Heinz 95-845: Classifying Fake News

Alton Lu altonl/lualton@cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

Sam Hobbs shobbs/shobbs@andrew.cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

Michael Pan zpan1/michaelpan@cmu.edu
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

1. Abstract

The impact of misleading information across social media platforms, online newspapers, and blogs have made it difficult to identify trustworthy news sources. Many platforms have begun to utilize computational methods to provide insight into the trustworthiness of content while research in this area has expanded. However, one of the key difficulties in these computational methods is the generalizability to different domains and content-styles. A well-trained classifier on 2016 political news may struggle with 2018 elections or with smaller statements. In this paper, we focus on an experiment of aggregating different fake news datasets to test the value of adding data from different domains (sports, news, politics, etc) and content-styles (news, blogs, statements). The results show that generalizability across domain can be achieved by adding more domain data. However, generalizability across content-styles is still difficult and is a key area to continue research.

2. Introduction

Fake news has been a big issue since the 2016 Presidential election and the Brexit vote. This notion of fake news continuous to be a recurring topic through the 2018 election and has continued to spread across social media platforms. Social media continues to become commonplace for US adults to source their news, estimating rate of 62 percent for US adults (Gottfried and Shearer 2016). Social media users tend to place higher trust for content shared by contacts who are part of their social media contacts (Bakshy et. al 2015) creating echo-chambers where ideas and opinions are likely propagated and reinforced, irregardless to the truthfulness supporting the ideas (Pelley, 2017).

Governments and society have struggled in determining what kind of intervention and who is responsible for intervention - or if any intervention is indeed appropriate - in combating the propagation of fake news. Organizations have attempted to use computational

methods to address the proliferation of fake news on their platforms. However, this is not always easy.

Misleading information, or fake news, can differ widely in domain and content-style. Politics and celebrity news are common areas of fake news. And these often come in the form of a news story or a short 280 character tweet. This issue also questions the robustness behind underlying assumptions in many common fake-news-detection machine learning initiatives. Most fake news detection projects focus on either one domain (politics) or of a single content-style (news article or blog). Further steps (Rosas et. al 2017) have been made to diversify fake news detection among differing topic domains.

The goal of this paper is to detect fake news in a way that more represents how users might consume content on any given platform. This includes considering diverse topic domains and differing content-styles, including news, statements, or blogs. Weve aggregated multiple well-structured and previously-used data sets that reflect various topic domains and content-styles. Topic domains include business, technology, celebrity news, and politics. The content style includes news articles, blogs, and short statements by politicians.

Section 2 in this paper describes the background overview in natural language processing (NLP) tasks, as well as some descriptions of well-used methods in detecting fake news. Section 3 reviews the current literature and background work in more detail. In Section 4, we discuss the experimental setup, preprocessing, and feature engineering, as well as evaluation criteria. Section 5 finishes the the results of the experiment. Section 6 concludes and describes the direction of future work in detecting fake news.

3. Background

To approach our task of detecting fake news, we employ several NLP techniques that enable machine learning on a body of text. These techniques attempt to represent characteristics of the text as we transform the text to fit a computational model. Representation of the text is considered as a feature.

Features of the text that we evaluate include punctuation usage, polarity and sentiment, objectivity, ease of readability, exclusive word usage, and possessive pronoun usage. Many of these features are evaluated in terms of a numerical score. Processing words from text into numerical framework to represent document vocabulary usually includes techniques such Bag of Words or Term Frequency Inverse Document Frequency (TF-IDF). For this research, we have employed Word2Vec as it creates word embeddings in a vector space to represent document vocabulary and attempts to capture relationships of a given word to other words in the corpus.

The previously mentioned processing techniques transform the data to then train classification models in efforts to detect fake news. We train a Nave Bayes, K-NN, and Random Forest classifiers and employ ensembling techniques of boosting and bagging.

4. Related Work

There has been a wealth of related literature and work in this area of fake news. Many researchers argue that this is an extension of work around misinformation and disinformation. Before understanding fake news, we had to try and understand a little more about the

nature of deception. Newman et al., (2003) ran multiple experiments to test how people try to deceive others in real life. Its unlikely for deception to be an outright falsehood. The authors stated that convincing stories involve the subtle manipulation of language and the careful construction of stores that are sincere. The authors found that deceptors use fewer first-person pronouns, third-person pronouns, more negative emotions, few exclusive words (but, except, without). These findings have been similarly replicated in Bond et al., (2005). Bond researched the nature of deception in prison and found similarly that less pronouns and exclusive words were used while negative words were more common.

There are several assumptions for these linguistic differences in deception. DePaulo et al., (2003) found that liars was to dissociate themselves from the lie. Liars may also tend to feel guilty about their roles and use more negative words. The lack of exclusive words also suggests simpler ideas. Words such as however or except tend to imply a distinction about a particular statement and is typically a more complex statement.

Another area of study to understand is spam detection. We look to the development of classifiers for email spam to approach classification of fake news. Aski et al., (2017) tries to bring advanced machine learning algorithms to filter spam with low error rates. They considered a rule-based approach to assign scores to spam emails. There were certain rules associated with linguistic features to be indicative of spam, from the number of apostrophes, the amount of white space, or unclear subject field. The thinking behind these rules can be extended to the work in fake news detection.

Shu et al. (2017) described an approach to fake news detection on social media. This paper discusses two overarching methodology in detecting fake news. The first is a news content approach, which relies on features within headlines, articles, and images. The models rely on either knowledge-based approaches or syntactic style-based methods. This is the approach taken in our paper.

Shu also describes a social context models, which relies on metadata around the news. This includes social-media data such as likes, network propagation, and user-reactions. As the dataset we work with does not include social media data, we do not consider this approach. However, our understanding, corroborated by other studies, show that these social features are extremely powerful in classifying fake news.

In Yang (2017), a new dataset is created from a decade-long, manual label of political statement from PolitiFact.com. This serves as another context of fake news, focusing on short statements without as many social features. Its one of the largest sets of data created on fake news. This paper showed some limitation of surface-level linguistic patterns in sentence-long detection, but an approach with metadata around the politician can be useful.

Perez-Rosas et al., (2017) contributed two novel datasets (both of which are used in this paper) and a set of experiments to detect fake news. This paper used several linguistic properties such as Ngrams, punctuation, psycholinguistic features (positive emotion, subjectivity) to perform reasonably across news from 7 different domains. The key aspect of this paper is around trying to extend the context of news, from politics to celebrities to test computational methods for detection.

5. Experimental Setup

The goal of our experiment is to test classification of fake news across different domains and context, extending the work done in Rosas-Perez 2017. There are several datasets that we obtain in different domains, contexts, and sources.

We utilize a variety of common classification algorithms to test our experiment. These are random forest, AdaBoost with Random Forest, Naive Bayes, and K-Nearest Neighbors.

5.1 Fake News Datasets

In this section, we describe the five datasets we aggregate in order to test our experiment.

News Dataset (Rosas-Perez) This dataset comes from Rosas-Perez 2017 and is split evenly across real and fake news stories. They began by collecting legitimate news from six different domains (sports, business, entertainment, politics, technology, and education). These news came from publishers (ABCNews, CNN, USAToday, New York Times, Fox News, Bloomberg, and Cnet.

The authors then generated fake news using Amazon Mechanical Turk (AMT). They had workers create a fake mirror story for each of the legitimate new stories. The authors note that the quality of these fakes may be low due to AMT paying by quantity of tasks completed.

> **Real**: *Ryan Seacrest not only takes the New York subways, but he also appreciates the occasional performances on the trains. On Tuesday, the Live co-host recorded and expressed how much he enjoyed a subway singer performing*

> **Fake**: *Jennifer Aniston is making Friends fans very happy! The actress hinted she would be open to appearing in a reunion of the hit NBC show. "The only thing I can think of doing is maybe for fun doing a Thanksgiving episode," Aniston said...*

Celebrity Dataset (Rosas-Perez) This is another dataset from Rosas-Perez 2017 that is around the domain of celebrities in the United States. This is a domain where fake content tends to occur naturally and in large numbers. The data from this come from tabloid publishers such as Entertainment Weekly, People Magazine, and RadarOnline. The veracity of data was evaluated from gossip-checking sites such as GossipCop.com.

> **Real**: *Apple is losing its grip on American classrooms, which technology companies have long used to hook students on their brands for life.Over the last three years, Apple's iPads and Mac notebooks - which accounted for about half of the mobile devices...*

> **Fake**: *Alex Jones, purveyor of the independent investigative news website Infowars and host of The Alex Jones Show, has been vindicated in his claims regarding the so-called "Pizzagate" controversy...*

Fake Kaggle (Risdal and McIntire) This dataset was provided by Kaggle in the aftermath of the 2016 election. It contain largely mostly fake news sources compiled by Risdal.

McIntire expanded this data in 2017 with a curated list of authentic news stories to create a more comprehensive dataset.

> **Real**: *'U.S. Secretary of State John F. Kerry said Monday that he will stop in Paris later this week, amid criticism that no top American officials attended Sundays unity march against terrorism...*
>
> **Fake**: *'Daniel Greenfield, a Shillman Journalism Fellow at the Freedom Center, is a New York writer focusing on radical Islam. In the final stretch of the election, Hillary Rodham Clinton has gone to war with the FBI...*

Liar Liar (Wang-PolitiFact) This corpus was created by Wang 2018 and utilizes data from PolitiFact. It contains 12.8k statements from politicians and classified on a scale of pants fire, false, barely true, half-true, mostly true, true. These are short statements that come from speeches, Facebook, Twitter, press releases, etc. and thus represent text from a wide-variety of contexts. This is particular valuable as it includes statements that are prepared (press releases) as well as potential off-the-cuff statements from interviews.

In our combined dataset, we undersample the Liar dataset, taking only 2500 observations witt 50 percent real and 50 percent fake.

> **Real**: *'Has created 60,000 net new jobs since taking office.'*
> **Fake**: *'The incandescent light bulb has no effect whatever on the planet.'*

BuzzFeed Fake News Corpus 2016 (BuzzFeed) This is a corpus of news stories from nine publishers within a week of the US 2016 presidential election. It includes three known hyper-partisan left-wing publishers, three hyper-partisan right-wing publishers, and three mainstream publishers. The data were published on February 20, 2018. Every single publisher in this data earned a blue checkmark from Facebook. The publishers in this data are: ABC News, CNN, Politico, Addicting Info, Occupy Democrats, The Other 98%, Eagle Rising, Freedom Daily and Right Wing News.

> **Real**: *'With the Hillary Clinton-Donald Trump debates upon us, the quadrennial question comes begging: Do these showdowns matter? The chances for impact seem ripe this year. The two most unpopular major-party candidates...*
>
> **Fake**: *'ABC News is bringing its original livestream series Strait Talk with Matt & LZ to the University of Detroit Mercy in front of a live audience on Monday, Sept. 19 at 7 p.m. ET for an open discussion about the biggest issues facing America this election cycle. ABC News Political Contributors Matthew Dowd...*

5.2 Preprocessing

All analysis in this paper were performed in Python, with scikit-learn, nltk, textblob, and gensim packages.

Because we are aggregating several different datasets all with different features and ways of labeling veracity of a news story, there are several important choices made to combine

Dataset	Counts	Real	Fake	Avg Words	Max Words	Min Words
Risdal-McIntire	3171	3164	6335	776	20891	0
Wang-PolitiFact	1250	1250	2500	18	61	2
Buzzfeed	1264	151	1627	560	5451	5
Rosas-Celebrity	250	250	500	413	14515	28
Rosas-News	240	240	400	120	287	55

Table 1: Data Summary

the data. These choices will have large ramifications on the results of our experiment and should be taken into account.

Subsetting Labels
Each dataset had a different labeling methodology. The Kaggle dataset included a variety of labels from fake, real, biased, or bs labels. The Liar dataset had far more granularity with the following levels: pants-fire, false, barely-true, half-true, mostly-true, true. Buzzfeed had four levels of the following: mostly false, mixture of true and false, mostly false, and no factual content.

To keep the problem simpler, we removed half-true, biased, and mixture labels. We then compressed labels such as mostly-false or mostly-true into FAKE and REAL, respectively. This allowed the problem to focus on binary classification.

Removing metadata
While previous research has shown metadata to be quite valuable in detecting fake news (Shu et al. 2017), we remove much of the metadata, especially around social networks. Because of the different context that these data are found (online, public statement, etc), the metadata is not consistent. We remove those to focus only on linguistic features.

Missing Data
Due to errors in construction or unavailability, there were some missing data. The author or the title may have been missing in some cases. In all missing text values, we replaced the null with an empty string . This means most of our feature extraction values are 0 when applied to the empty data, but allows us to get a value.

5.3 Feature Engineering

To create the classifier, we attempt to engineer several features from the text and title of the stories.

Punctuation Ratio (Ott, 2011): In the realm of spam detection, the use of punctuation was often a powerful predictor of spam. In an exploration of our datasets, we also see that fake news often has many exclamation points (!) or ellipsis (...). Our assumption is that liberal use of punctuation marks is often related to biased and fake news stories.

Polarity/Sentiment (Shu 2017): Sentiment appears to have some useful features. An understanding of a large part of fake news is a negative viewpoint towards whatever

topic that is being written.

Objectivity: We utilize the TextBlob package in Python to calculate a subjectivity score for each text. This takes on a value from 0 - 1 (subjective - objective) on how subjective a text is. Our assumption is that fake news tends to be more subjective about the subject matter while proper news stories are more objective in nature.

Readability Ease: This is a method developed in Farr et al., (1951). Reading ease is calculated with the following formula: Reading Ease = 206.835 - (1.015 x Average Sentence Length) (84.6 x Average Syllables per Word).

Readability Grade: This method was developed in Kincaid et al., (1975) to score the difficulty of material. The Grade-level is calculated with the following formula: Grade = (0.39 x Average Sentence Length) + (11.8 x Average Syllables per Word) - 15.59.

Exclusive Word Usage: As stated in the the related work section, previous work in deception found that people are less likely to use exclusive words, such as, however or although (Newman 2003). We create a dictionary of common exclusive words in the english language and compute a ratio of usage for each story.

Possessive Pronouns: Again, as stated in the the related work section, previous work in deception found liars try to disassociate themselves from the lie by using less possessive pronouns like we and I (Newman 2003). While fake news is a different context than lying to a person face-to-face, we believe there are similar patterns. The goal of fake news is deception and this would typically be to create a they theme in the story. There would a less focus on what the in-group is doing and more on why an outside group is bad.

Word2Vec (Mikolov, et al. 2013): Word2Vec uses a shallow neural network to transform a given corpus of text into word embeddings mapped in a vector space. We use Word2Vecs continuous-bag-of-words (CBOW) model architecture. The intention behind using Word2Vec over traditional text processing techniques, such as Term Frequency Inverse Document Frequency (TF-IDF) is to capture relationships such as document vocabulary and word context, semantic, and syntactic representation of words within a document through word embeddings

5.4 Evaluation Criteria

We evaluate based on pure accuracy. There are numerous discussions and opinions around the cost of false positives and false negatives in fake news. We similarly could not come to a conclusion. The cost of a false positive is a journalists work - which can be substantial - being removed from a platform. Thats a cost to the publisher, as well as potential readers. The cost of a false negative is similarly difficult. A single false negative is fine, but our opinion is that many fake news articles can be extremely damaging to a public news space.

We also breakdown our results into two separate subanalysis. The first is a confusion matrix by each dataset. And then the same confusion matrices in a model that did not include the Liar dataset by Wang.

The reason for analyzing with and without the Wang Liar-PolitiFact dataset is that the structure of the data is so different. Were interested in understanding how well our model can generalize to different domains as well as content-styles, but recognize that its a very difficult problem.

6. Results

In this section, we present the results of our experiment to classify fake news. The complete model accuracy results are found in table 2. Each subsection describes a separate evaluation of this task. We find reasonable success at classifying fake news, especially with the boosted

Model	Accuracy w/o Liar	Accuracy w/ Liar
Random Forest	78.01	71.69
AdaBoost+RF	78.79	72.55
Naive Bayes	69.13	65.11
KNN	64.76	60.68

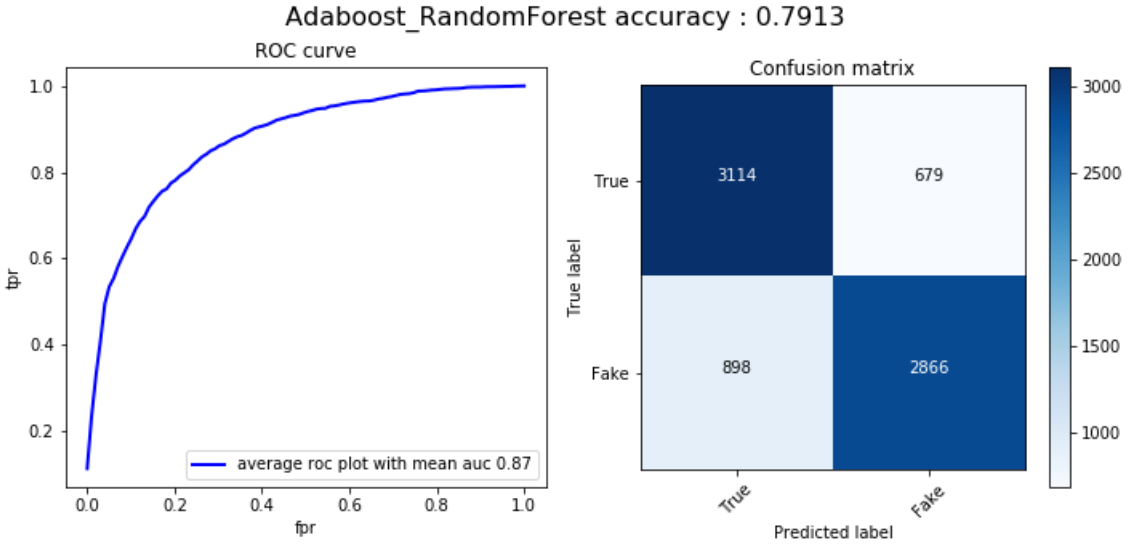
Table 2: Model Results

random forest. The key insight though is that adding a new datasets with a very different content style drastically reduced the quality of our classification.

In the next sections, we look more in-depth at the results of the boosted random forest for each sub-part of the classification.

6.1 Results without Liar

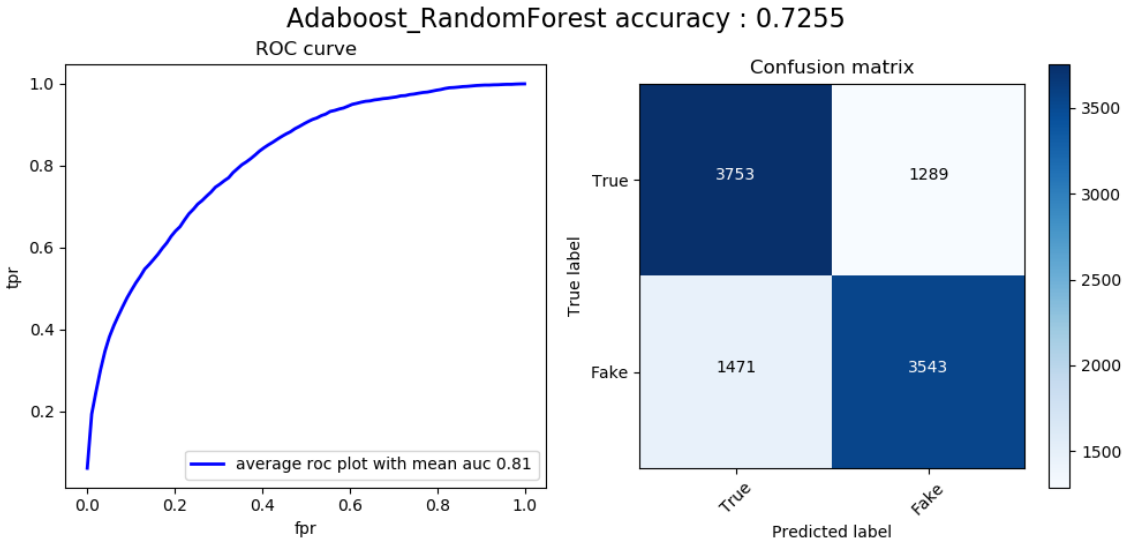
This section describes the model results without utilizing the liar dataset. We see that the model performs well at around 80% accuracy. This is paired with a 77% sensitivity and an 80% specificity. Both false positives and false negatives are roughly the same.



6.2 Results with Liar

This section includes the Liar dataset, which are the short true-false statements labeled by PolitiFact and curated in Wang 2017. We re-train our model on this new combined dataset to get final results.

We see that with the addition of this completely new content-style (statements), accuracy drops by around 8%. Sensitivity and specificity both see drops, with sensitivity at 71.84% and sensitivity at 73.32%.

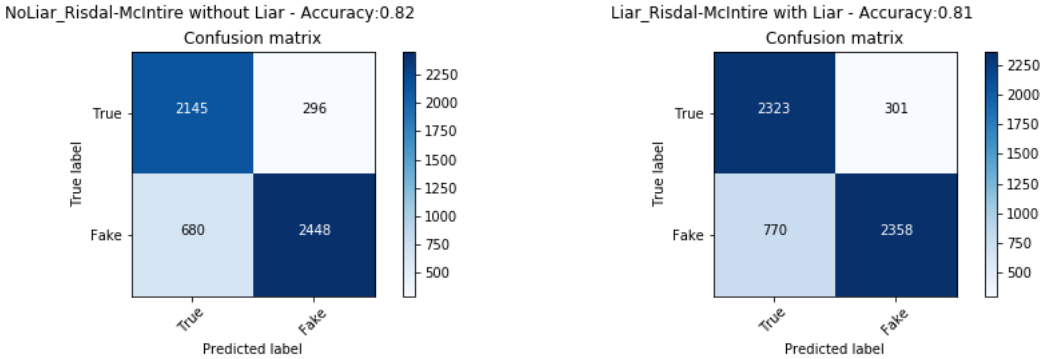


6.3 Results for each dataset

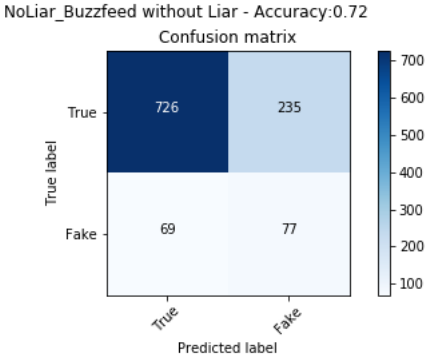
From the previous table, our assumption is that the Liar dataset was not well predicted. In this section, we show confusion matrices for each separate dataset within the model.

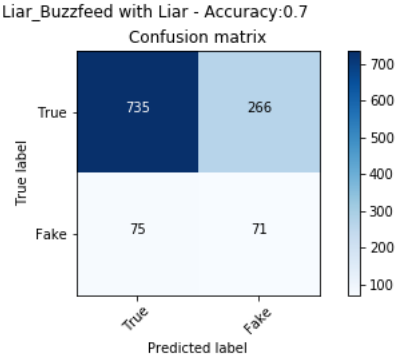
Risdal-McIntire (No Liar, Liar)

With the Risdal-McIntire dataset, we see that performance decreases slighlty with the addition of the liar dataset. There is something within attempting to capture information in short statements that reduces some quality in news stories.

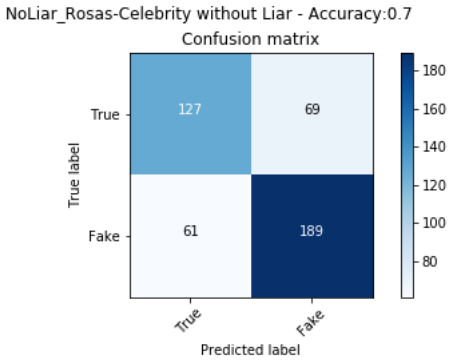


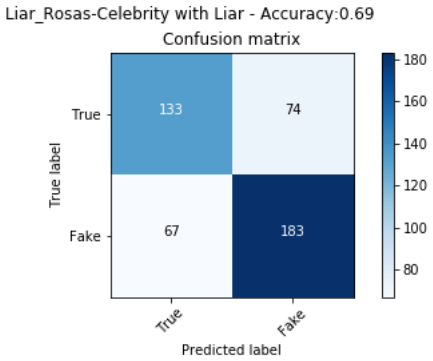
Buzzfeed (No Liar, Liar)// Our Buzzfeed dataset seems similar performance, with an accuracy drop of 2%.



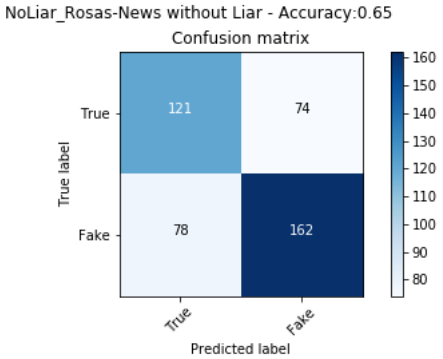


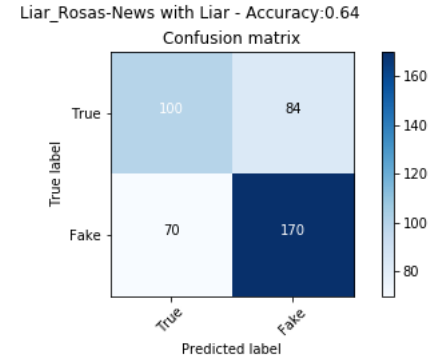
Rosas-Celebrity (No Liar, Liar) In Rosas-Celebrity, we see that same 1% accuracy decrease. However, the key insight here is that the celebrity dataset was near the performance of the Buzzfeed news dataset, meaning that domain extension was valuable. We're able to get a completley different domain just by adding data. This is intuitive and makes sense (more data is better).





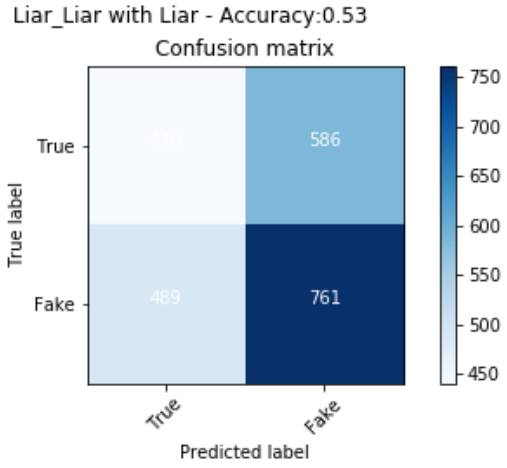
Rosas-News (No Liar, Liar) With the Rosas-News data, we see that performance drops and is generally inaccurate. This may be because of the fake news construction with the AMT workers creating false stories. Their intent may not be to deceive as other fake news, thus some features may not be as informative. The fake news was also created from real news, meaning much of the syntactic structure was the same.





Liar Confusion Matrix We see our hypothesis confirmed in the liar confusion matrix. We have a 53% accuracy, which is essentially flipping a coin. This was also the misclassified values in our full confusion matrix, meaning that the inclusion of this new content-style did not change our algorithms performance. It simply found the information not-valuable for performance purposes.

The prediction power of our classifier tends to come from the other four datasets. This is reaosnable, as the Liar dataset is extremely limited in content. It's difficult to tell when someone is lying based on a single sentence.



One key insight is how the decrease in accuracy across all datasets is less than the proportion of liar data added. Our 53% accuracy in liar is slightly better than random chance while every other dataset only lost about 1% in accuracy.

7. Conclusion and Future Work

This paper showed that domain extensibility in fake news classification is possible just by adding data from different domains. We saw that performance in the celebrity dataset was similar with the larger news datasets. However, this paper showed the difficulty of extending a classifier to different content styles. This was demonstrated by the performance with the Liar dataset by Wang.

Furthemore, much of this application is simplified by removing the middle veracity (half-true, half-false) news. While this was done due to limitations of the data, its identifying half-true stories that could be the most valuable. Someone trying to generate fake news with malicious intent is more likely to disguise within some level of truth.

These two demonstrations show where future work needs to move.

The first is content-style extensibility. Models trained on certain types of news do not work very well on different content types. We need to work on including data around different content-styles, from blogs, to statements, to press releases. This can also be extended to public forums, such as congressional hearings. Working around these different content styles can lead to more generalizable fake news detectors.

The second is around data quality and data volume. The datasets used in this paper are some of the gold-standard available at the current time, but do not follow any standardized structure. For research in this area to continue, there needs to be an emphasis on collecting news from a variety of sources and a variety of labels. For half-truth news to be properly evaluated, there needs to be more data. Or that research in this area should start focusing on biased news.

For organizations attempting to pursue more computational methods for detecting fake news, our research shows how generalizing to different domains can be achieved by adding more data about different domains. However, it is not enough to simply add data about different content-styles. We see that there may be more value in creating separate classifiers based on each separate content-style instead.

References

[1] Ali Shafigh Aski, Navid Khalilzadeh Sourati. 2016. Proposed efficient algorithm to filter spam using machine learning techniques. Pacific Science Review A: Natural Science and Engineering Vol 18, Issue 2, pages 145-149.

[2] Charles Bond and Bella DePaulo. 2006. Accuracy of Deception Judgments. Personality and Social Psychology Review. Vol. 10, No. 3, 214234

[3] Gary D. Bond, Adrienne Y. Lee. 2005. Language of lies in prison: linguistic classification of prisoners' truthful and deceptive natural language. Applied Cognitive Psychology Volume 19, Issue 3.

[4] Niall. J. Conroy et al. 2016. Automatic Deception Detection: Methods for Finding Fake News, ASIST 2015.

[5] Bella M. DePaulo, et al. 2003. Cues to Deception. Psychological Bulletin Vol. 129 Issue 1, pages 74 - 118.

[6] James N. Farr, James J. Jenkins, Donald G. Paterson. 1951. Simplification of Flesch Reading Ease Formula. Journal of Applied Psychology, Volume 35(5), page 333-337

[7] C.J. Hutto and Eric E. Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).

[8] J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. Institute for Simulation and Training Paper 56.

[9] Matthew L. Newman, James W. Pennebaker, Diane S. Berry, Jane M. Richards, 2003. Lying Words: Predicting Deception from Linguistic Styles. Personality and Social Psychology Bulletin, Vol. 29, pages 665-675.

[10] Veronica Perez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea. 2017. Automatic Detection of Fake News. International Committee on Computational Linguistics.

[11] Martin F. Porter, 1980, An algorithm for suffix stripping, Program, 14(3) pp 130137.

[12] Giovanni C. Santia, Jake Ryland Williams. 2018. BuzzFace: A News Veracity Dataset with Facebook User Commentary and Egos. Association for the Advancement of Artificial Intelligence ICWSM 2018.

[13] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter: Social and Information Networks.

[14] Mikolov, Tomas; et al. 2013 "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781

[15] William Yang Wang. 2017. Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection. Annual Meeting of the Association for Computational Linguistics.