Greetings everyone! Welcome to our project.

My name is Kien Nguyen, and I am the speaker of our group. You will find me in the photo on the right. And on the left is Son Nghiem, the other member of our two-person team.

As you already can see, the topic of our project is Natural Language Processing with Disaster Tweets. In this project we focus on predicting which Tweets are about real disasters and which ones are not. However, our research aim is more about understanding the usage as well as the performance of the language models we use to accomplish that goal.

Let us talk about the motivation behind this project. The rise of Twitter as one of the biggest social networks on the planet and the omnipresence of smartphones enable people to announce an emergency they are observing in real-time. Because of this, more agencies show interest in monitoring Twitter for quick real-time emergency responses, like issuing warnings in news channels or dispatching help when necessary.

You will see how that can be achieved in the right diagram. In the first layer the Tweets will be preprocessed, classified, and analyzed to gain key information about the disasters, which will then be utilized by the second layer to provide services to its subscribers. The scope of our project lies mostly in the red bounding box.

To accomplish the project goal, we will use a dataset from Kaggle which consists of approximately eleven thousand Tweets. Seventy percent of them are training data that was previously hand classified by Kaggle. The remaining unlabeled thirty percent test data is going to be classified by us. All of them share the same data format: the text of each Tweet is always given; its keyword and location are occasionally so.

Movin on to the models at our disposal. Son will use the TF-IDF, Word2Vec Embedders and the Logistic Regression Classifier to perform the given task, while I possess Bag-of-Words, Glove, BERT models and Multilayer Perceptron, Random Forest Classifiers in my toolbox. Our work orientates itself by the shown pipeline, which includes four phases: Data Analysis, Data Preprocessing, Applying Models, Evaluation and Conclusion.

As Evaluation Metric we choose F-Beta-Score - the weighted harmonic mean of precision and recall. This score weighs recall more than precision by a factor of Beta, which will be set smaller than 1, as precision is more important than recall in the previous diagram.

That ‘s it for our project. Thank you for listening. We welcome any questions and feedback.