Fake news identification: a comparison of parts-of-speech and N-grams with neural networks

Brandon Stoick, Nicholas Snell, Jeremy Straub

Brandon Stoick, Nicholas Snell, Jeremy Straub, "Fake news identification: a comparison of parts-of-speech and N-grams with neural networks," Proc. SPIE 10989, Big Data: Learning, Analytics, and Applications, 109890D (28 May 2019); doi: 10.1117/12.2521250

SPIE.

Event: SPIE Defense + Commercial Sensing, 2019, Baltimore, Maryland, United States

Fake News Identification: A Comparison of Parts-of-Speech and N-grams with Neural Networks

Brandon Stoick, Nicholas Snell, Jeremy Straub
Department of Computer Science, North Dakota State University, Fargo, North Dakota, USA

ABSTRACT

The rise of the internet has enabled fake news to reach larger audiences more quickly. As more people turn to social media for news, the accuracy of information on these platforms is especially important. To help enable classification of the accuracy news articles at scale, machine learning models have been developed and trained to recognize fake articles. Previous linguistic work suggests part-of-speech and N-gram frequencies are often different between fake and real articles. To compare how these frequencies relate to the accuracy of the article, a dataset of 260 news articles, 130 fake and 130 real, was collected for training neural network classifiers. The first model relies solely on part-of-speech frequencies within the body of the text and consistently achieved 82% accuracy. As the proportion of the dataset used for training grew smaller, accuracy decreased, as expected. The true negative rate, however, remained high. Thus, some aspect of the fake articles was readily identifiable, even when the classifier was trained on a limited number of examples. The second model relies on the most commonly occurring N-gram frequencies. The neural nets were trained on N-grams of different length. Interestingly, the accuracy was near 61% for each N-gram size. This suggests some of the same information may be ascertainable across N-grams of different sizes.

Keywords: fake news, neural networks, machine learning, part-of-speech, N-gram, natural language processing, social media reliability

1. INTRODUCTION

Fake news identification is of growing importance as more people turn to social media and websites for news and information. According to the PEW research center, a 2017 survey revealed two-thirds of adults consume news on through social media[1]. Content posted on social media, especially, is often less rigorously vetted than traditional news outlets. Furthermore, the barriers to market entry, or in this case the ability to distribute content to a large audience, tend to be lower when aided by the internet[2]. Modern social media platforms have streamlined processes to create an account and post content with ease. This lack of barriers has enabled a greater number of people to distribute their thoughts to audiences across the world. Combined with the lack of a review process and information vetting, the ease of distribution has created an environment susceptible to disinformation. Unfortunately, disinformation can have widespread negative consequences, especially when recognition as such is low[3]. Validity of information can be critically important as people make economic, political, health and other important decisions based on information they consume. Information is a powerful[4] resource but significantly less so when tainted by unidentified deception.

To combat fake news, a variety of approaches have been explored. Fact checking sites such as Snopes post findings on the validity of facts to help readers assess the information they read[5]. However, readers often don't take the time to examine each piece of information for validity and could benefit from classification of the article as a whole. The proposed classifier could be embedded in a platform and flag intentionally deceptive news[6]. In development of a classifier, a range of extracted linguistic features have been explored. Conroy, et al. developed a model to differentiate celebrity news articles leveraging a neural network and extracted features including part-of-speech frequencies, N-grams, psycholinguistic features and readability scores[7]. Other approaches focus on how content disseminates through social networks[8] or semantic analysis of the text[9]. A variety of machine learning techniques have been employed with success including random forests, SVM's, Naïve Bayes, and neural networks[10]. Within neural networks, techniques including the use of LSTM's and the attention mechanism[11] have been utilized to capture the sequential nature of article text.

A major challenge in classifying news as fake or real stems from diverse intent within the news space. Satire, for example, is not constrained by validity and would be classified as fake if the only factor considered was the accuracy of the information presented. In data collection, articles gathered for the fake designation were chose for their intent to deceive.

Big Data: Learning, Analytics, and Applications, edited by Fauzia Ahmad,
Proc. of SPIE Vol. 10989, 109890D · © 2019 SPIE · CCC code:
0277-786X/19/$18 · doi: 10.1117/12.2521250

Satire and other humous forms are generally not meant to be interpreted as factual[12] and were not included in the data collected. Real news was gathered from a variety of sources. Preliminary neural network training showed the model to be sensitive to the source. Classification was exceptionally accurate for the limited sources the network was trained on but low on articles from a priori sources. To combat this lack of generalization, the bulk of the dataset was collected from many different sources.

Intent, such as deception, may drive differences in the structure of sentences and in word choice. Part-of-speech frequencies can reflect such structural differences and can correlate with malicious intent. The purpose of this paper is to evaluate and compare the effectiveness of part-of-speech frequencies and N-gram frequencies in training a neural network to classify news articles as fake or real. So, the first model developed uses solely part-of-speech frequencies in the binary classification of news articles. The accuracy of this approach demonstrates a relationship does exist and that part-of-speech frequencies could be an important part in more complex ensemble models. The second model explores the effectiveness of using common N-grams as extracted features. The neural net was trained on the forty most common character group frequencies. N-gram sizes were varied from a single letter up to five letters per group.

2. METHODOLOGY

The methodology first discusses the reasoning for using part-of-speech frequencies in training the classifier. Then, the tagging process used to label the dataset is outlined. The second model, using N-gram frequencies, is presented next. Finally, the feature vectors and neural network structures used in both models are discussed.

2.1 Part-of-Speech Frequencies

To develop a classifier, features must be selected for use as inputs. Part-of-speech frequencies are one such linguistic feature and have been shown to useful in the classification of news articles[13]. Key to this selection is determining how part-of-speech frequencies relate to the validity of a news article. To investigate this relationship, an examination of the article development process and writer intent is needed. Writing, no matter the style or platform, is driven by a purpose. Sometimes, a writer may generate content for the sake of generating content. Quotas, deadlines, and budgets must be met, and writing is done to fulfill this need. Often, however, individual writers have another or additional purpose when writing: they seek to affect their audience in some way. An author may seek to inform, entertain, persuade or deceive their readers. This difference in intent manifests itself in a variety of ways, prominently of which is word choice[14]. Words are the fundamental foundation of English written communication. Meaning is derived from their usage and structural arrangement. In attempt to extract the meaning from news articles, the first model examines part-of-speech frequencies within the text. Each word in the English language falls under some part-of-speech category based on its meaning or functional interaction with other words. The first model developed assumes some relationship between these part-of-speech frequencies and the validity of the article. The premise is a difference in intent will drive a difference in diction discernable in the distribution of part-of-speech frequencies. The model doesn't, however, account for the arrangement of the part-of-speech frequencies. As discussed, both the diction (word choice) and the arrangement (syntax) influence meaning in the English language.

Table 1. Sample sentence and part-of-speech labels from fake article (see Table 3 for key).

I	*guess*	*its*	*all*	*he*	*knew*	*how*	*to*	*do.*
PRP	VBD	PRP$	DT	PRP	VBD	WRB	TO	VB

Table 2. Sample sentence and part-of-speech labels from real article (see Table 3 for key).

In	*it*	*,*	*he*	*explores*	*complex*	*issues*	*through*	*comedy.*
IN	PRP	,	PRP	VBZ	JJ	NNS	IN	NN

Tables 1 and 2 present sample sentences from fake and real articles, respectively. In the first sample, the author uses first person and diction not usually characteristic of vetted reporting. The differences in diction are partially visible in the part-of-speech frequencies. The first person "I" in the fake sample is obscured by other prepositions, but the use of past tense

is evident. The examples, while specifically selected from the data, highlight the possibility for differences in part-of-speech frequencies. Of note, the fake sample neglects to include an apostrophe in "its" and the corresponding part-speech-label reflects the choice.

2.2 Part-of-Speech Tagging

In the English language, words can have different part-of-speech designations based on their usage. Consequently, context clues and sentence structure should be examined to facilitate accurate classification. The part-of-speech tagger used in the model is part of Python's natural language processing toolkit. The toolkit examines the context of each word to more accurately tag parts-of-speech. The dataset was tagged using 34 possible designations, as outlined in Table 3.

Table 3. Part-of-Speech tags used as inputs.

Code	Part-of-Speech
CC	Coordinating Conjunction
CD	Cardinal Digit
EX	Determiner
FW	Existential There
IN	Foreign Word
JJ	Preposition/Subordinating Conjunction
JJR	Comparative Adjective
JJS	Superlative Adjective
LS	List Marker
MD	Modal
NN	Singular Noun
NNP	Singular Proper Noun
NNPS	Plural Proper Noun
NNS	Plural Noun
PDT	Predeterminer
POS	Possessive Ending
PRP	Personal Pronoun

Code	Part-of-Speech
PRP$	Possessive Pronoun
RB	Adverb
RBR	Comparative Adverb
RBS	Superlative Adverb
RP	Particle
SYM	Symbol
TO	To
VB	Verb
VBD	Past Tense Verb
VBG	Present Participle Verb
VBN	Past Participle Verb
VBP	Present Tense Verb
VBZ	3rd Person Verb
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Wh-possesive pronoun
WRB	Wh-adverb

2.3 N-gram Frequencies

The effectiveness of N-gram frequencies was also explored for classifying articles. Again, the difference in intent between fake and real articles may drive differences in diction. With N-grams, certain words or even constituent sounds may be more likely to appear when attempting to deceive the audience. Appeals to emotion, for example, may accompanied by a tendency toward dissonance or harsh sounding language. If dissonance was more likely to be found in fake articles, the N-gram frequencies may display such a bias and be ascertainable by the neural net. With larger groups, entire words are often the highest frequency character combinations. With these vectors, the classifier may be able to learn what words, or what proportions of what words, show up more frequently in fake or real articles.

Table 4. Fifteen most common N-grams in dataset from highest to lowest frequency.

1 Char	2 Char	3 Char	4 Char	5 Char
t	th	the	that	Trump
a	to	and	with	Their
s	an	tha	said	about
o	of	for	have	Ameri
w	in	The	from	would
,	co	con	this	Democ
i	re	wit	Trum	state
c	ha	was	they	peopl
.	be	thi	thei	which
h	wh	pro	abou	repor
b	fo	not	Amer	other
p	wa	hav	been	offic
f	on	are	woul	Presi
d	wi	sai	will	gover
m	pr	has	stat	again

To train the neural net, 40 of the most common N-grams within the dataset were identified. Then, the frequency of occurrence of each group was determined for each training example. The classifier was trained on these one-dimensional vectors and the results were compared across N-gram sizes.

2.4 Neural Network Structure

The models developed use one-dimensional vectors as inputs. The first model uses part-of-speech frequency vectors of length 34. The second model uses the 40 most common N-gram frequencies. The effectiveness of single up to five-N-grams were explored in separately trained neural nets. To mitigate the impact of the length of the article, each part-of-speech or N-gram input is represented as a percentage of its occurrence within the body of the article. Often, articles lacked certain parts-of-speech or N-grams. However, such findings were not always unfavorable as the neural network may be able to learn a relationship between the lack of a part-of-speech and the validity of the article if it existed. Exploration of different numbers of hidden layers and nodes revealed this smaller network to be better for the limited data available. Other models proposed in literature include the use of recurrent neural network structures such as LSTM's and GRU's[15]. The model developed only used feed-forward fully connected layers but could be readily modified to different neural network structures and techniques. Overfitting was of concern, but the smaller network combined with high dropout rates helped mitigate problem.

3. RESULTS

First, the performance of the neural network trained on part-of-speech frequencies is presented. The subsequent relationship between the proportion of the dataset used in training and prediction accuracy is then discussed. Next, the accuracy of the second model is presented for N-grams of different length.

3.1 Part-of-Speech Model Performance

The first model was trained for 500 epochs. Above 500 epochs, overfitting became especially troublesome. Even with high dropout and other overfitting mitigation techniques, the model consistently achieved 100% training classification above 1000 epochs. Validation on the test set revealed extreme overfitting. Also, shuffling the training and test data typically varied the accuracy by a couple percent in either direction. The rates shown in Figure 2 are a simple average of multiple training runs on different random, but balanced, training-test splits. Further splitting the dataset into train, test, and validation was considered, but initial testing and the limited size of the dataset encouraged only a binary split.

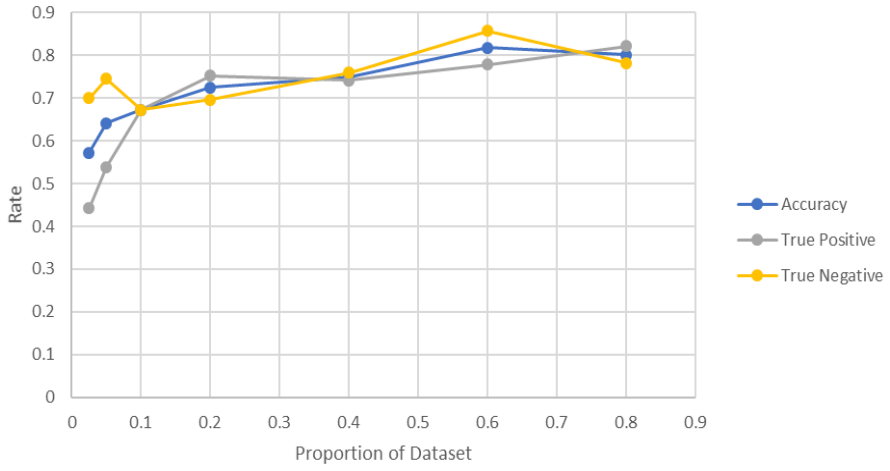

Figure 2. Accuracy, True Positive, and True Negative rates as a function of the proportion of the dataset used in training

As expected, the accuracy of the classifier drops off as the portion of the dataset used for training decreases. Of note, the change in accuracy is consistent from twenty to eighty percent of the dataset. Below twenty percent, however, the accuracy drops off in a manner reminiscent of the sigmoid function. Also, the true positive rate and the true negative rate closely

match the accuracy for portions of the dataset greater than 0.1 (approximately 27 articles). Interesting, however, is the separation between the true positive rate and true negative rate when the neural net is trained on an exceptionally limited subset of the data. The true negative rate holds approximately constant at around 0.7 where as the true positive rate falls sharply off to .45. This suggests some aspect of the false articles may be readily discernable by the classifier, even with just a few examples. This may be result of syntax or diction more commonly found in fake articles. The subsequent differences in part-of-speech frequencies are then identifiable by the classifier. It should be noted, however, that 0.45 is below random guessing on a balanced binary dataset. Likely, the network is increasing the true negative rate at the expense of the true positive rate. As further discussed in regard to the second model, this may be somewhat of a function of the gradient landscape and optimizer used.

Table 5. Confusion matrix for part-of-speech model trained on entirety of dataset.

		Predicted	
		Positive	Negative
Actual	Positive	22	4
	Negative	6	20

When the entire dataset was used for training, the classifier consistently achieved above eighty percent accuracy in testing and the confusion matrix was close to balanced for both correct and incorrect predictions. The true positive rate deviated by a maximum of four classifications from the true negative rate across training runs. Similarly, the false positive and false negative rate deviated only by a maximum of three classifications. Such balance, when using the entirety of the dataset, indicates the model is not favoring one classification at the expense of the other. These findings are favorable as excessive bias would detrimental to the use case of embedding the system in a social media platform.

3.2 N-gram Model Performance

Across N-gram sizes, accuracy held near constant. In training, accuracy was consistently near 60% for many neural networking structures, hyperparameter settings, and training runs. Unless the network was exceptionally small or large, the accuracy deviated minimally. These results suggest some aspect of fake and real articles is identifiable in the appearance of some common N-grams. The exceptional consistency in accuracy across different extracted features suggest some redundancy or similarity of information. At first glance, the varied separation of true positive and true negative rates suggest a difference of information valuable for classification. It posits that the classifier struggles (and excels) with different articles in each extracted feature case. However, this may be due to the gradient decent process. The classifier could increase its true positive rate at the expense of its true negative rate just by consistently moving the predictions to positive without any additional training or net improvement of the constituent weights. With slightly different extracted features, the gradient landscape could change causing the model to converge to a solution minimizing loss to a similar magnitude but consistently favoring one binary classification, especially in cases of more ambiguity. The nearly constant accuracy, however, suggests that this is only partially the case. Some the same information valuable for classification appears to be available across N-gram sizes.

For some of the N-gram sizes, such as length four (shown in Table 6), there was large imbalance between the true positive and true negative rates. Furthermore, the false negative rate was significantly lower than the false positive rate. This tendency to classify the articles as positive may indicate some of the fake articles have a characteristic that is readily identifiable in the N-gram frequencies but that others lack it. In the absence of this characteristic, the classifier errs on the side of true classification since it may struggle to otherwise differentiate the articles. Also, the optimizer and gradient decent process may influence such separation. The gradient landscape might be structured such that the classifier can achieve the lowest loss when favoring the true designation at the expense of the correct negative classification rate.

Table 6. Confusion matrix for N-grams of length four

		Predicted	
		Positive	Negative
Actual	Positive	25	7
	Negative	17	15

Evidently, part-of-speech frequencies are more informative, or at least have a more readily learnable relationship in classification, than N-grams for the collected dataset. The part-of-speech model uses external information learned from a

different dataset—it has the benefit of apply previous knowledge to a new use case in attempt to learn a new pattern. Such transfer learning, while not a direct application of the same skills in one domain to another, can be valuable for many machine learning tasks. Some knowledge (gained in learning to classify parts-of-speech) makes the classification of news articles easier to learn, at least with the limited dataset and neural network developed. Further work could explore the use of other language characteristics in classification. Classifiers could be trained on larger external datasets to recognize features such as personification, simile, juxtaposition, irony and other linguistic features. Then the occurrence of these features could in news articles could then be used in training a neural net classifier.

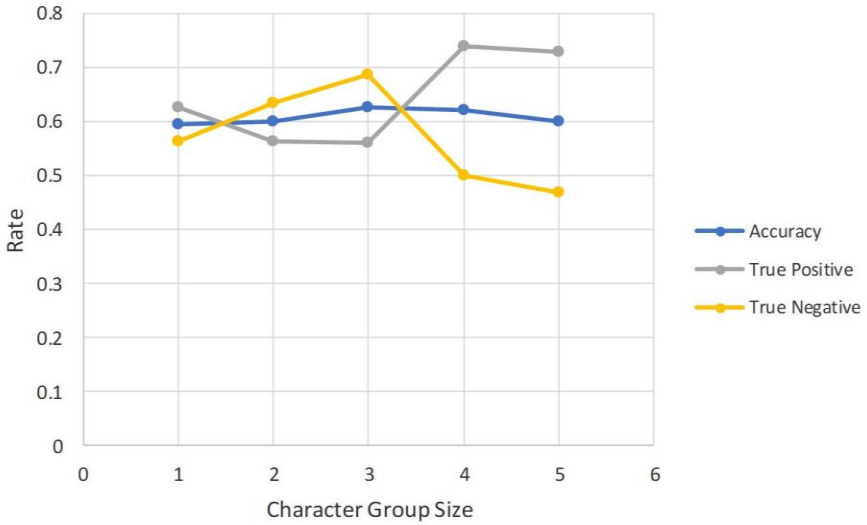

Figure 3. Accuracy as a function of the N-gram size.

4. CONCLUSIONS

Fake news identification is challenging to implement at scale without an automated solution. Part-of-speech frequencies and character N-grams have some correlation to the validity of the article, as demonstrated by the classifiers developed and presented herein. Only using part-of-speech frequencies as input, the neural network consistently achieved above eighty percent accuracy on the dataset collected. Interestingly, when trained on less than ten percent of the data, the classifier still did well identifying fakes. It maintained a high true negative rate, but the true positive rate fell to the accuracy of random guessing. Such findings indicate some aspect of the fake articles is readily ascertainable with a limited dataset using part-of-speech frequencies. N-grams consistently achieved around sixty percent accuracy for character each character group size ranges one to five. The exceptional consistency in accuracy indicates some same information useful in the classification may be present in each of the N-gram sizes. However, some differences exist highlighted by the large changes in true positive and true negative rates across character combination sizes. Combining these frequencies with other inputs on larger training sets should allow classifiers to achieve even greater accuracy. If such models were good enough, they could be integrated into social media platforms to inform users of the likely validity of what they are reading. Future work should explore the deficiencies of this linguistic approach to classification and what supplemental feature extraction best fills the gaps resulting in incorrect classification.

ACKNOWLEDGEMENTS

Thanks are given to Zak Merrigan, Brian Kalvoda, Brandon Stoick, and Bonnie Jan for aiding in collection of the fake news dataset. Thanks are also given to Terry Traylor for his input and feedback on this work. Data collection was partially supported by the U.S. National Science Foundation (NSF Award # 1757659). Facilities and some equipment used for this work were provided by the North Dakota State University Institute for Cyber Security Education and Research.

REFERENCES

[1] Shearer, B. Y. E. and Gottfried, J., "News Use Across Social Media Platforms 2017" (2017).

[2] Granik, M. and Mesyura, V., "Fake News Detection Using Naive Bayes Classifier," 900–903 (2017).

[3] Lewandowsky, S., Ecker, U. K. H., Cook, J. and States, U., "Journal of Applied Research in Memory and Cognition Beyond Misinformation : Understanding and Coping with the ' Post-Truth ' Era," J. Appl. Res. Mem. Cogn. **6**(4), 353–369 (2017).

[4] Volkova, S. and Jang, J. Y., "Misleading or Falsification ? Inferring Deceptive Strategies and Types in Online News and Social Media," 575–583 (2018).

[5] Newman, N., "Journalism, Media, and Technologoy Trends and Predictions 2017," 37 (2017).

[6] Rubin, V. L., Conroy, N. J., Chen, Y. and Cornwell, S., "Fake News or Truth ? Using Satirical Cues to Detect Potentially Misleading News .," 7–17 (2016).

[7] Pérez-Rosas, V., Kleinberg, B., Lefevre, A. and Mihalcea, R., "Automatic Detection of Fake News," 3391–3401 (2017).

[8] Wu, L. and Liu, H., "Tracing Fake-News Footprints," 637–645 (2018).

[9] Ahmed, H. and Sc, B., "Detecting Opinion Spam and Fake News Using N-gram Analysis and Semantic Similarity" (2012).

[10] Shu, K., Mahudeswaran, D. and Liu, H., "FakeNewsTracker: a tool for fake news collection, detection, and visualization," Comput. Math. Organ. Theory (2018).

[11] B, S. S., Fernandez, N. and Rao, S., "3HAN : A Deep Neural Network for Fake News Detection 3HAN : A Deep Neural Network for Fake," 0–10 (2017).

[12] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S. and Choi, Y., "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking," Proc. 2017 Conf. Empir. Methods Nat. Lang. Process., 2931–2937 (2017).

[13] Horne, B. D. and Adalı, S., "This Just In : Fake News Packs a Lot in Title , Uses Simpler , Repetitive Content in Text Body , More Similar to Satire than Real News," 759–766 (2016).

[14] Conroy, N. J., Rubin, V. L. and Chen, Y., "Automatic deception detection: Methods for finding fake news," Proc. Assoc. Inf. Sci. Technol. **52**(1), 1–4 (2015).

[15] Ruchansky, N., "CSI : A Hybrid Deep Model for Fake News Detection."