Table 3 reports the average F1-measure, precision, and recall for all the models across the six topics in PHM2017 dataset. The results show that the main improvement of *WESPAD* comes from the higher recall, i.e., detecting additional true health mentions. Table 3 also shows that the highest precision is achieved by the simple *ME+lex* model, since this model only relies on the lexical features. On the other hand, *LSTM+GRNN* has the lowest precision, and this can be attributed to the complex structure of the network which expects to be fine-tuned during the training.

Model	F1	Precision	Recall
ME+lex	0.572	0.834	0.462
ME+cen	0.536	0.819	0.429
ME+lex+emb	0.593	0.833	0.483
ME+lex+cen	0.594	0.827	0.493
LSTM-GRNN	0.615	0.638	0.605
FastText	0.630	0.802	0.538
CNN	0.673	0.794	0.610
WESPAD	**0.695**	0.803	**0.628**

Table 3: Average F1-measure, precision, and recall in PHM2017 dataset.

Table 4 reports F1-measure, precision, and recall of all the baselines in comparison to *WESPAD* in FLU2013 dataset. The results show that *WESPAD* outperforms all the baselines, even though there are considerable differences between PHM2017 and FLU2013 datasets (in terms of the proportion of the positive tweets). The results also show that *WESPAD* performs slightly better than the disease-specific *Rules* classifier, implemented according to the descriptions in reference [22]. More detailed analysis revealed that the syntactic subtrees that we use in our model, to some extent, can also automatically capture the manually designed patterns reported in [22]. It is also worth mentioning that, all the improvements of *WESPAD* model over the lexical baseline *ME+lex* in both datasets are statistically significant using paired t-test at $p<0.05$.

The comparison between the relative improvement of *WESPAD* in PHM2017 and FLU2013 datasets shows that our model performs significantly better in PHM2017 dataset. The improvement can be attributed to the inherent differences between these two datasets, and the fact that PHM2017 is highly imbalanced and FLU2013 is nearly balanced. We discuss this issue further in the next section.

6.2 Discussion

We now analyze the performance of *WESPAD* in more detail, focusing on the effects of the word embeddings partitioning, contribution of different features, and the ability of *WESPAD* to generalize from few positive examples in training.

Word embeddings partitions: in Section 3.3 we argued that large partitions can increase recall, and degrade precision. To support the argument, we fixed the values of α and α_2; and experimented with different values for K (the number of the partitions in the regular embeddings space) and K_2 (the number of the partitions in the distorted embeddings space). Figure 3 illustrates the result of this experiment. To be able to easier interpret the results, we also set K to be equal to K_2. The experiment confirms that by decreasing the number of partitions (and thereby increasing the partition sizes),

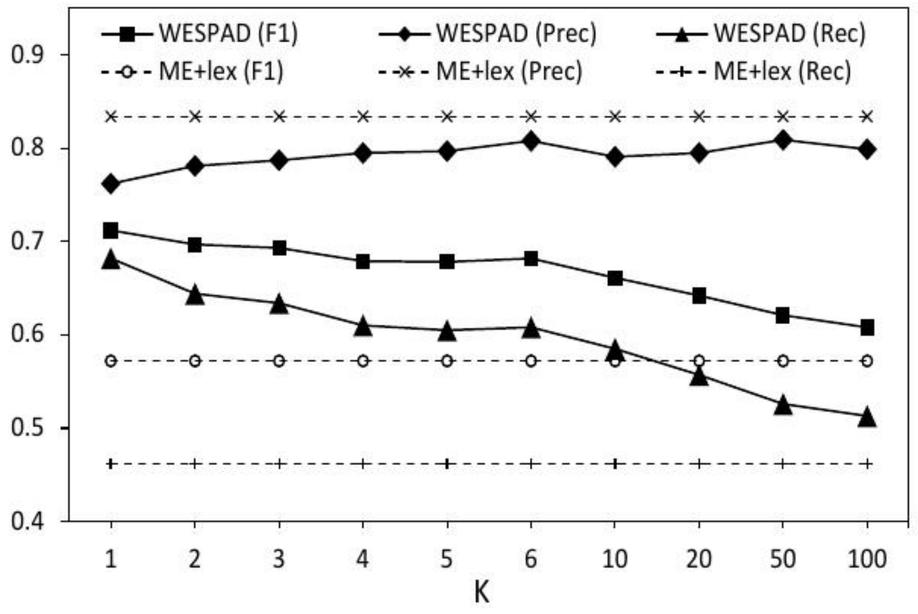

Figure 3: Impact of the number of partitions K on *WESPAD* on F1-measure, precision, and recall (PHM2017 dataset).

the *Recall* of *WESPAD* improves. However, this comes at the cost of degrading the *Precision* (specifically at $K = 1$).

Feature ablation: Table 5 reports the result of the ablation study on the features in *WESPAD* model in PHM2017 dataset. The experiment shows that *we_distortion* and *we_partitioning* feature sets have the highest impact, in terms of F1-measure. We also observe that, in terms of precision, *we_partitioning* performs better than *we_distortion*. One possible explanation is that due to the small size of the positive sets, IG-weighting may fail to accurately assign the weights to the word vectors, and thus, the tweet centroid is drifted.

Effect of the number of positive examples: in Section 6.1 we observed that the relative improvement of *WESPAD* in PHM2017 dataset is considerably higher than its relative improvement in FLU2013 dataset. We argue that since FLU2013 dataset is nearly balanced, and also has a substantially larger set of positive tweets, simple models such as *ME+lex* can perform relatively well. To analyze the effect of the size of the training data, and specifically the availability of true positive examples, we varied the number of the positive examples in the training folds, by randomly sampling from 10% to 90% of the positive examples (and keeping all of the negative examples), and re-trained *WESPAD* , *Rules*, and *ME+lex* in the reduced training sets in FLU2013 dataset. Figure 4 reports the values of the F1-measure for *ME+lex*, *Rules*, and *WESPAD* at varying fractions of the positive tweets used in the training data. The experiment shows that at smaller fractions of available positive tweets (10%-30%), *WESPAD* dramatically outperforms the *ME+lex* baseline, demonstrating that *WESPAD* is able to generalize from fewer positive training examples. *WESPAD* also significantly outperforms *Rules* at small fractions of positive tweets (10%-20%), signifying that the rule based models highly depend on their lexical based counterparts. We also observe that learning from just 20% of the available positive examples, the F1-measure for *ME+lex* model is 0.564, and for *WESPAD* model is 0.658. These F1 values are comparable to the F1 values that these models achieved in PHM2017 dataset, which also contains only 19% of the positive class in the training and test data (on average, across the different disease topics).

In summary, our results show that WESPAD is able to outperform the state-of-the-art baselines for both datasets and under variety of settings, and even outperforms a disease-specific classifier in the prominent FLU2013 benchmark dataset. This is striking, as