第一页

Good afternoon everyone. We are group Pangolin. I'm Beichen Zhang from the School of Natural Resources. Yinglu & Bin. Today, we are going to present our final project: deep learning based nature disasters monitoring from twitter.

第二页

I will start with an introduction of the backgorund and define the primary learning problem and goal, following with a data description and some exploratory analysis. Yinglu will continue explaining our experiment setup for the models. Finally, Bin will discuss the final results and give a summary of the project.

第四页

Our project is focused on natural disaster. Especially with ongoing climate change, natural disasters are not far from us. Actually, we just experienced a severe flooding in 2019 caused milions dollars damage in Nebraska. And the Texas winter storm just happend in this March took at least 56 people's lives and lead to over 195 billion damage. Some other common disasters such as wildfire, drought, earthquake, and hurricane.

第五页

During these disasters, Twitter, being one of the largest social media platform, has been widely used to distribute information and disasater relief supplies. For example, left picture shows the Presidential Coomunicaiton Development and strategic planning office tweeted about the information of typhoon in 2012. It made us think about using Tweets as a primary inputting data set to identify information related or unrelated to natural disasters, which might be helpful to analyze the needs, impacts, damages. And help svae lives and provide aid.

第六页

However, it is very challenging to achieve a such goal. First, the data from Twitter are extremely huge. About 500 million tweets would be published every day. So, it is not realistic to manually identify those information. Second, understanding the language we speak that is called natural language is a very hard task for a computer. For example, the tweet in the picture mentioned a word ablaze that can be understood as burning fiercely, however, can also be interpreted as very bringhly colored or lighted, or even filled with anger or another strong emotion. Based on the sentiment of tweets, it was used to described the sky. Now, with the developing of deep learning algorithm, now computers are able to learn the correct meaning of the sentences.

第七页

Therefore, in this project, our learning problem is applying Tweets and deep learning to identify whether they would be related or unrelated to natural disasters, we definied it as a natural language processing problem. To be more specific, it is a binary text classification problem by deploying deep learning algorithms. And our primay goals are building and testing several RNN and BERT models, investigating the differences between pre-trained and self-trained word-embedding models. in the end, comparing and discussing results.

第九页

The data set was downloaded from Kaggle. Training set has 7613 tweets and test set has 3263 tweets. except for id, we have four features: keyword, location, tweets content, and target. Target is the label of whether the tweets are related or not to natural disasters.

第十页

Folowing, we did some exploratory data analysis. Left figure is the histogram of the data distritbuion in the training data set. Tweets are label as unrelated to natural disasters are slightly more than related. However, based on the size of the data set, we considered the data set was balanced. And the right figures are two histrograms for missing values in both test and training data set. Over one third of the location were missing while, only about 0.8% keyword were missing. Therefore, we dropped the feature of location and kept keyword. For the missing values in the keywords, we use no keyword to fill up.

第十一页

Besides, based on the word clouds for the keyword. We could see that keywords related to natural disasters can provide some addtional and useful information such as forest, fire, typhoon. And the keywords for tweets unrelated to natural disasters were mostly far from natural disasters.