Fake News Detection

INLS 613 Text Mining

Jessica Qiu Sirui Li Sarah Ganci

Introduction & Motivation

- Topic: Detect whether a piece of news is fake or not
- Fake news
  - Verifiably false
  - Intentional misleading
- Motivation
  - Mistrust in news sources, prevalence of fake news, concerning consequences of spreading misinformation, and the vast quantity of information shared on the internet
  - A need to produce an automated detection of fake news and deceptive content
- Goal:
  - Classification model
  - Compare effect of title and text on classification task

In Washington Pizzeria Attack, Fake News Brought Real Guns

Overview of Related Works

- **Shenoy, G., Dsouza, E., & Kübler, S. (2017). Performing Stance Detection on Twitter Data using Computational Linguistics Techniques. Retrieved from https://arxiv.org/ftp/arxiv/papers/1703/1703.02019.pdf**
    - Building a stance detection for tweets, which is to detect whether a certain tweet is in favor or against a certain particular target
    - Bag-of-words model
    - Involving sentiment scores to optimize the features
- **Ljungberg, B. F. (2017). Dimensionality reduction for bag-of-words models: PCA vs LSA. Retrieved from http://cs229.stanford.edu/proj2017/final-reports/5163902.pdf**
    - Bag-of-Words Model: Curse of Dimensionality & Computational Inefficiency
    - Dimension reduction technique
        - PCA(Principal Component Analysis): projects a set of points onto a smaller dimensional affine subspace of that represents most proportion of the variance.
        - LSA(Latent Semantic Analysis): naively apply SVD(singular value decomposition) to reduce the dimension of the feature space

Dataset

- Fake_Or_Real_News gathered by George McIntire with 6413 news articles in 2017
    - Original dataset: DocumentID+ Title+ Text+ Label of either fake or real
    - Around evenly split fake and real news
    - Fake news: Kaggle's fake news dataset
    - Real news: news articles from All Sides. News are published by various media such as WSJ and Bloomberg between 2015 and 2016

```
df['label'].value_counts()
```

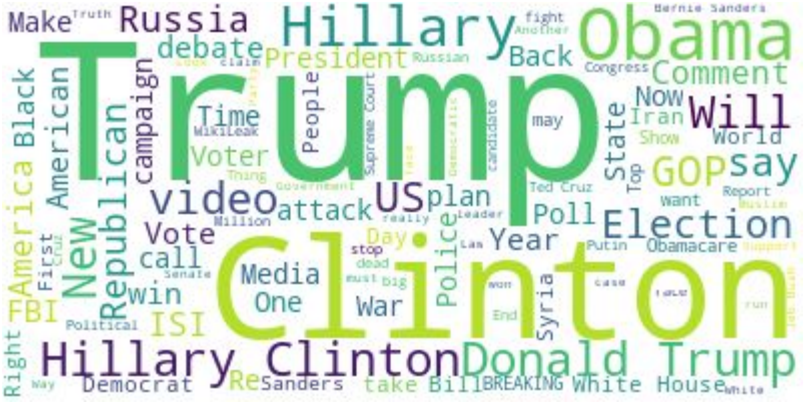
```
REAL    3171
FAKE    3164
Name: label, dtype: int64
```

```
df.head()
```

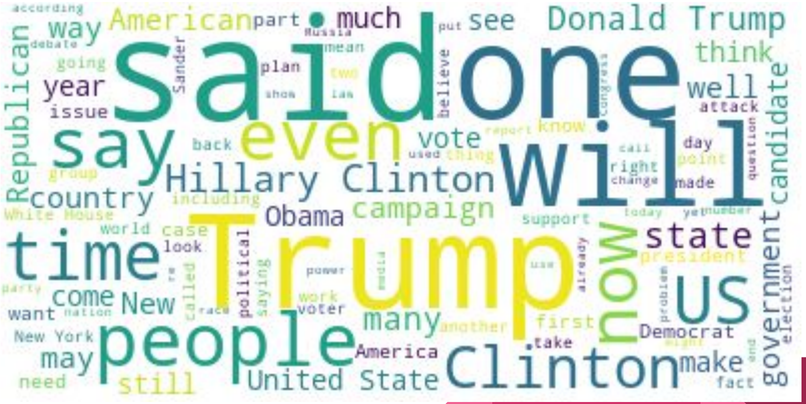
	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL

Exploration: Title and Text Word Cloud

Title Word Cloud



Text Word Cloud



Overview of Experimental Methodology

Train Test Split	Preprocessing	Feature Extraction	Model Fitting	Evaluation
Split data with 80% train and 20% test	Downcasing	F1: TFIDF of Title	Naive Bayes	Accuracy
	Removal of stopwords	F2: TFIDF of Text	Random Forest	Recall
	Converting labels	F3: TFIDF of Concatenated Text and Title	Support Vector Machine	Precision
	Modifying titles	F4: Combined F1 and F2 Vectors	Logistic Regression	

Feature Extraction

All feature spaces built based on python sklearn TFIDF Vectorizer

Four Different Feature Sets:

F1: TFIDF of Title

Fit on training data titles

50 features

F2: TFIDF of Text

Fit on training data text

50 features

F3: TFIDF of Concatenated Title and Text

Fit on concatenated prefixed title and text training data

600 features

F4: TFIDF Combined Vectors

Combined F1 and F2 Feature Space

100 features

Top 50 TFIDF of Title and Text

Title

world', 'debate', 'isis', 'iran', 'house', 'hillary', 'gop', 'fbi','emails', 'email', 'election', 'donald', 'deal', 'won', 'cruz', 'court', 'comment', 'clinton', 'campaign', 'bush', 'black', 'big', 'bernie', 'america', 'just', 'media', 'new', 'news', 'wikileaks', 'white', 'war', 'vote', 'video', 'trump', 'syria', 'state', 'says', 'sanders', 'russia', 'rubio', 'right', 'republicans', 'republican', 'presidential', 'president', 'police', 'people', 'party', 'obama', '2016'

Text

'world', 'war', 'people', 'think', 'just', 'year', 'like', 'don', 'new', 'america', 'country', 'say', 'way', 'time', 'government','obama', 'did', 'news', 'american', 'states', 'president', 'united', 'said', 'know', 'told', 'right', 'state', 'percent', 'white','political', 'going', 'make', 'years', 'clinton', 'campaign', 'democratic', 'donald', 'election', 'party', 'hillary', 'house', 'media', 'national', 'presidential', 'republican', 'sanders', 'support', 'trump', 'voters', '2016'

Overlap

'2016', 'america', 'campaign', 'clinton', 'donald', 'election', 'hillary', 'house', 'just', 'media', 'new', 'news', 'obama', 'party', 'people', 'president', 'presidential', 'republican', 'right', 'sanders', 'state', 'trump', 'war', 'white', 'world'

F3 Features

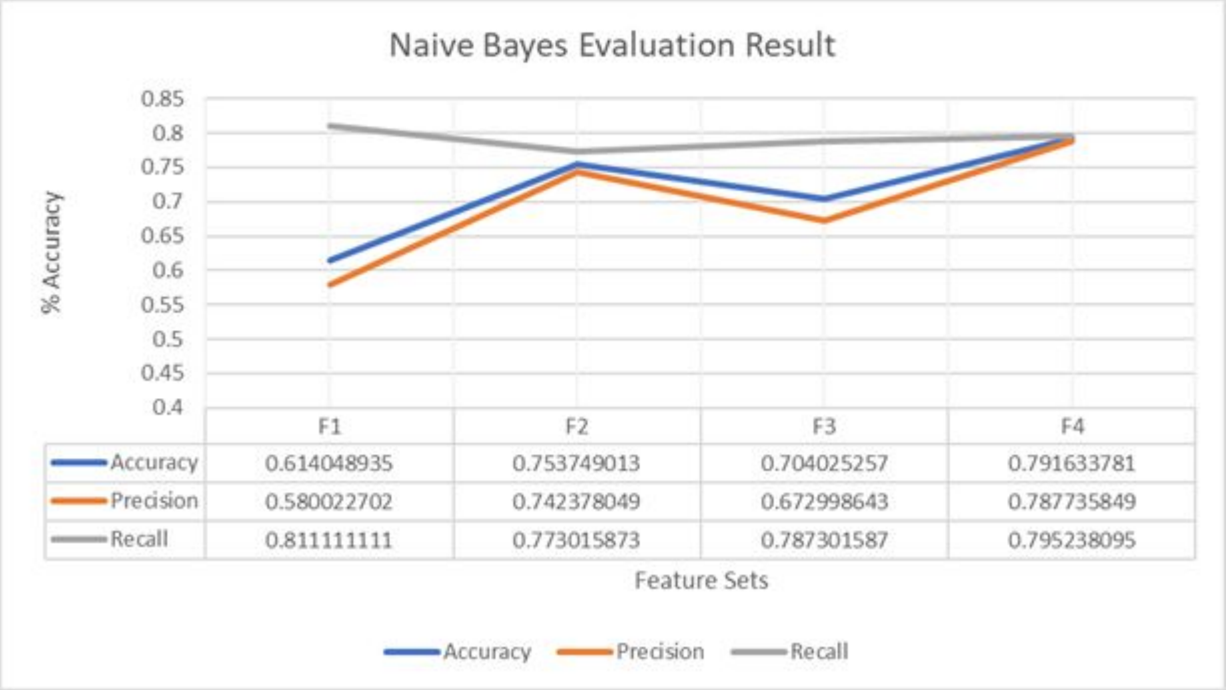
world, says, war, health, people, isn, response, global, human, countries, happened, look, talk, problem, million, public, death, doesn, 000, 10, think, bad, ve, international, come, just, year, kind, person, means, want, number, race, fear, like, vice, things, don, better, today, decades, problems, new, similar, work, good, need, america, care, simply, getting, control, country, course, say, months, way, used, help, likely, really, americans, times, crisis, foundation, rules, staff, title_the, time, killed, happen, government, wasn, day, 2014, entire, team, pay, shows, obama, did, news, american, chief, needs, states, wrote, director, line, 20, sure, face, ll, comes, administration, early, history, president, little, big, past, days, use, far, united, said, know, told, right, state, western, nuclear, book, experience, leadership, 25, field, al, choice, results, ways, member, syria, 50, victory, senior, expected, efforts, play, sent, thousands, speaking, leading, bring, turned, result, millions, position, talking, lives, tell, half, men, black, personal, effort, began, released, small, committee, job, private, idea, nearly, fight, future, large, attack, reason, led, known, close, title_of, possible, reported, military, money, different, set, question, lot, major, percent, title_in, ago, second, took, best, didn, issue, general, case, making, place, high, real, clear, policy, point, candidate, does, white, according, political, going, make, years, attacks, air, federal, gave, ahead, george, feel, army, announced, foreign, florida, area, focus, follow, following, force, final, article, friday, fighting, anti, fox, free, financial, asked, freedom, forces, gold, agreement, held, 2015, 2013, 2012, hard, having, head, 2008, 15, 12, hampshire, hillary, 11, hold, home, 100, hope, hours, house, hand, 2016, agency, act, address, added, given, actually, actions, gop, got, action, gov, 30, governor, great, ground, group, groups, gun, access, able, attention, company, fbi, immigrants, climate, deal, class, claims, debate, debt, claim, civil, decision, city, citizens, defense, christie, delegates, democratic, democrats, china, department, despite, clinton, data, change, current, conference, congress, congressional, conservative, conservatives, common, continue, comments, convention, comment, coming, comey, com, college, cnn, court, crime, criminal, cruz, children, chance, away, enforcement, especially, business, establishment, bush, billion, europe, event, evidence, example, executive, bernie, believe, community, based, barack, facebook, fact, failed, family, california, energy, chairman, end, doing, center, donald, cause, earlier, carolina, candidates, campaign, east, came, economic, economy, calling, called, elected, election, elections, email, emails, david, young, immigration, start, short, sign, single, social, source, south, speaker, special, speech, spent, stage, stand, started, sex, statement, stop, story, strategy, street, strong, sunday, support, supporters, supreme, syrian, taken, share, service, tax, russian, reporters, reports, republican, republicans, research, rights, role, romney, rubio, run, running, russia, ryan, sense, sanders, saturday, saudi, saying, school, secretary, security, seen, self, sen, senate, senator, taking, ted, religious, weeks, voters, votes, voting, wall, wanted, wants, washington, watch, water, weapons, wednesday, week, went, violence, west, wikileaks, win, winning, woman, women, won, words, worked, workers, working, wrong, vote, view, term, trade, terrorism, terrorist, texas, thing, thought, threat, thursday, title_, title_for, title_on, title_to, title_trump, tried, video, true, trump, truth, try, trying, tuesday, turn, twitter, understand, union, university, using, report, record, important, love, legal, let, level, liberal, life, list, live, local, long, york, looking, lost, majority, leave, makes, man, march, market, marriage, matter, mean, media, meeting, members, message, middle, left, leaders, minister, israel, including, information, instead, intelligence, interview, investigation, involved, iowa, iran, iraq, isis, islamic, issues, leader, jeb, jobs, john, justice, kasich, key, late, later, latest, law, laws, lead, mind, moment, recently, presidency, plan, plans, podesta, points, police, policies, politics, poll, polls, post, potential, power, presidential, party, press, primary, probably, process, program, putin, questions, rally, read, reality, received, recent, paul, parties, monday, north, month, morning, movement, mr, nation, national, nations, near, night, nomination, nominee, non, november, paid, numbers, october, office, officers, official, officials, ohio, oil, old, open, order, outside, longer

Supervised Learning Models

- Perform ML Models on 4 feature sets and discuss which feature sets with which kind of the models give us the best result
    - Major Evaluation Metrics: Test Set Prediction Accuracy
    - Other metrics as a reference: Precision and Recall
- Models Included
    - Naive Bayes
    - Logistic Regression
    - Random Forest
    - Support Vector Machine

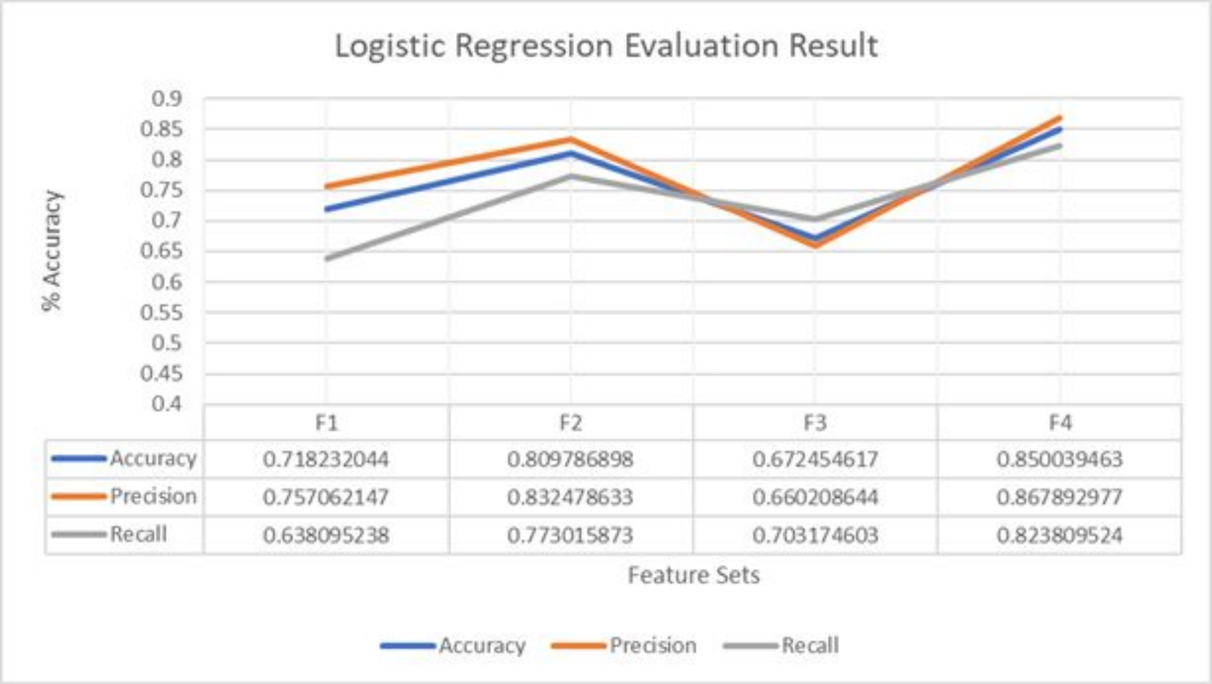
Naive Bayes

- Best Feature Set:
  - F4 (combined TFIDF title and text vectors) with 79.16% test accuracy
- Worst Feature Set:
  - F1 (title TFIDF) with 61.40% accuracy



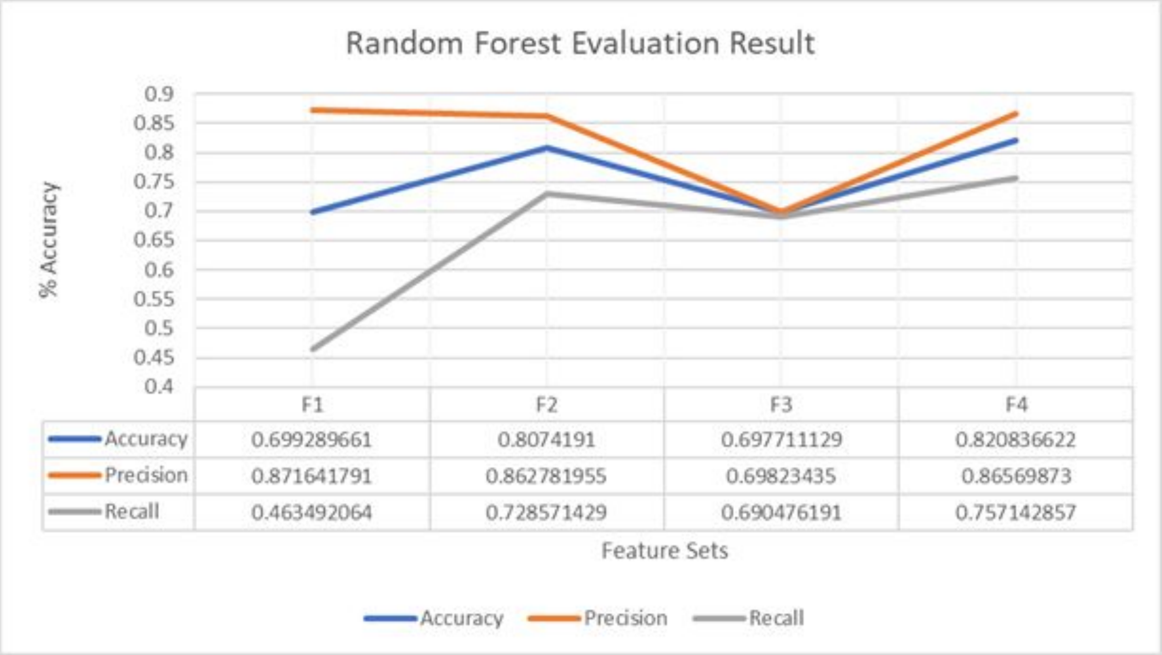
Logistic Regression

- Best Feature Set
  - F4 (Combined Text and Title vectors) with 85.00% test accuracy
- Worst Feature Set
  - F3 (Concatenated Text and Title) with 67.25% test accuracy



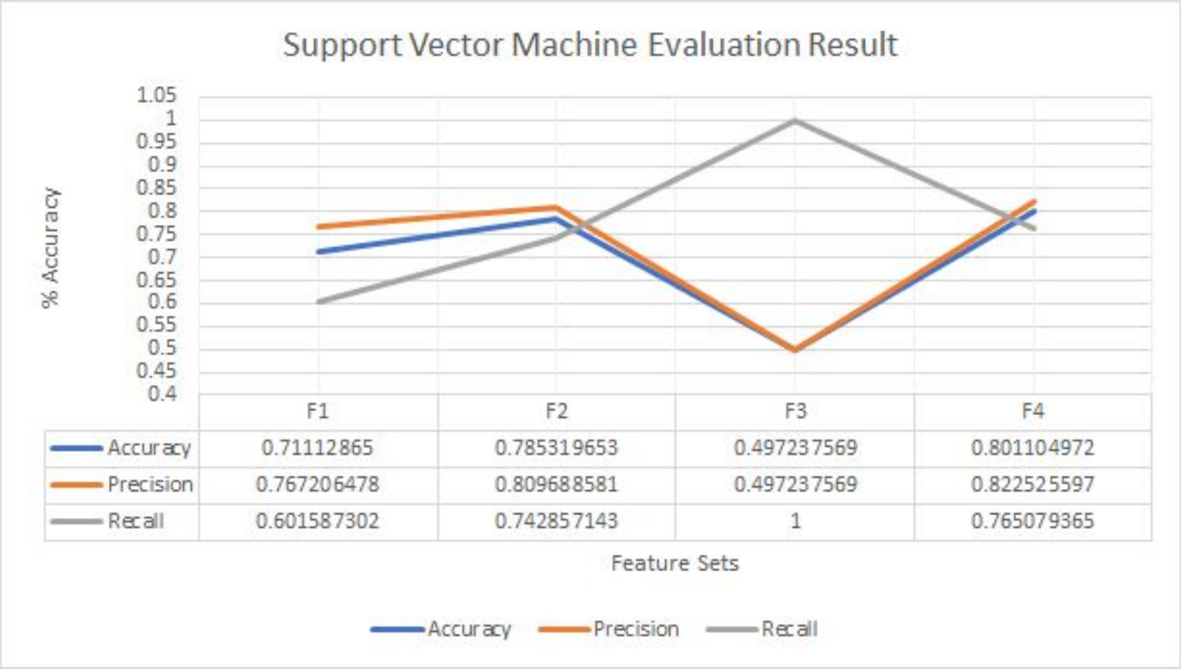
Random Forest

- Best Feature Set:
  - F4 (combined title and text vectors) with accuracy of 82.08%
- Worst Feature Set:
  - F3(concatenated title and text) with accuracy of 69.77%



Support Vector Machine

- Best Feature Set
  - F4 (Title TFIDF) with 80.11% prediction accuracy
- Worst Feature Set
  - F3 (Concatenated Title and Text with accuracy of 49.72%
- Why F3 performs bad?
  - Probably Curse of Dimensionality



Discussion

- Average rating of feature set for each model:
  - F1: 3.25
  - F2: 2
  - F3: 3.75
  - F4: 1
- Best Model and Feature Combination: Logistic Regression with F4
- For F3 with high dimensionality, Naive Bayes works the best among all models

Model	Feature Set	Accuracy	Precision	Recall
LR	F4	**0.8500394633**	**0.8678929766**	**0.8238095238**
RF	F4	0.8208366219	0.8656987296	0.7571428571
LR	F2	0.8097868982	0.8324786325	0.773015873
RF	F2	0.8074191002	0.8627819549	0.7285714286

Conclusion

- Title Matters!
  - Models based on text features outperformed models based on title features
  - Models performed best on feature sets containing elements from the text and title
- Limitations
  - F3 Feature set had to be large to avoid losing title features, and thus performed worse
  - Explored only unigrams
  - Limited scope of data set (collected articles published during an election year)
- Future Improvements
  - Explore other feature categories for both title and text (sentiment scores, part of speech tagging)
  - Reduce dimensionality using PCA(Principal Component Analysis) and LSA(Latent Semantic Analysis)
  - Adjust F3 such that titles are added in a weighted manner

Potential Applications

- Browser Extension :
  - The model could be integrated into a browser extension to help users evaluate or reflect on the validity of any given article
- Social Media:
  - To help combat the issue of spreading fake news via social media platforms like facebook, a model like ours could be used to help filter out or identify fake news posts

Questions?