Video transcript:

Hi, everyone! My name is Nancy. Today, I’ll present our project plan and initial progress.

The glut of information and the opaqueness of the sources make it difficult for people to distinguish between true and fake news. So the purpose of our research is to give people clues to differentiate between fake and real news. Our research question is what indicators distinguish real and fake news. Specifically, we want to know what common features appear in real and fake news respectively and how we can make a model for people to determine if an article is fake news or not.

We find a dataset containing a total of 44,043 news texts on Kaggle.All texts are collected from 2015 to 2018. These images show the details of the dataset.

Our plan can be divided into two parts. First, we will analyze the texts of fake and real news and find their features or differences. This part will analyze sentiment, word frequency, topic, and geographic location of texts through methods such as NLTK, Topic Modeling, TF-IDF, and LIWC. Second, we will build a model based on what we find for users to determine the truth of the news. And, we will use machine learning and greedy algorithms to achieve this goal.

We initially tried to find the sentiment of titles and contents, the common words, and tf-idf scores in fake and real news, respectively. Then, we decided to use word cloud, NLTK and tf-idf methods to get these results.

We firstly created two word clouds to show common words in real and fake news. This is our code to get word clouds.

Comparing these two-word clouds shows that the term “Trump'' appears in both real and fake news, which is reasonable because Donald Trump was president during 2017 and 2018. On the other hand, the names of presidents such as ‘Obama’ and ’Hillary’ are more frequently mentioned in fake news than in real news. Interestingly, ‘video’ is one of the most common words in fake news and we need to do more work to find out what video represents? In addition, real news has more common political words than fake news. For example, the terms ‘US’, ‘North Korea, ‘republican,’’and Russia’’ frequently appear in real news and are political-related words. It implies that the focus of real news is more centered than fake news, while fake news is more casual and its focus lies more on people.

Then, we applied the NLTK method to compare the emotion of content and titles through line charts, column charts and pie charts. This is our code for NLTK.

These line charts show the emotion of true news. The blue graph shows the emotional score of the text, and the orange one shows the score of the title. From the graphs, we can see that the fluctuation of emotion in content is larger than in the title, but the title graph has a more extensive range of emotional scores, which is from 0.4 to -0.8. The true news is mostly positive (in content and neutral in title) from Jan. in 2016 to September in 2017 as the line is above 0, but text in September. But emotion becomes neutral to slightly negative, and fluctuation is minor than before after December 2017 . The emotion of the title is primarily neutral but has become negative since September. In 2017, fluctuation at that time also becomes small.

For fake news. The blue and orange graph, which shows the emotional score of text and title respectively, has higher volatility from March in 2015 to Jan. in 2016. However, both fluctuations become smaller after that. Moreover, the emotion of the content is primarily neutral , and the content is primarily negative， which is below 0.

These pie charts show the emotional ratio of true news and separate them by years. For example, from the left graph, we can see that the positive emotion in content decreases while negative emotion increases from 2016 to 2017. The major emotion in the content of true news is still positive，but it decreases from 2016 to 2017. And, Title is mainly neutral but has a decreasing trend. Therefore, it decreases and negative emotion increases from 2016 to 2017.

For the fake news, the lower graph shows that neutral emotion in content is increasing and this trend is consistent with in title. Compared with pie charts of real news, fake news has more negative content and more neutral titles.

Finally, we calculate the TF-IDF score of fake and true news. But we had a session crash problem while running the code because of limited RAM volume. We tried many times to fix this problem, but unfortunately, we could not solve it before the class, so we decided to analyze a few news texts.

Here are lists of 20 most common of-idf words in fake and real news. We can find that fake news uses more adjectives and verbs such as ‘wish’, 'happy ',and ’new’. Real news uses more nouns such as ‘investigation’,’ republicans’, and ‘administration’. It may imply that real news is more official than fake news, but we still need more research to test this conclusion.

We raised four questions in our initial progress that we need to solve in the future. Like

That’s all we have right now. Thank you for watching! Have a good night!

Hi, everyone! My name is Nancy. Today, I’ll present our follow-up progress.

Our research question is What…..

We plan to find the differences firstly……. And then we will create a model to test ……..

Last time we did sentiment analysis……

And this week, we did sentiment analysis through a scatter plot…..

These two pictures show the title and content emotion of real news. We found that the fluctuation of emotion in content is larger than in the title. Like real news, fake news has a wider fluctuation of emotion in content than in the title. By comparing the scatter plots of true and false news, we found that the emotional distribution of true news is relatively uniform, but the content distribution of false news is relatively extreme. The false news content is mainly gathered in 1 and -1, 1 represents positive and -1 represents negative, while the news gathered in 0 is less than the actual news.

The following table shows the named entity recognition results of real and fake news. The first column is real news and the second column is fake news. The GPE (Geopolitical entity)of true news is much higher than false news. Real news also has a higher LOCATION than fake news does.

Depending on named entity recognition, topic classification analysis is carried out. Under the same classification standard, True news has fewer subjects than fake news. the true news only has two news sources: political news and world news. Real news also has more GPE, organizations, and persons. The news sources in fake news are more than in real news. there are government news, Middle-East news, US news, left-news, and politics. Fake news also has many GPE, organizations, and people, but its number of GPE is less than real news.

Last time, we could not run the tf-idf score for all news texts because we have limited random access memory. We decided to shuffle into 2000 texts for each fake and real news and these lists are the 20 most common of-idf words in fake and real news respectively. The words in both lists are pretty similar, but the frequency of words is different.

then, we did topic modeling. We set 10 topics to analyze news and visualize our topic distribution using the LDA method. From this visualization, we find that the biggest topic in both real and fake news in politics and social media. Both fake and real news includes many political words such as “trump”, “election”,” policy”, and ‘twitter’. However, real news has more diverse political words than fake news has.

Finally, We use the clustering method to find the content difference between fake and real news. We also create a word cloud to show the representative words for each cluster. These are word clouds for real news

And These are fake news’ word clouds

Depending on our analysis from this week and last week, we conclude five differences between true and fake news. First, we find that the focus of real news is more centered than fake news and real news uses diverse aspects to describe a single topic; it may imply that real news is more dedicated to creating reflect to reflect a wide range of perspectives across many different groups or organizations. Second, we find that first- and second-person pronouns are used for more unreliable or deceptive types of news. Third, Words used to exaggerate - subjective, superlative, and modal adverbs - are used by fake news, and more official

words are used in real news. Last but not least founding is that we are really surprised how similar fake and real news are. Although we applied many methods trying to get an explicit indicator differing in fake and real news, it seems impossible to analyze only from the contents of this news. We will continuously and more detailedly analyze our results next week.

Depending on our analysis from this week and last week, we conclude five differences between true and fake news. First, we find that the focus of real

news is more centered than fake news, and real news uses diverse aspects to describe a single topic, it may imply that real news is more dedicated to creating reflect to reflect a wide range of perspectives across many different groups or organizations. Second, we find that first- and second-person pronouns are used for more unreliable or deceptive types of news. Third, Words used to exaggerate - subjective, superlative, and modal adverbs - are used by fake news, and more official words are used in real news. Last but not least founding is that we are surprised how similar fake and real news are. Although we applied many methods to get a clear indicator differing in fake and real news, it seems impossible to analyze only from the contents of this news. We will continuously and more detailedly analyze our results next week.

That is all we have right now. Thank you!