation is that the embedding file with 75 dimension outperforms others.
2. I cannot help but miss some discussion of the semantics of the new embeddings, not just in terms of model results, but a more direct comparison of their semantics.

## Review #4

### Summary of Weaknesses\*

1. Unfortunately, the paper does not provide any hypotheses about why the combination of static and contextualized embeddings would be informative and what it would mean to combine them. I find it very difficult to understand how the combined embeddings relate to other, contextualized-only solutions in terms of the information they can provide.
   * Combined embedding can capture both static and context-specific meanings
   * It will be helpful to reinforce the Out of Vocabulary (OOV) words of static embedding
2. It is not quite clear to me how the proposed solution addresses low-resource scenarios, as the method explained in Section 2.2 seems to rely on pre-trained BERT embeddings after all. It is possible that I misunderstood something.
   * We are not making any new embeddings but extracting the embeddings from the pre-trained BERT for different datasets. And throughout the process, we are reducing the embedding file size that can be usable on cutting edge low-resource devices.
   * The low resource scenario also implies data scarcity, where we artificially create data scarcity for our datasets and perform this experiment.
3. The low-resource solution proposed in the paper has only been evaluated on English datasets. As far as I understand, no low-resource scenarios have been simulated. I find it difficult to draw conclusions about low-resource scenarios based on such experiments.
   * We make an artificial scarcity of data by randomly selecting 500 samples where 250 samples are positive (1) and 250 samples are negative (0). Then, we split these 500 samples into 80-20 ratios for train and testing. The reason for this method is to make an artificial environment where we do not have enough information or data, but with our models and procedures, it can work almost accurately.
   * From the technical environment aspect, we tried to keep track of the computational units, CO2 emission, and power consumption (some values are shown in Table 1). Physical tests on cutting-edge devices are part of our next work plan.

Overall, I think the approach presented in the paper may have potential, but at this point, I find it very difficult to understand what the experiments can show, in particular with respect to low-resource scenarios. I could imagine that the experiments might actually offer insights about the nature of static and contextualized embeddings and what information they can provide, but this is not really apparent from the current version of the paper. Perhaps these questions could be taken into account in a revised version.