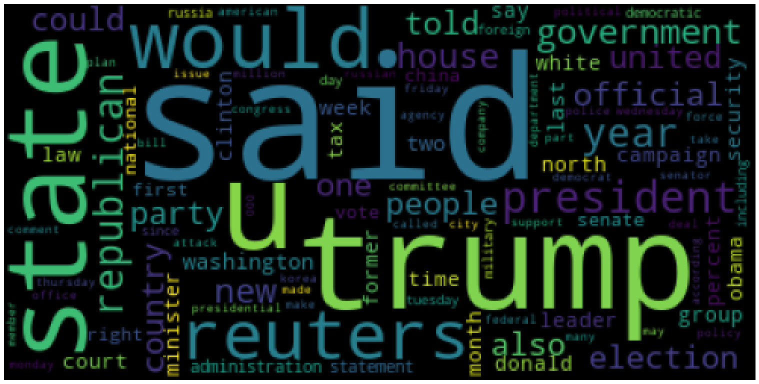
Sirut Buasai
CS 525
Prof. Xiaozhong Liu
Assignment 1

**Task 1**

1. Most common 100 words in real news, fake news, and collection of both
    a. Real news

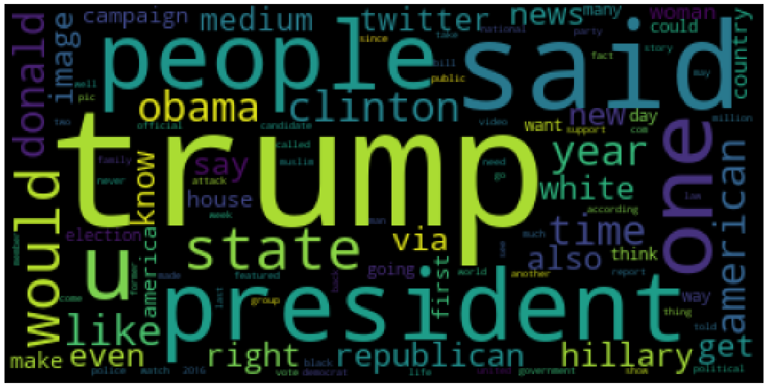
| Word | Freq | Word | Freq | Word | Freq | Word | Freq |
|---|---|---|---|---|---|---|---|
| said | 99062 | leader | 10575 | day | 8095 | made | 6422 |
| trump | 54732 | security | 10466 | russia | 8064 | city | 6389 |
| u | 47110 | court | 10460 | presidential | 8039 | department | 6364 |
| state | 37677 | donald | 10456 | wednesday | 8014 | issue | 6342 |
| would | 31605 | percent | 10012 | democrat | 7984 | 000 | 6246 |
| reuters | 28976 | say | 9949 | may | 7842 | company | 6234 |
| president | 28728 | north | 9912 | political | 7723 | make | 6188 |
| republican | 23007 | time | 9699 | support | 7675 | part | 6179 |
| year | 22622 | law | 9665 | thursday | 7664 | comment | 6143 |
| government | 19992 | tax | 9653 | million | 7661 | according | 6142 |
| house | 17030 | white | 9618 | bill | 7618 | police | 6088 |
| new | 16917 | clinton | 9570 | policy | 7589 | take | 6086 |
| also | 15954 | minister | 9569 | american | 7536 | attack | 6041 |
| united | 15590 | obama | 9406 | plan | 7407 | | |
| people | 15356 | month | 9275 | member | 7363 | | |
| party | 15294 | senate | 9253 | friday | 7332 | | |
| election | 14759 | right | 9229 | korea | 7299 | | |
| official | 14620 | vote | 9105 | monday | 7101 | | |
| told | 14245 | china | 8866 | force | 7095 | | |
| country | 14161 | first | 8810 | office | 6968 | | |
| one | 13750 | national | 8582 | committee | 6889 | | |
| could | 13711 | statement | 8528 | deal | 6884 | | |
| washington | 12988 | administration | 8427 | called | 6804 | | |
| last | 12776 | democratic | 8387 | many | 6724 | | |
| two | 12711 | since | 8334 | agency | 6577 | | |
| campaign | 11155 | foreign | 8270 | congress | 6503 | | |
| group | 11129 | tuesday | 8268 | senator | 6502 | | |
| week | 10658 | military | 8171 | federal | 6457 | | |
| former | 10603 | including | 8123 | russian | 6456 | | |




b. Fake news

| Word | Freq | Word | Freq | Word | Freq | Word | Freq |
|---|---|---|---|---|---|---|---|
| trump | 79519 | twitter | 11722 | show | 8378 | attack | 6636 |
| said | 33763 | campaign | 11640 | black | 8340 | man | 6620 |
| president | 28310 | make | 11639 | featured | 8261 | support | 6532 |
| people | 26657 | woman | 11552 | last | 8258 | another | 6489 |
| one | 25389 | country | 11449 | group | 8209 | member | 6448 |
| u | 24545 | house | 11292 | according | 8076 | called | 6431 |
| state | 23658 | america | 11254 | united | 8011 | family | 6368 |
| would | 23562 | first | 10612 | take | 7952 | since | 6359 |
| clinton | 19826 | election | 10302 | see | 7923 | never | 6326 |
| time | 19214 | could | 10246 | report | 7888 | pic | 6322 |
| year | 19073 | day | 10153 | come | 7820 | candidate | 6266 |
| obama | 18797 | many | 9945 | fact | 7704 | 2016 | 6265 |
| like | 18649 | think | 9916 | may | 7700 | muslim | 6262 |
| american | 18120 | want | 9886 | political | 7654 | | |
| donald | 17681 | going | 9808 | life | 7464 | | |
| republican | 16726 | way | 9768 | world | 7463 | | |
| say | 15782 | government | 9704 | vote | 7450 | | |
| also | 15403 | law | 9381 | national | 7345 | | |
| right | 14860 | thing | 9183 | former | 7303 | | |
| news | 14629 | video | 9169 | democrat | 7248 | | |
| new | 14394 | told | 9122 | need | 7227 | | |
| image | 14319 | police | 9120 | million | 7203 | | |
| hillary | 14127 | made | 9119 | much | 7135 | | |
| even | 14012 | two | 9116 | story | 7021 | | |
| white | 13566 | back | 9039 | bill | 6826 | | |
| via | 12776 | go | 8721 | public | 6779 | | |

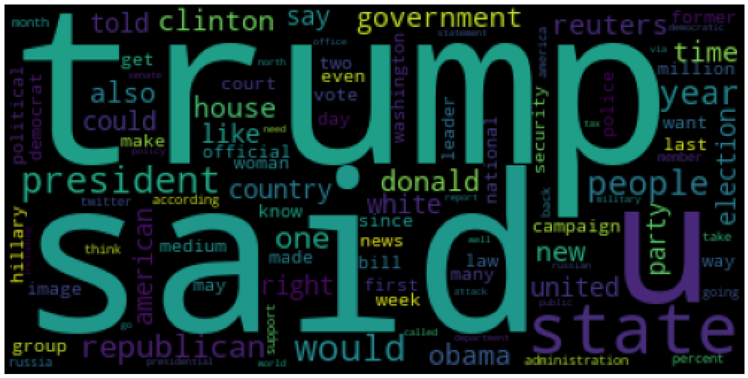
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| get | 12368 | well | 8619 | week | 6678 | | |
| know | 12055 | party | 8548 | watch | 6656 | | |
| medium | 11801 | com | 8541 | official | 6653 | | |




c. Collection of real and fake

| Word | Freq | Word | Freq | Word | Freq | Word | Freq |
|---|---|---|---|---|---|---|---|
| trump | 134251 | white | 23184 | police | 15208 | senate | 12930 |
| said | 132825 | campaign | 22795 | leader | 14940 | report | 12830 |
| u | 71655 | two | 21827 | image | 14923 | well | 12742 |
| state | 61335 | official | 21273 | million | 14864 | attack | 12677 |
| president | 57038 | last | 21034 | since | 14693 | including | 12626 |
| would | 55167 | news | 20661 | know | 14565 | north | 12583 |
| people | 42013 | first | 19422 | way | 14467 | world | 12301 |
| year | 41695 | group | 19338 | bill | 14444 | public | 12226 |
| republican | 39733 | law | 19046 | percent | 14429 | go | 12190 |
| one | 39139 | washington | 18729 | back | 14338 | department | 12151 |
| also | 31357 | day | 18248 | month | 14313 | need | 12119 |
| new | 31311 | even | 17943 | administration | 14299 | russian | 11984 |
| government | 29696 | former | 17906 | twitter | 14253 | military | 11949 |
| reuters | 29425 | make | 17827 | according | 14218 | | |
| clinton | 29396 | week | 17336 | support | 14207 | | |
| time | 28913 | hillary | 16870 | going | 14192 | | |
| house | 28322 | get | 16793 | think | 14165 | | |
| obama | 28203 | many | 16669 | take | 14038 | | |
| donald | 28137 | vote | 16555 | russia | 14034 | | |
| say | 25731 | security | 16349 | member | 13811 | | |
| american | 25656 | medium | 16319 | america | 13811 | | |
| country | 25610 | court | 16253 | presidential | 13783 | | |
| election | 25061 | national | 15927 | statement | 13602 | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| right | 24089 | want | 15658 | tax | 13524 | | |
| could | 23957 | may | 15542 | democratic | 13288 | | |
| party | 23842 | made | 15541 | via | 13285 | | |
| united | 23601 | political | 15377 | called | 13235 | | |
| like | 23448 | woman | 15267 | policy | 13050 | | |
| told | 23367 | democrat | 15232 | office | 12999 | | |



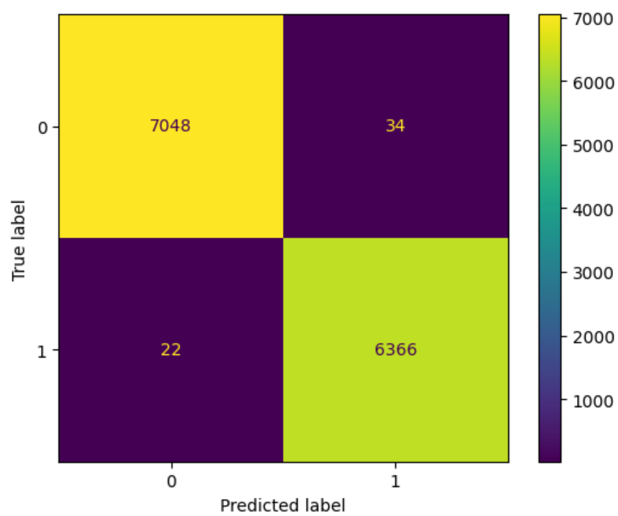

2. Difference between real news and fake news
    a. Given the top 100 most common words, both real and fake news seem to be using similar set of words. This is reasonable given that fake news are attempting to mimic the real content but with a twist of their own, therefore we can expect both real and fake news to report on the same topic, using the same set of words. However, fake news differs from the real ones in their word frequency distribution. The top 5 most common words from real news are *"said", "trump", "u", "state", "would"* while the top 5 from fake news are *"trump", "said", "president", "people", "one"*. Real news tend to utilize more verbs and focus their reports on actions. Fake news, on the other hand, tend to focus on more striking nouns. In fact, the word *"trump"* appears 1.5 times as much in fake news as real news. Given these observations, there seems to be a differing word frequencies between real and fake news, thus, a strong feature set would be certain nouns that more frequently than others in certain document.

**Task 2**

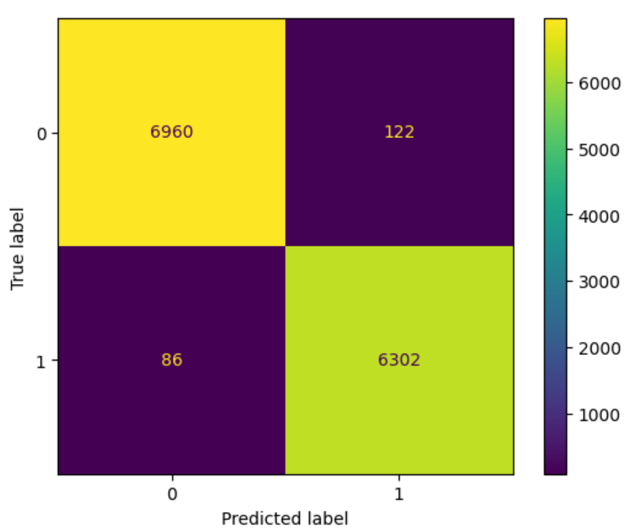
1. Algorithm Performance

| ML Model | Feature | Precision | Recall | Accuracy |
|---|---|---|---|---|
| Logistic Regression | TF-IDF | 0.981 | 0.987 | 0.985 |
| Multinomial Naive Bayes | TF-IDF | 0.935 | 0.933 | 0.938 |
| Logistic Regression | Bag of Words | 0.995 | 0.997 | 0.996 |
| Multinomial Naive Bayes | Bag of Words | 0.945 | 0.954 | 0.952 |

a. Rank 1: Logistic Regression on Bag of Words Feature Set




b. Rank 2: Logistic Regression on TF-IDF Feature Set



2. Error Analysis
    a. In Task 2, the top 2 best performing models are both logistic regression models on bag of words feature and TF-IDF feature. Both of these models yielded high precision, recall, and accuracy score. According to the two confusion matrices, the models were able to predict the labels correctly in both real and fake categories. This may be attributed to the fact that logistic regression models are extremely good at finding the decision boundary for linearly separable dataset. As seen from task 1, the word distribution between real and fake news seems to be separating by the frequency of certain words. Moreover, we can attribute this high performance metrics to the extremely clean data of real and fake news from Kaggle. Since the dataset had relatively small number of NaN values and that the string formatting of the tabular data cells were easy to parse through, data processing tasks were much easier than expected.

       Furthermore, it is worth mentioning that multinomial naive bayes models also performed extremely well on this data set, although not as well as logistic regression. This is again attributed to the clean dataset with elementary data preprocessing. One reason that logistic regression performed relatively better than multinomial naive bayes model may be attributed to the linearly separable nature of the dataset. As such, the logistic regression were able to find a better decision boundary. However, we cannot overlook the fact that logistic regression model may be overfitting the dataset given the limited amount of data that we have.

**Task 3**

1. Algorithm Performance

| ML Model | Feature | Filter | Precision | Recall | Accuracy |
|---|---|---|---|---|---|
| Logistic Regression | TF-IDF | Nouns | 0.982 | 0.979 | 0.981 |
| Logistic Regression | TF-IDF | Verbs | 0.932 | 0.941 | 0.939 |
| Logistic Regression | TF-IDF | Nouns, Verbs | 0.980 | 0.986 | 0.984 |
| Multinomial Naive Bayes | TF-IDF | Nouns | 0.921 | 0.912 | 0.921 |
| Multinomial Naive Bayes | TF-IDF | Verbs | 0.918 | 0.922 | 0.924 |
| Multinomial Naive Bayes | TF-IDF | Nouns, Verbs | 0.926 | 0.923 | 0.929 |
| Logistic Regression | Bag of Words | Nouns | 0.994 | 0.990 | 0.992 |
| Logistic Regression | Bag of Words | Verbs | 0.949 | 0.940 | 0.947 |
| Logistic Regression | Bag of Words | Nouns, Verbs | 0.995 | 0.993 | 0.995 |
| Multinomial Naive Bayes | Bag of Words | Nouns | 0.930 | 0.935 | 0.935 |
| Multinomial Naive Bayes | Bag of Words | Verbs | 0.921 | 0.954 | 0.939 |
| Multinomial Naive Bayes | Bag of Words | Nouns, Verbs | 0.938 | 0.947 | 0.945 |

2. Comparison with Task 2
   a. Compared to the machine learning models from Task 2, POS tagging seems to have an insignificant effect on both logistic regression and multinomial naive bayes models and both bag of words and TF-IDF features. This may be attributed to the fact that even before POS tagging and filtering, the data were already linearly separable given the abundance of nouns and verbs in both real and fake data with fake data having abnormal frequency of nouns. Thus, when filtering the data using only nouns, verbs, and nouns combined with verbs, the model showed no improvement due to the minimal change in the dataset after the filter.

      In this task, the top 2 best performing models are logistic regression on bag of words with filtering on nouns and nouns + verbs. This further cemented the assumption that the decision boundary between fake and real news may lie in the differing frequencies of nouns between the two classes.

**Task 4**

One interesting idea that inspired me from reading the papers is the usage of contextual data that can be extracted along with the text such as new sources credibility and contextual online environment. I believe that contextual knowledge is extremely relevant when it comes to determining the reliability news content. This is already the norm in the academia world given that researchers build up their reputation based on their peer reviews and research reputation. The same can be applied to the general media where the environment data such as social media platform or the information source such as authors can be included as a feature for machine to learn the relevance of those features.