Hi there, Hi Mike, Here is Group 1’s presentation.

I’m Binkai Tan.

As is shown in this flow chart, we have tried 4 different ways to explore the project.

First of all, we tried the base-line method, in which we used doc2vec plus xgboost,

Then considering Inter-sentence relationship and context we used Elmo to transfer doc into vectors

Thirdly, we tried Glove with several kinds of classifiers to compare the result

Finally, we tried BERT.

I did the first two pipeline

--------------------------------

So here comes the base line method, In this first pipeline , we will show you the basic text processing method and basic hyperparameters.

And here is the flow.

Firstly, we cleaned the data, using lemmatize, lower the words and removed stop words and all of the punctuations. The tile and text is like this

And then EDA, we counted the words and We got the top 15 words in both fake news and real news, then comes the word cloud ! also both in fake news and real news . To be honest, they are very similar and can’t explain much.

Then Doc2Vec, we set the dimension of text is 300 and dimension of title is 50, here is the array.

Then Smote, it is used to balance data, and as is shown the number of each label is the same both in training data and test data.

Finally XGBoost, using grid search and cross validation to hypertune the model

The result is here f1 is 0.877, not bad.

For the previous model we did not take the context into account. so In the second pipeline, we searched the Internet and found elmo is a good way to convert the text into vectors. ELMo is a new type of deep contextualized word representation, which can model the complex features of words (such as syntax and semantics) and the changes of words in linguistic context (i.e., the modeling of polysemous words). Our word vector is a function of the internal state of the deep bi-directional language model (biLM), which is pre-trained in a large text corpus.

Due to the time limitation, Here shows the result:

Because running all of the samples will overflow the colab’s memory, so I tried 100 samples to display it symbolicly

Next give it to Yichen for the third pipeline here