Hi everyone, we are Group 1 from Hollis Lab, our team members are: Ji Haorui, Zhang Mengze and Tan Weihao. It’s a pleasure to present our final work and share with you our ideas about this project.

What we have done in general can be described as:

1. Combine title and text corpus together to retrieve more information from the dataset.

2. Concatenate vectors made from w2v, d2v and tfidf methods as final feature vectors (900d).

3. Using an Ensemble classifier (composed of five tuned estimators) to train and make predictions on the feature vectors we get.

4. Reach an over 0.94 f-1 score in test.

And here is the flow chart of our pipeline, which I’ll explain in no time.

We first preprocess the raw data by removing empty documents, removing punctuations and stopwords, finally stemming and lemmatizing using NLTK. In this way can we get a “cleaned” version of dataset used for training and testing.

Our second stage aims at getting informative feature embeddings. Just as the flow chart suggests, we split into two parts here, and each part has two concerns. First we would like to explore the relationship between titles and texts. How do we take fully advatange of information lies in the dataset especially for this particular type of FakeNews dataset? By simple concatenation of title and text corpus? Or by training them separately and taking their weighted average? Don’t worry, we shall see it soon. Second we use several embedding methods including w2v, d2v and tfidf to get our final 900-dimension feature vectors, whose detail will be explained later.

After the feature vectors are generated, our third stage is building our final classifiers and conduct hypertuning on it. We do this by training five different classifers including Logistic Regression, Random Forest, XGBoost, SVM and MLP separately and ensemble them using dynamic weights. After all those steps mentioned above, we get our final result of over 94% f1 score.

Next we’ll explain our pipeline in detail using our code.