IST736 Texting Mining

Yicun Deng

Prof. Bei Yu

**Topic: How to classify fake news**

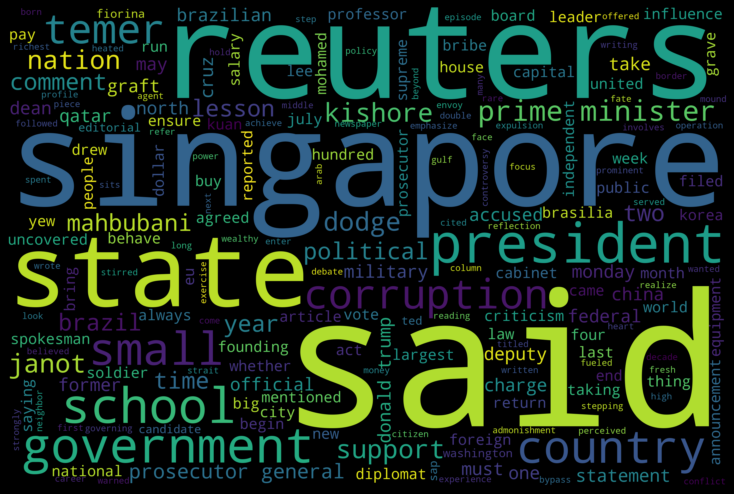
**Ethics Statement:** This study is done on an online dataset from Kaggle of true/fake news. We had not selected, processed, or modified any data before this study. As a result, the outcome of this study is only serving as a presentation for using different text mining methods such as data processing, vectorizing, and modeling to increase the accuracy of our model’s prediction. The outcome must not be used in any commercial setting. Also, the output model that predicts with the highest accuracy might not work the same in real life due to potentially overfitting and other issues.

**Data Cleaning:** Frist, we use the “panda” package from python to read in the truenews.csv and fakenews.csv. Then, we use the “panda” package to blend these two datasets. One thing that I want to draw your attention to is the importance of shuffling the data after merging the data. Otherwise, you are looking for an extremely low accuracy model since the training data is consisting only of true data and the testing data is consisting only of fake data.

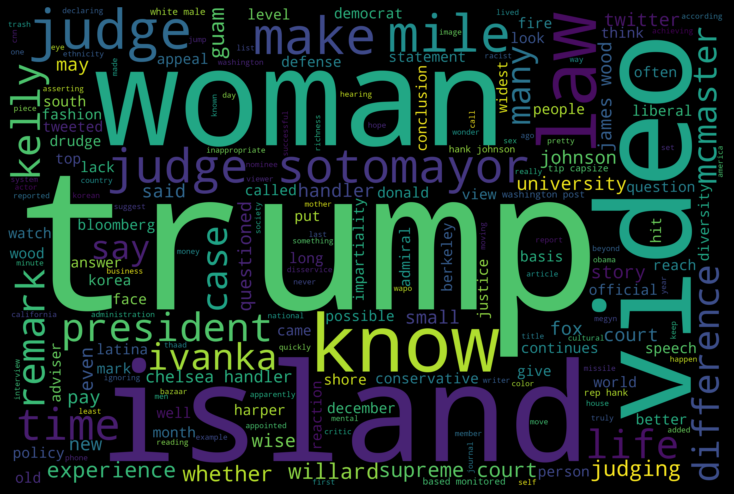
Then, we used the package of “texthero” to remove digits, punctuations, diacritics, brackets, and stopwords from the text for data processing purposes.

Also, we had used “WordNetLemmatizer” to lemmatize the data. We had also generated these word clouds:

**Top true news words:**



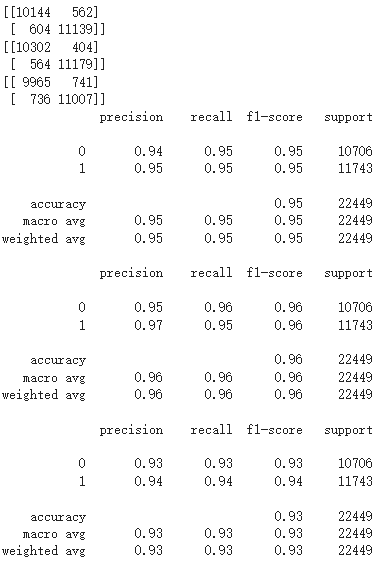
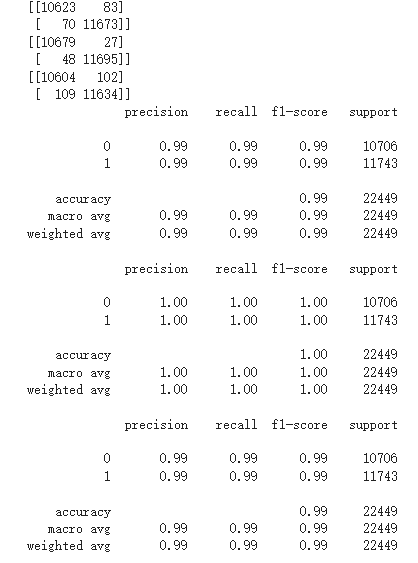
**Top fake news words:**



**Vectorizing:** We used count vectorizer, boolean vectorizer, tfidf vectorizer for the SVM, MNB, LTSM model

**Accuracy for SVM and MNB models:**

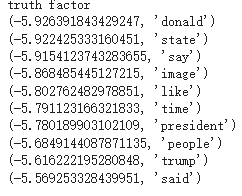
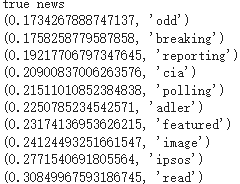
**SVM: MNB:**



Both the boolean vectorizer works the best for the SVM and MNB models. And the 100% accuracy of SVM is stunning. Though, an accuracy this high normally means over-fitting.

**Features:** Due to unknown reasons, my model is only giving me top features relating to true news. These are the top10 features of true news from SVM and MNB:

**SVM: MNB:**



**Error Analyzes:**

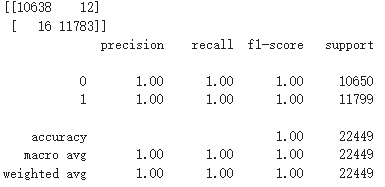
**SVM true as fake (partial since the full text is extremely long): “government grant taxpayer dollar beyond top done fallen police officer victim domestic terrorism obama always talking everyone equal well obviously equal indulgent since work million per family washington reuters justice department fast track sending million south carolina help family victim mass murder nine churchgoer historic black church charleston justice department spokesman said friday unspecified portion money allocated government national crime victim assistance formula grant program used provide service family victim shooting emmanuel ame church spokesman kevin lewis said via yahoo“**

**SVM fake as true (partial since the full text is extremely long): “sacramento calif guarantee california vote legalize recreational marijuana november political operative father four daniel conway already staked future conway left job chief staff sacramento celebrity mayor former phoenix sun nba basketball star kevin johnson help start marijuana investment company truth enterprise one hundred populous u state already pushing ahead plan”**

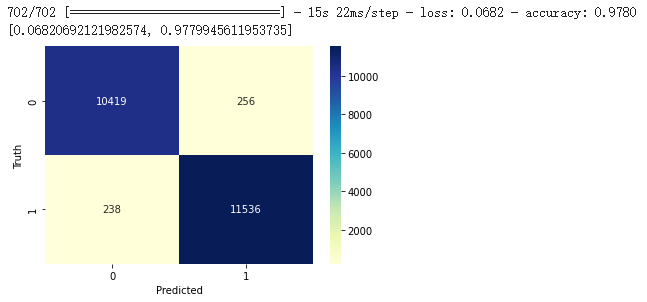
**MNB true as fake (partial since the full text is extremely long):”salt lake city reuters president donald trump said monday want u senator orrin hatch utah run election next year putting trump collision course republican rival mitt romney want run hatch seat hatch made noise retiring senate seat held since republican official say romney utah resident republican presidential nominee preparing run hatch seat next year trump clashed romney past said want hatch run election true fighter orrin say trump said event salt lake city utah capital hope continue serve state country senate long time come asked going seek election hatch told reporter see called trump endorsement certainly nice thing say whether would influence decision romney former governor massachusetts spends great deal time utah expecting run hatch seat state republican typically win romney made clear took dim view trump”**

**MNB fake as true (partial since the full text is extremely long):”washington reuters u senator john mccain recently diagnosed aggressive form brain cancer given hero welcome return capitol tuesday quickly seized opportunity blister party president partisan politics bruised scarred recent surgery flashing time characteristic self deprecating humor mccain spoke length senate floor delivering passionate rebuke fellow republican congress administration shown result republican president donald trump first six month office getting nothing done year old senator lamented mccain made dramatic return arizona home cast critical vote keep alive one trump top”**

**Accuracy for BERT and Bi-LSTM:**

**BERT: **

**Bi-LTSM:**



**Error Analysis:**

**BERT:**

**Trueasfake error(partial):”new york ifr jared kushner divested equity interest fifth avenue story office retail building manhattan famed shopping area according spokesperson kushner company kushner said would step ceo kushner company family owned real estate company begin divest substantial asset made senior white house advisor u president donald trump father law mr kushner divested equity interest fifth avenue role management operation property kushner company spokesperson said emailed statement ifr mr kushner ownership interest sold using third party appraisal fair market value family trust beneficiary spokesperson said neither ivanka trump kushner child beneficiary family trust spokesperson also confirmed asked ifr kushner company declined give detail sale price equity stake declined discus aspect outstanding debt property kushner bought property u 8bn highest price ever paid single office building sale united state time according kroll bond rating agency last valued well level last appraisal completed part debt restructuring valued building u 820m kroll said building vacant july valued property u 1m kushner partner fifth took u 2bn senior debt buy property later packaged sold three cmb deal clear divestiture mean debt property senior debt restructured extended february part restructuring kushner brought vornado realty partner call email vornado immediately returned”**

**Fakeastrue error(partial):“Fakeastrue error:white house wednesday disclosed group former lobbyist working president donald trump administration issued ethic waiver following request u government ethic agency list least waiver include white house adviser kellyanne conway trump chief staff reince priebus according chart issued white house website bit ly 1lpkwx8 conway permitted participate communication meeting involving former client political advocacy trade non profit organization priebus former republican national committee chairman allowed communication meeting rnc document say shortly taking office january trump signed executive order barring lobbyist joined administration working issue related prior work administration power grant waiver particular hire exempting restriction also list michael catanzaro special assistant president former oil gas lobbyist cleared weigh energy policy daniel epstein associate counsel president may provide legal advice“**

**Bi-LTSM: Unfortunately, Bi-LTSM seems to work differently from BERT, SVM, and MNB. Thus I couldn’t access its error…**

**Performance debriefs:** Bert works the best for this given dataset, making only 26 errors out of 20,000 data points. Followed by SVM, making only 75 errors out of 20,000 data points. Then, Bi-LTSM is doing a good job of making 500 errors out of 20,000 data points. The least successful model is MNB, making 968 errors out of 20,000 data points. Though, the performances made me wonder if this dataset I am using is somehow modified or improved for modeling? It is kind of unusual to see all models past 95% accuracy.

**Something I had learned:** Shuffling your data after merging is really important. Previously, I got really low accuracy scores because I am using true-news data to train and bad-news data to test (60%..) This result even lead me to think of the abbreviation’s impact on the modeling so I had tried to either complete the abbreviation or deleting. Completing the abbreviation is a failure, despite that papers are providing theoretical solutions, I don’t really know how to do this in Python (Park & Byrd, 2001). Removing all the abbreviations seems only improve the models by littles. However, finding out that I need to shuffle the data saved in my project, my performance immediately jump to above 95% for all models.

**Over-fitting test:** I used text from the Onion News website to test BERT’s performance on new data. Here are the results: 100% accuracy for identifying news from Onion News to be fake:

**news1**:'NEW YORK—In an effort to keep customers apprised of what to expect on their trip, a United Airlines email with the subject line “Information About Your Upcoming Flight” reportedly informed passengers Friday that airplane will go in sky. “You go up, up, up!” read the informational email in part, confirming that sky was up where clouds are. “Sit down in seat on plane and you fly way high in air where are bird and balloon. Whoosh through big blue sky! Yippee!” At press time, the email assured concerned passengers that airplane not go so high that it crash into sun.'

**news2:**'PHOENIX—Describing the 15-minute incident as “a nightmare come to life,” local office workers at Martin Realty were reportedly terrorized Friday by an unhinged ex-employee dropping in to say hi. “We always knew he was a bit off by the way he’d say ‘good morning’ and ask about how our weekend was, but we never thought it’d escalate to anything like this,” said administrative assistant Andrea Foder, who was one of several employees who were forced to barricade themselves inside the bathroom as former associate Henry Bassett stalked his way through the office suite, openly carrying a large box of pastries. “He was definitely disgruntled after getting laid off, saying stuff like, ‘I’m really going to miss you guys,’ and, ‘Let’s keep in touch,’ but at the time I chalked it up to the heat of the moment. I guess I should have taken the threat more seriously.” At press time, a rattled Foder stated, “That could have been me,” after learning that Bassett had taken two of her coworkers to lunch.'

Unfortunately, I didn’t find a way to run these tests with SVM, MNB, or Bi-LTSM. But BERT is really doing a good job here.

**Conclusion:** BERT is performing the best–getting the highest accuracy, predicting right for the new fake news found on OnionNews’ website. I believe that BERT is working the best for more complex jobs based on the transformer inner structure it uses. But for a simpler dataset like predicting if someone lives in Florida using her/his address information, SVM should be a quick and accurate choice.

’

Reference List:

Park, Y., & Byrd, R.J. (2001). Hybrid Text Mining for Finding Abbreviations and their Definitions. *EMNLP*.