THE UNIVERSITY OF CHICAGO

Sources Matter:
A Comparison of Fake News Datasets on Linguistic Feature Performance

By

Miaohan Wang

November 2021

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Chenhao Tan
Preceptor: Sanja Miklin

Sources Matter: A Comparison of Fake News Datasets on Linguistic Feature Performance

Miaohan Wang

November 2021

1 Introduction

The issue of fake news has become an increasingly alarming since the 2016 Presidential Election, when a massive amount of unwarranted news articles, some intentionally misinforming, overwhelmed the social media and the internet as a whole [4, 42]. As humans are not necessarily good at detecting fake news, many researchers and social media companies resorted to developing a machine algorithm that helps human readers differentiate and filter out intentionally false information [51]. Computer researchers have developed many fake news detection models, which started at a humble better-than-chance accuracy but recently advanced to an accuracy around 90% [54]. A variety of methods have been adopted for fake news detection, from content analysis to tracing the spread patterns of the news [55]. Overall, fake news detection models rely on one of the four aspects of the fake news phenomenon: (1) text and non-text content of news, (2) the social context of the news, its platform, and distribution, (3) the sender accounts of news, their credibility, and lastly (4) the targeted (credulous) victim's characteristics for early intervention [51]. The first data source,

news content, has been investigated the most because of its easy access and algorithmic efficiency [54]. While models relying on non-text content (i.e. images [19]) have been developed lately thanks to deep learning models [43], text-only content models continue to occupy a major proportion of fake news research, given its simplicity and well-formed theories. More importantly, text-based models require only textual input to classify a news article, so that an alarming fake news article can be detected as soon as it is posted online, providing a possible early intervention before it goes viral [54].

A text-only fake news model relies on feature extraction from text data. These features can be set manually (e.g. sentiment score [10]) or by algorithm (e.g. word embeddings [28]). The manual features usually concern linguistic or psychological cues latent in news text, such as recurring word use [15], reading ease, or anger-related word count [54, 33]. Researchers have been avidly searching for new features and feature combinations that can best reflect the veracity of news pieces in the field of fake news content models. So far, popular features are lexicon features, including bag-of-words [54, 2] and Term Frequency-Inverse Document Frequency (TFIDF) [33], grammar-related features, including Part-of-Speech (POS) tags [14] and syntax tree structures [33], sentiment-related features, including the common sentiment score and emotion-related word counts from LIWC [5, 44], and also cognitive features like readability scores (e.g. Flesch-Kincaid) and subjectivity scores [34, 54]. The comprehensive text-only model of Zhou et al. [54] surveys the greatest number of manual features and achieves a very high accuracy of .892. However, their model was only run on their own two datasets [41], one from POLITIFACT.COM and another from celebrity news sources. While these two datasets are both extracted from the

real world and their diffusion are observed on Twitter, they might not prove that Zhou et al.'s model [54] can maintain its superb performance across news data from all sources, topics, geographical regions, and time frames. Most text-based models developed so far were tested on the researcher's "own" fake news dataset, that is, news pieces collected upon the need to test the model. Although some researchers borrow their dataset from another renowned paper, the field has not yet established a universal dataset that can test the performance of every content-based detection model. Without a standard, potential bias from each data collection can jeopardize the validity of model performance results of detection models, bringing into question the prestigious statuses of detection models in many academic works.

With this problem in mind, this study will look at the potential biases across fake news datasets, and investigate whether different datasets might influence the performance of different model structures. The focus of this study will be on supervised news content models that examine news text (mostly news body). Specifically, I am to examine how the differences in dataset sources affect the performance of general manual features like BOW, readability, grammar, and psychological cues. If a detection feature performs differently across available datasets, then I can say that these datasets differ in the quality this feature measures to various degrees. For example, when a readability feature performs well on a dataset, then more disparity in reading ease is observed between the set of fake news and the set of real news. The fake news pieces are likely more readable than real news pieces [17]. Nevertheless, when the fake and real news pieces are less differentiable by their reading ease, the readability feature will yield a worse performance and might eventually become a redundant part of the model. Hence, some detection models might be performing well because they

had the right features to represent the disparity of fake/real news in the dataset. The model is then overfitting to the selected dataset instead of the general "fake news" phenomenon.

As mentioned before, there is little to no "standard" on the construction of fake news datasets. Researchers usually emphasize the size and extensiveness of news data instead of paying attention to the quality of news sources, in particular, that of fake news pieces. While real news articles are chosen from a range of famed sources like Reuters and New York Times (NYT) [3, 2], fake news articles are extracted from limited or more obscure ones. Buzzfeed fake news set [42] has been a choice of many researchers as their fake set, but it does have questionable parallelizability with the chosen real news sources: real NYT articles might have little possibility of appearing at the same time and on the same platforms with the fake articles. As a result, this study wants to confirm if the gaps between fake and real news vary across published papers. In this research, I ask the question: Do fake news datasets share the same set of effective predictive features? I specifically concern the performance of text-based manual detection features and hypothesize that fake news datasets from published papers vary in performance when run under the same text-based predictive features.

2 Literature Review

2.1 Definition and Characteristics of Fake News

The definition of fake news has not yet been fully agreed upon. Among journalists, "fake news" commonly refers to false news [55]. However, false news can be many things, including satire news [38], rumor [57] and disinformation [22]. Earlier work

tends to miss a formal definition of the exact kind of fake news the model are classifying. "Fake news" usually refers to whatever new pieces are deemed false by fact-checking websites like POLITIFACT.COM and SNOPES.COM [17, 33]. Later work places a significant emphasis on elaborating this definition; for example, Gravanis et al. [16] defines fake news as "the online publication of intentionally or knowingly false statements of facts" and Nasir et al. [28] states that fake news are made to "achieve ... fraudulent or illegal [goals]". This paper will use the narrower definition of fake news raised by [55], "Fake news is intentionally false news published by a news outlet." Therefore, I will not consider categories of false news that are not made intentional and truth-bending, like satires, rumors, and misinformation.

Fake news articles are commonly perceived to be sensational false pieces of information that take advantage of the readers' confirmation biases [55] within the "echo chamber" of a certain belief [17]. Hence, to differentiate fake news from real news, researchers have been drawing on deception theories like the Undeutsch hypothesis [45], which states that the style and quality of a person's writing depend on whether he/she has the true experience of the event. They have indeed found stylistic differences such as less logical, technical, and sound arguments in fake news, which demonstrate that fake news articles use heuristics than persuasive arguments [17] and display some extent of extremism [35]. Nevertheless, the characteristics of fake news and the model performances alike depend on the quality of the fake news corpus of interest. As Gravanis et al. [16] has claimed, a large proportion of fake news datasets concentrates on solely one news area (e.g. politics). Such homogeneity results not in detection models for general fake news, but ones for political fake news. The identified relevant characteristics might not apply to the fake news of all areas,

but only to political fake news.

2.2 Previous Generations of Content-based Models

As in the narrower definition, fake news is usually conceived as journalistic lies [22], where the writer knows the truth but intentionally chooses to fabricate the news away from truth. From this conception, early fake news detection models borrowed their structures from linguistic lie studies [17, 34], in the rationale that the fake news classification is just a special case of lie detection. Humans are notoriously bad deception detectors with an average accuracy of 0.54, only slightly better than chance [6]. In fact, trained lie catchers like police officers do not necessarily outperform laymen when posed with deceptive textual messages [47]. Hence, with the rise of computing power, researchers like Fuller et al. [15] developed the first generation of the automated textual deception detection algorithm, which compiles features engineered from past deception research based on psychological deception cues [30, 44] and law enforcement deception transcript data. Following Fuller et al., Afroz et al. [1] explored the problem of stylistic deception, where deception can be identified through the varying linguistic structure between lies and truths [30]. Afroz et al. built a style classifying algorithm featuring writing-style (Writeprints) features [52] (e.g. lexicons/syntax of habit and context), lying-detection features [8] (e.g. stylistic uncertainty) and authorship attribution features [7] (e.g. writing's readability). This model was trained on two datasets, one with professional writers imitating other styles and nonprofessional writers altering their style according to instructions. The Writeprints feature set was found to powerfully predict forced stylistic changes in professional and non-professional writers. With the link between deception and sty-

lometry established, Potthast et al. [35] takes stylistic features further, in context of detection of hyperpartisan fake news. They used n-grams, stopwords, part-of-speech tags, and readability scores as style features, and found that within the scope of U.S. news, hyperpartisan (left or right) articles can be easily identified from balanced news in their writing styles, in that they share common extremism. The uniqueness of hyperpartisan news gave researchers confidence in using stylistic features for fake news detection, not only on political news but also in other categories of news like celebrity news [54].

Along with stylistic interpretation of fake news, the psychological perspective has also generated detection feature ideas. The reality-monitoring theory [20] states that episodic memories (of real events) provide more perceptual, contextual, and affective information than "imagined" memories. Deception researchers have constructed effective criteria for textual lie detection based on this theory [12, 48, 49]: for example, a truthful speech contains more words on affect and space details, and less words on cognitive processes (e.g. "think", "believe"). Later, such criteria were automated by Bond et al. [5] with LIWC dictionaries to form the first algorithmic political lie detector. Although this detector performed with poor accuracy, future fake news research [54] was largely inspired by Bond et al.'s employment of cognitive/perceptive processes in deception detection. In addition, research on satirical news [38] has further shown that text-based style features, readability/complexity features, and psychological features perform well in differentiating intentionally fake news from satirical news and real news.

2.3 Textual Detection Models

Possibly the first formal attempt on fake news detection was conducted by Horne and Adali [17] with the Buzzfeed election news dataset [42] containing fake and real political news from the mainstream, hyperpartisan left, and hyperpartisan right sources. Horne and Adali's study inherited the satire detection feature set from Rubin et al. [38] and performed a three-way comparison between real news, fake news, and satirical news. The combination of stylistic, complexity, and psychology features in Horne and Adali's model [17] shows promising accuracy between fake and real, as well as fake and satire, but does not demonstrate as much predictive power between satire and fake news. In the same year, Pérez-Rosas et al. [33] expanded on previous work by selecting features from models of both Rubin et al. [38] and Potthast et al. [35]. While keeping the readability (i.e. complexity) and psychology (LIWC) features, the findings in Pérez-Rosas et al. [33] emphasize the n-gram feature [35] and the types of punctuation used. Importantly, Pérez-Rosas et al. [33] observed that existing datasets are concentrated on political news, so that they created two new datasets, one built by Amazon Turk workers recreating fake news from legitimate news, covering news from 6 domains, and another collected from real-world celebrity news. The resulting model performed with an accuracy of 0.74 on the crowd-sourced dataset and 0.73 on the celebrity dataset. Although the results do not look as promising as that of the detection models soon to come, Pérez-Rosas et al. [33] pointed out the monotony of popular fake news datasets that centers on the 2016 Presidential Election. The subsequent works in the field of fake news detection hence started to pay close attention to the diversity in their news dataset.

After the many detection models burgeoned in the light of the fake news problem

of the 2016 Presidential Election, Gravanis et al. [16] conducted an evaluation of fake news features [8, 30, 53] on their more-rounded dataset consisting of individually fact-checked articles from a variety of news categories. They found that the features that stood out in predictive power were, first of all, complexity-related features (e.g. # of syllable/word/difficult- word/sentence, Flesch-Kincaid grade level). First-/third-person narrative and affect-related terms follow in predictive power. Importantly, Gravanis et al. [16] tested their model on several existing datasets and found that their model performs differently depending on how the dataset was built. If two datasets were created similarly, their top predictive features are similar. As the field of fake news detection matured, Zhou et al. [54] developed the top-of-the-line detection model with the most comprehensive set of textual content features so far, with two even larger and more carefully made datasets covering politics and celebrity news. The model considers 4 levels of analysis, lexicon (e.g. n-gram), syntax (e.g. POS), semantics (e.g. complexity, affect), and discourse (e.g. rhetorical style). Particularly, the authors dissect the rich semantics level into two feature groups, one clickbait-related (i.e increase click-through rate) and one disinformation-related (i.e. deceptive styles). The accuracy of the lexicon + syntax + semantics model was the highest (0.892/0.879 in the political and the celebrity datasets respectively), and the disinformation features (0.729/0.667) performed better than clickbait-related ones (0.604/0.638).

All attempts of fake news detection employ a binary classifier, for example support vector machines (SVM) [38, 17, 33, 16], random forests (RF) [35, 54] and boosted trees [10, 54] (e.g. XGBoost [9]) . More recently, as deep learning grows in popularity, more researchers started moving the fake news detection problem into a neural network

(NN) model [28]. NN models skip the feature engineering process and take word embeddings of news text as input directly, and thus becoming less interpretable. Since this paper specifically studies the feature's behavior under varied datasets, we will not discuss NN fake news detection models.

2.3.1 Available Datasets

The fake news crisis in the 2016 Presidential Election provided rich building materials for fake news classification datasets. The Buzzfeed team [42] was the first to compile a fake news dataset from political news on Facebook. They first collected election-related articles with high user engagement and then classified the ones from known fake news sources as "fake", and ones from well-known and trusted sources as "real". To tailor this dataset toward fake news classification, Horne and Adali [17] filtered out the satire news and the highly opinionated articles (71 instances out of the original 120). However, as Horne and Adali [17] pointed out, this Buzzfeed dataset might be affected by hidden selection bias (e.g. only high-reading-volume articles were selected) and was solely concentrated on political news. Horne and Adali [17] thus created an additional dataset of political news (50 instances) collected directly from news website, so that the bias of social media volume can be neglected. At the same time, the KDNugget author George McIntire [24] created a larger dataset (6310 instances) by combining Kaggle's collection of election news[1] [37] (fake only, sourced by the BS Detector[2]) [40] and scraped his real news from ALLSIDES.COM in the same election period from reputable sources like New York Times, Wall Street Journal, and The Guardian. A shortcoming of McIntire's dataset was later pointed out by Gravanis et

[1]https://www.kaggle.com/mrisdal/fake-news
[2]https://github.com/selfagency/bs-detector

al. [16] that its fake set was not fact-checked by humans; instead, the news articles are traced from questionable sources that have a history of producing fake news. However, the mere size of this dataset makes it valuable for the diversity of political news it has.

Apart from collecting straight from online sources, Pérez-Rosas et al. [33] were creative in their method that they employed crowdsourcing (AmazonTurk) to create their fake news set (480 instances). The Turkers were instructed to write their fake news according to the real news article given as journalistically as possible. Although this dataset does not fully resemble real-world journalism, Turker's attempt to deceive in writing while mimicking reporters might inform the study about fake news writers' general behavior and psychology. For topical variety and more naturally occurring fake news, Pérez-Rosas et al. [33] built another dataset (200 instances) from celebrity news on the fact-checking website GOSSIPCOP.COM. Ahmed et al. [3] took it further in the collection of real and fake news from POITIFACT.COM and built a mega-sized dataset of 21,417 real pieces and 23,481 fake pieces of world news as well as U.S. political news, named *ISOT*. This is the largest fake news dataset available so far.

In an attempt to enrich dimensions of fake news datasets, Shu et al. [41] integrated news articles with Twitter's social engagement data (e.g. retweets, replies, spatio-temporal information) to form their new dataset, *FakeNewsNet*. Articles are extracted from fact-checking websites POITIFACT.COM and GOSSIPCOP.COM and their related tweets are linked with the article's URL. In terms of articles with text, there are 948 instances of political news and 21,641 instances of celebrity news. Importantly, the model of Zhou et al. [54] was trained and tested on the political news of *FakeNewsNet*, so that the quality of this dataset decides the validity of Zhou et al.'s

model performance. Recently, more international news sources are welcomed to the field of fake news detection. Salem et al.'s [39] dataset *FA-KES* (802 instances), a collection of English news on the Syrian War, has been readily adopted by Elhadad et al. [11] and Nasir et al. [28]. In addition to the mentioned datasets, the *LIAR* dataset by Wang [50], the *CREDBANK* dataset by Mitra and Gilbert [26], and the *NELA-GT* dataset by Norregaard et al. [32] are reputable datasets that contributed significantly to fake news studies or deception studies overall. Unfortunately, these datasets either are not binary (fake vs. real) or do not stick closely to the definition of fake news and therefore they will not be compared in this study.

Until now, we have seen fake news datasets extracted either (1) by the reputation of sources or (2) directly from fact-checking websites. These datasets have many news topics, but they are predominantly political. Most news articles concern U.S. affairs, in particular U.S. politics. Such a collection does cover the centerpiece of the fake news crisis, the election, but it does not include every kind of fake news available. When fake news detection models are tested on these datasets, the model performances can only be guaranteed within the dataset or news articles similar to ones in the dataset. Once the testing dataset becomes biased or skewed, the detection study which employed the dataset becomes questionable in its conclusions (e.g. top features). The proposed features that work perfectly in one specific dataset might not work as well in other datasets of different construction. Therefore, this study need to compare model performances between differently constructed datasets, specifically the feature set performances (e.g. complexity features, psychology features), as the detection model shares different structures but similar components. If a feature set works consistently across the different datasets, then the characteristic of this feature

set qualifies as a quality for fake news in general. However, if the feature set has inconsistent performances, then this study need to reconsider if it only concludes the qualities of a portion of fake news.

Because of the variety of dataset construction observed earlier, this study hypothesizes that *common feature sets across detection models work differently when employed on different datasets*. If this hypothesis is supported, then some detection models, that claim to classify fake news in general, only account for a certain set of fake news (e.g. political news). The authors of the models might need to use more caution when presenting the performance of their model. In addition, some may suggest that all fake news can be classified by a comprehensive enough model that contains all features available to use. Nevertheless, this claim is unpractical, since a long list of features will only increase the algorithmic complexity of a fake news detection model. The computational efficiency of the model will be compromised, undermining the model's value in real-world practice. Generally speaking, this study evaluates two aspects of fake news detection models: whether the models understands "fake news" as intentional disinformation instead of politically charged false news, and whether the models detect news veracity in genres other than that of the model's training dataset. In order to be put into real practice (e.g., social media platforms), good fake news detection models need to "survive" in the diverse pool of real-world fake news. As fake news continued into the current age of COVID-19, many lives can be taken by false information on healthcare and disease treatments. Fake news detection has therefore become ever more important than in the 2016 Presidential Election. This study hence aims to help develop better detection models by examining past fake news detection practices, in particular the ability for popular features to generalize

across news categories and the competence for the datasets to represent the "fake news" phenomenon.

3 Method

This study aims to examine if fake news detection feature sets perform consistently across various datasets. Nine existing datasets were selected to test on four groups of features (Bag-of-words, Readability, Shallow Syntax, and Psychological Cues). I first compare the prediction performances of feature sets on different train and test datasets pairs (a cross-comparison). Then, I extract the top features, to explore deeper trends in performance consistency/inconsistency across datasets. Finally, focusing on readability features, I perform an independent two-sample t-test between its fake and real instances, to inspect the reasons for potential inconsistency in feature performance.

3.1 Data

This study intended to compare all fake news dataset used in the previous literature of detection models. However, only nine were readily available and successfully extracted. These includes *FakeNewsNet* [41] (2 datasets included), *FA-KES* [39], *ISOT* [3], the processed Buzzfeed dataset [42] and the "random political news" dataset by Horne and Adali [17], the two datasets by Pérez-Rosas et al. [33], and the dataset by McIntire [24]. Table 1 is a summary of the nine datasets.

FakeNewsNet *FakeNewsNet* [41] is the first dataset I would like to consider because it was used in Zhou et al.'s textual model [54], a top-notch textual fake news detection model. The dataset covers two areas of news that are heavy with fake news problems, political news, and celebrity news. The political dataset is crawled from POLITIFACT.COM and the celebrity dataset from GOSSIPCOP.COM. This dataset was designed to cover the social context information (e.g. retweet rate) of the article along with its content, but this study will only take out the textual content from both datasets. *FakeNewsNet* included 528 real articles and 420 fake articles in the political set, and 16,694 real articles and 4,947 fake articles in the celebrity set. However, I were only able to retrieve 523/385 from the political set and 13,265/4,185 from the celebrity set (after data processing) because many news articles were removed or re-addressed. (Given the data size, *FakeNewsNet* provided a data retriever [21] to collect articles directly from stored links. Unfortunately, some of the links were expired.)

Horne and Adali Horne and Adali [17] employs two datasets[3] in their study. The first dataset was inherited from Craig Silverman's fake election news dataset [42] collected by BuzzSumo[4], a search engine for news stories with high Facebook engagement. Silverman filtered these high engagement news articles then by their sources. If the source is known to produce fake news, the article is then considered fake news; if the source is reputable and well-known, then the article is considered real news. 36 real articles and 35 fake articles were retrieved in Horne and Adali's discretion (expanded to 53/48 to the date, after data processing). Considering that Silverman did not fact-check his news, Horne and Adali collected 75 real articles and

[3]https://github.com/BenjaminDHorne/fakenewsdata1
[4]https://buzzsumo.com/

75 fake articles (75/75 after data processing). The real articles come from a list of trusted sources [13] and the fake from a list of problematic sources [56], with fake articles fact-checked by websites like SNOPES.COM. The dataset is named *Random Political News*.

Pérez-Rosas et al. Pérez-Rosas et al. [33] created two datasets for their paper. The first one concerns news (predominantly news in the U.S.) of six domains, including "sports, business, entertainment, politics, technology, and education." The authors first collected legitimate news from reputable sources like ABCNews, USAToday, and NewYorkTimes and then created the fake news articles by crowdsourcing (Amazon Mechanical Turk (AMT)). They trained AMT workers to mimic a journalistic style and re-write the legitimate news into a fake one, maintaining topic and length. 240 pieces of real news and 240 pieces of fake news are included in this dataset (240/240 after data processing), referred to as *FakeNewsAMT*. The second dataset consists of real-world celebrity news that comes in pairs, one real and one fake, both on the same topic. Each piece of news was fact-checked using GOSSIPCOP.COM, a total of 100 real news articles and 100 fake news articles were collected. This dataset is referred to as *Celebrity*. The authors made sure that the length and writing style of news pieces are consistent in each dataset, as well as their time frame and delivery purpose. In 2018, the authors updated their *Celebrity* dataset to 250 real and 250 fake news articles (250/250 after data processing). This study will make use of this larger dataset. [5]

[5]https://lit.eecs.umich.edu/downloads.html#Fake News

McIntire Although from a non-academic source, the dataset by McIntire [24] has been adopted by many fake news studies due to its size. [36, 31, 18] The dataset[6] consists of 3,171 real articles and 3,164 fake articles (3,070/2,989 after data processing), where the real articles are scraped from reputable sources like New York Times, Wall Street Journal, and Bloomberg and the fake articles borrowed from Kaggle's fake news dataset [37], all within the 2016 election period.

ISOT Similar to McIntire's dataset, Ahmed et al. [3] collected 12,600 real articles from REUTERS.COM and 12,600 fake articles from a fake news dataset on Kaggle (source not mentioned), which are collected from lists of known fake news sources on POLITIFACT.COM and Wikipedia, all articles are from the same timeline around 2016 and are longer than 200 characters. To the date, the *ISOT* dataset[7] expanded to 21,417 real articles and 23481 fake articles (21,190/17,453 after data processing). Although both McIntire's [24] and Ahmed et al.'s [3] datasets are not strictly fact-checked and matched in topics, their great sizes are valuable for this study.

FA-KES Observing that most fake news datasets center on U.S. news, Salem et al. [39] constructed a dataset[8] that concentrates on the Syrian War. With the help of the Syrian Violation Documentation Center (VDC), an independently funded website that keeps accurate records of conflicts and casualties, the authors were able to fact-check news articles on the Syrian war. Specifically, the authors extracted high-casualty events from VDC and then scraped down related news articles from various middle-east news outlets, representing several party factions in the war. There are a

[6]https://github.com/lutzhamel/fake-news
[7]https://www.uvic.ca/ecs/ece/isot/datasets/fake-news/index.php
[8]https://doi.org/10.5281/zenodo.2607278

Name	Paper	Year	Category	Source	#Real	#Fake	#Total
FakeNewsNet_PolitiFact	[41]	2018	Politics	PolitiFact.com	528	385	913
FakeNewsNet_GossipCop	[41]	2018	Celebrity	GossipCop.com	13,265	4,185	17,450
Horne_Buzzfeed	[17]	2016	Politics	[42]	53	48	101
Horne_RandomPoliNews	[17]	2017	Politics	SNOPES.COM	75	75	150
Perez-Rosas_AMT	[34]	2017	Various	Amazon Turk Workers	240	240	480
Perez-Rosas_Celebrity	[34]	2017	Celebrity	GossipCop.com	250	250	500
McIntire	[24]	2017	Politics	[37] & NYT, etc.	3,070	2,989	6059
ISOT	[3]	2018	Politics	Kaggle & Reuters	21,190	17,453	38,463
FA-KES	[39]	2019	War	VDC	418	371	789

Table 1: A summary of datasets used in this study

total of 426 real articles and 378 fake articles (418/371 after data processing). It is worth noted that *FA-KES* shares a different sense of "fake news" comparing to other datasets. The event reported on fake articles is not necessarily fabricated; instead, they might simply possess a different date or location of the event, or mention a different kind of weapons used. These details, however, can be used by news agencies to blame the wrong party for the conflict. Therefore, *FA-KES* provides an opportunity for fake news detection models to evaluate non-characteristic fake news and expand their use outside of peaceful times.

3.1.1 Data Processing

From each dataset, I retrieve the label (real/fake) and the article's body text content. News titles are not studied here because they have to do more with sensationalism and not deceptive/disinformation styles. The entries with empty and duplicated body text are first removed from each of the nine dataset. For each remaining entry, non-text characters are removed from text using regular expression. The cleaned texts

are then tokenized[9] into a list of lower-case words, and rejoined into one string for algorithmic simplicity. If necessary, the veracity label is re-indexed into 0 and 1, 0 for real news and 1 for fake news. The resulting cleaned dataset is saved into *.csv* form for further usage.

3.2 Feature Sets

The following four kinds of features have been reoccurring throughout the literature on textual fake news models. Hence, they are used as four dimensions to examine if the nine datasets above support the same detection model performances. The following feature sets will each be constructed as a detection model on their own, following the structure used by Zhou et al. [54].

Standardized Bag-Of-Words (BOW) Model BOW model concerns fake news detection on a lexicon level by capturing the absolute frequency of words within a news article. However, given the varying length of text throughout the 9 datasets, I will use the relative frequency of words within an article, which is the absolute frequency standardized by the total number of words within the article. The term frequency-inverse document frequency (TFIDF) method is not employed here because fake news articles might differ from real articles in their use of stopwords (e.g. "its," "for," "that"). The feature consists of *Scikit-Learn*'s *TfidfVectorizer* (with *use_idf* parameter set to *false*) along with a standard scaler.[10]

[9] I used the *word_tokenize* function from the package Lucem Illud. https://github.com/Computational-Content-Analysis-2020/lucem_illud_2020

[10] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

Shallow Syntax Model (POS) Rising from lexicon to syntax, the part-of-speech (POS) tags [46] were used in models of both Pérez-Rosas et al. [33] and Zhou et al. [54]. Part-of-speech tags, as the name suggests, mark the grammatical role of a word in a sentence (e.g. nouns, pronouns, verbs), and in our algorithm, the POS marking is achieved by the *pos_tag* function in the *nltk* package.[11] I then transfer the count of distinct POS tags in one body text into a scaled feature through *Scikit-Learn*'s *DictVectorizer*[12].

Readability Scores Model Semantic interpretation of fake news has always been popular and effective. In particular, the complexity or readability (reading ease) of fake news is shown to differ from legitimate news [17, 16]. Zhou et al. [54] applied readability features in their attempt to study the relationship between fake news and click-baits. In this study, I recruited the Flesch Reading Ease Index (FREI), Flesch-Kincaid Grade Level (FKGL), Automated Readability Index (ARI), Gunning Fox Index (GFI), Coleman-Liau Index (CLI), as well as the total number of words, syllables, polysyllables (syllable ≥ 3), characters and long/difficult words present in one body text as readability features (scaled), achieved through the Python *Readability* package.[13]

Psychological Cues Model (LIWC) Psychological cues are another kind of semantic-based feature frequently employed in the literature [35, 33, 54]. The marking of these cues is usually done through the LIWC Language Tool [44]. The tool counts the number of, for example, positive emotion words, and then I turn the count

[11] https://www.nltk.org/
[12] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.DictVectorizer.html
[13] https://github.com/andreasvc/readability/

of cue words into a list of scaled counts with *DictVectorizer* as in the POS feature. To consider as many psychological aspects of fake news as possible, this study employs all semantic features Zhou et al. [54] achieved through LIWC. This includes the number of affect-related words like positive words, negative words, anxiety words, anger words, and sadness words, informality-related words like swear words, netspeak, assent, nonfluencies, and fillers, as well as cognition-/perception-related words on insight, causation, discrepancy, tentative, certainty, differentiation, see, hear, and feel.

3.2.1 Classifier

This study hires XGBoost Classifier[14] [9] as binary classifier because of its superb performance in Zhou et al. [54], with all parameters set to default value as Zhou et al. did not mention any specification.

3.3 Analysis

Performance Evaluation Since I am studying how feature performance changes by different datasets, our analysis starts with a dataset × model performance evaluation. Each of the nine datasets will be used as the train/test set (0.8/0.2) for each of the four feature models. The test accuracy, precision, recall, and the F1 score, as well as 5-time cross-validated training accuracy, will then be derived as the evaluation metric. The precision and recall curve will also be drawn for each dataset × model instance. In addition, I will test each fake news classifier trained on one of the nine dataset with the other eight datasets, to get a cross-comparison table of test accuracy

[14]https://xgboost.readthedocs.io/en/latest/python/python_api.html

for generalizability conclusions.

Top Features Analysis To examine whether a feature model works similarly in different datasets, this study conducts a top feature (10) analysis for every dataset $\times$ model instance. Under each feature model, I compare the top features of the nine 9 datasets. If the datasets are of similar natures, similar rankings might be observed for the top features. If the top features vary drastically across datasets, this study can then conclude that these datasets are constructed with very different content and topics.

Deceptive Style Comparison Finally, this study would like to attempt operationalizing the difference between fake articles and real articles within a dataset. If a feature model performs differently for our datasets, then this "gap" between real and fake has to be conceptually different to some extent for each of the 9 datasets. Since the readability feature model makes use of reading ease metrics that come in single scalar values, one metric can be picked from the model, and the resulting scalar values can be put into an independent two-sample t-test with pooled variance on the fake and the real set. This study can then gain an idea of "difference" present between them. Unfortunately, other feature models make use of count features, which cannot be straightforwardly compared to the readability feature. Hence the t-tests will only be administered with the readability features. (Real News − Fake News)

	Test Accuracy	Training Accuracy	Precision	Recall	F1 Score
ISOT	**0.989**	**0.989**	**0.994**	0.981	**0.988**
McIntire	0.920	0.922	0.920	0.921	0.920
Horne_RandomPoliticalNews	0.867	0.808	0.818	**1.000**	0.900
FakeNewsNet_Politifact	0.852	0.854	0.836	0.803	0.819
FakeNewsNet_GossipCop	0.841	0.851	0.797	0.479	0.598
Perez-Rosas_CelebrityNews	0.670	0.738	0.685	0.698	0.692
Perez-Rosas_AMT	0.635	0.654	0.633	0.745	0.685
FA-KES	0.513	0.512	0.614	0.531	0.570
Horne_Buzzfeed	0.476	0.825	0.600	0.462	0.522

Table 2: Single-dataset performance summary of bag-of-words (BOW) model

4 Result

4.1 Data×Model Performance Results

This section displays the performance evaluation results of the dataset differences upon feature models. This study has fake news coded as 1 and real news coded as 0. The model's accuracy implies how many times out of all the model identifies fake news as fake news, while the precision score and recall score refer to the number of correctly identified fake news overall fake news predictions by the model, and the number of correctly identified fake news overall fake news present in the dataset, respectively.

Bag-Of-Words (BOW) Feature From the 9 run-through (Table 2, Appendix B. Figure 1), I found that the BOW model performs very well in *ISOT* [3] and McIntire's dataset [24] , both in terms of accuracy and F1 score. The ≥ 0.9 test accuracy indicates that the model is performing at a very high level within the field of fake news detection. The ≥ 0.9 precision, recall, F1 score of the run with *ISOT*

and McIntire's dataset show that the BOW model is very powerful in separating fake news from real news, without many false negatives and false positives. While in *FakeNewsNet*'s (*FNN*) GOSSIPCOP dataset [41], the 0.598 F1 score does not match the quite decent test accuracy of 0.841. Particularly, the *FNN*'s GOSSIPCOP dataset has a recall rate of 0.479. Such a low recall rate indicates that too few fake news articles were identified as fake in the BOW algorithm, and indeed this dataset has a *fake: real* ratio of about 1 : 3. In addition, as shown in graph (Appendix B. Figure 1), the BOW models run in *FA-KES* [39], Horne and Adali's Buzzfeed dataset [17], and Pérez-Rosas et al.'s crowdsourced dataset [34] do not differentiate themselves much from the no-skill classifiers in performance. In particular, the BOW model run in Horne and Adali's Buzzfeed set does not even have a test accuracy higher than by-chance (0.5).

In the cross-comparison table of the BOW model (Table 3), a different story is shown where the *ISOT*-trained model performs less effectively in prediction. While the *ISOT*-trained model keeps test accuracy of 0.773 when predicting the Buzzfeed dataset, it yields a worse-than-chance accuracy in *FA-KES* (0.470) and the two *FakeNewsNet* dataset (0.450/0.309). Thus, despite its high self-test accuracy, the *ISOT*-trained BOW model lacks generalizability in classifying different fake news datasets. The model trained by McIntire's dataset, second-best in self-test accuracy, holds better generalizability. The only under-chance (< 0.5) test accuracy is found in predicting the *FNN*'s GossipCop dataset. Interestingly, *FNN*'s GossipCop dataset and Pérez-Rosas et al.'s *Celebrity* dataset are quite effective in predicting each other with their trained models. The *FNN*'s GossipCop model predicts *Celebrity* with an accuracy of 0.736, shortly after its self-test accuracy of 0.841. In the other direction,

	Train Data				
Test Data	FNN_Politifact	FNN_GossipCop	FA-KES	ISOT	McIntire
FNN_Politifact	**0.852**	0.574	0.488	0.450	0.612
FNN_GossipCop	0.347	**0.841**	0.519	0.309	0.363
FA-KES	0.527	0.474	**0.513**	0.470	0.537
ISOT	0.573	0.551	0.530	**0.989**	0.745
McIntire	0.628	0.434	0.434	0.640	**0.920**
Horne_Buzzfeed	0.584	0.455	0.356	0.574	0.663
Horne_Random	0.547	0.387	0.487	0.773	0.687
Perez-Rosas_AMT	0.548	0.512	0.492	0.619	0.571
Perez-Rosas_Celebrity	0.520	0.736	0.488	0.546	0.582

	Train Data			
Test Data	Horne_Buzzfeed	Horne_Random	Perez-Rosas_AMT	Perez-Rosas_Celebrity
FNN_Politifact	0.573	0.463	0.569	0.576
FNN_GossipCop	0.378	0.281	0.507	0.754
FA-KES	0.515	0.518	0.466	0.468
ISOT	0.439	0.651	0.752	0.627
McIntire	0.635	0.617	0.557	0.572
Horne_Buzzfeed	**0.476**	0.545	0.386	0.574
Horne_Random	0.620	**0.867**	0.573	0.560
Perez-Rosas_AMT	0.490	0.506	**0.635**	0.560
Perez-Rosas_Celebrity	0.532	0.512	0.560	**0.670**

Table 3: Cross comparison: accuracy result of single-dataset-trained BOW classifiers tested on other datasets

the *Celebrity* model predicts *FNN*'s GossipCop set with an accuracy of 0.754, even higher than its self-test accuracy of 0.67. Such good two-way prediction reveals that the BOW model performs with the best generalizability when the test dataset shared the same topic as the training dataset (e.g. celebrity). Among datasets of different topics, however, the model's performance worsens because the fake and real articles differ in the choice of the word quite a lot; that is, fake articles or real articles in this dataset consistently use a set of words that is less present in the other kind of articles. Hence, as Table 3 demonstrates, the BOW model has less generalizability when varying topics occurs in the test set. The homogeneity of news topics in the training set can therefore jeopardize the performance of the BOW feature in fake news detection models.

Looking into the top features of the BOW model in each dataset (Appendix A. Table 11), for *ISOT*, words like "care," "law," and "tweeted" are the stronger predictors that points to fake news articles, given positive features contribute to a prediction of 1 (fake) instead of 0 (real). McIntire's dataset also has "policy" and "nonsense" to be its strong predictors for fake news. Among these words, "nonsense" likely points to unwarranted speech while "tweeted" points to unwarranted sources since fake news particle can easily accuse political actors of declaring non-existent statements. As for the two celebrity news trained models, the feature "herself" stands out from the *FNN* Gossipcop model and "she wants," "she was," and "like she" stands out from the *Celebrity* set of Pérez-Rosas et al.. All of the four features possess positive coefficients, indicating that celebrity fake news tends to mention more about female celebrities. Nevertheless, both celebrity datasets were sourced from GOSSIPCOP.COM, so that this female-oriented characteristic might be special for news from GOSSIPCOP.COM

	Test_Accuracy	Training_Accuracy	Precision	Recall	F1_Score
ISOT	**0.884**	**0.886**	**0.897**	**0.839**	**0.867**
McIntire	0.799	0.793	0.795	0.808	0.801
FakeNewsNet_GossipCop	0.773	0.775	0.587	0.267	0.367
Horne_RandomPoliticalNews	0.767	0.783	0.867	0.722	0.788
FakeNewsNet_Politifact	0.736	0.826	0.689	0.671	0.680
Horne_Buzzfeed	0.667	0.800	0.714	0.769	0.741
Perez-Rosas_Celebrity	0.610	0.635	0.625	0.660	0.642
Perez-Rosas_AMT	0.521	0.591	0.549	0.549	0.549
FA-KES	0.399	0.477	0.506	0.438	0.469

Table 4: Single-dataset performance summary of shallow syntax model (POS)

instead of celebrity news in general.

Shallow Syntax Feature Table 4 and Appendix B. Figure 2 shows performance results of the shallow syntax (i.e. part-of-speech (POS) tags model) by dataset. As with the BOW Model, the POS model continues to perform the best with the *ISOT* [3] and McIntire's dataset [25]. However, the accuracy of the POS model on these two datasets is slightly lower than that of the BOW models. Compared to the BOW model, the POS model performs much better under the Horne and Adali dataset [17], with test accuracy increased from 0.476 to 0.667. However, for the *FA-KES* [39] dataset, the POS model adapts poorly, with only a 0.399 accuracy and a 0.469 F1 score. From the precision-recall plot (Appendix B. Figure 2), we see that only in *FA-KES* the POS classifier is performing worse than the no-skill classifier. Meanwhile, the two datasets of Pérez-Rosas et al. [34] show poorer performance in the syntax model. Again, the model recall rate is low under the *FNN*'s GossipCop dataset [41] because of the uneven fake: real ratio, sabotaging the model's F1 score under this dataset.

	Train Data				
Test Data	FNN_PolitiFact	FNN_GossipCop	FA-KES	ISOT	McIntire
FNN_PolitiFact	**0.736**	0.587	0.406	0.417	0.644
FNN_GossipCop	0.450	**0.773**	0.493	0.365	0.435
FA-KES	0.508	0.473	**0.399**	0.464	0.520
ISOT	0.538	0.569	0.490	**0.884**	0.494
McIntire	0.624	0.501	0.456	0.516	**0.799**
Horne_Buzzfeed	0.772	0.535	0.386	0.505	0.683
Horne_Random	0.707	0.507	0.487	0.587	0.773
Perez-Rosas_AMT	0.523	0.521	0.521	0.467	0.502
Perez-Rosas_Celebrity	0.556	0.626	0.504	0.546	0.566

	Train Data			
Test Data	Horne_Buzzfeed	Horne_Random	Perez-Rosas_AMT	Perez-Rosas_Celebrity
FNN_PolitiFact	0.729	0.600	0.627	0.584
FNN_GossipCop	0.466	0.512	0.546	0.622
FA-KES	0.534	0.516	0.526	0.482
ISOT	0.548	0.544	0.428	0.577
McIntire	0.627	0.688	0.528	0.563
Horne_Buzzfeed	**0.667**	0.634	0.604	0.574
Horne_Random	0.647	**0.767**	0.540	0.533
Perez-Rosas_AMT	0.550	0.544	**0.521**	0.529
Perez-Rosas_Celebrity	0.518	0.544	0.552	**0.610**

Table 5: Cross comparison: accuracy results of single-dataset-trained POS classifier on other datasets

In the cross-comparison table (Table 5) of POS models, *FNN*'s GossipCop dataset and Pérez-Rosas et al.'s *Celebrity* dataset continues their "reciprocal" relationship as in the BOW model. Their prediction results of each other remain higher than other datasets. These two datasets also share similar top features (Appendix A. Table 12): negative coefficients for RP (particle), VBP (non-3rd person present verb), WRB (Wh-adverb). Thus, the two celebrity datasets are further confirmed with likeness in their shallow syntactical structures. In particular, the *Celebrity*-trained POS model predicts *FNN*'s GossipCop set (0.622) better than predicting itself (0.610). Such "over-performance" is also seen in the *AMT* set shares the situation with *Celebrity* where *AMT*-trained model predicts every other dataset except from *ISOT* better than predicting its own test set. Given that both of Pérez-Rosas et al.'s datasets are small in size, the trained models of each might be capturing fake news syntactical characteristics that are more representative in larger datasets. The small size limits consistency in news content and thus leads to low test accuracy. The similar "better-than-self" situation can also be observed in the BOW model (Table 3).

Still, the varying accuracy level in Table 2 and 3 requires explanation: why there exists disparity in POS model performance between datasets. The reason might be the different degrees of grammatical style differences between the fake and the real within each dataset. As we have observed before, Pérez-Rosas et al.'s crowdsourced dataset had its fake news modeled after real news. The AMT writers were asked to follow a journalistic style when writing the fake news, and very possibly, they have followed the grammatical style of the original real news article, since the writers (un-professional) might find it easier mimicking the style of the sample news in front of them than making up a distinct journalistic style of their own. Thus, the grammatical

structures of the original news and the mimicked fake news are similar. This mimicking process might have recreated the fake news fabrication process in the real world and preserved the outstanding syntax characteristics common in fake news, hence the over-performance in predicting, for example, the *FNN*'s PolitiFact set.

Additionally, *FA-KES* yields poor performance and generalizability in both self-test and cross-compared accuracy, possibly because the writing style of war news does vary much between fake and real news. War news usually situates in a less ideologically polarized area than political news. While fake political news attempts to provoke or reinforce confirmation biases in a particular population group, fake war news focuses more on mimicking the style of realistic news to reach a more general audience. Fake war news, as Salem et al. [39] highlights, is not far from real war news in content; only the small details are tweaked (e.g. who made the attack, what weapons were used). Political news, at least among our datasets, concentrates its fake articles on hyperpartisan sources. Apart from the few news pranks (e.g. the Obama assassination [29]), political fake news more often stems from an ideological purpose than a financial purpose when compared to the purely financial motivation of celebrity news [27]. Hence, the writing motivated by ideology might possess a very different style from the writing from other motivations (e.g. monetary), resulting in the contrasting result in Table 4 and 5. The top features analysis of the POS model does not provide us with a consistent pattern for the political/non-political split of the fake news dataset. Therefore, we cannot explain this split in detailed feature components.

	Test Accuracy	Training Accuracy	Precision	Recall	F1 Score
Horne_Buzzfeed	**0.810**	**0.838**	0.846	**0.846**	**0.846**
FakeNewsNet_GossipCop	0.753	0.750	0.493	0.079	0.136
FakeNewsNet_Politifact	0.747	0.763	0.697	0.697	0.697
ISOT	0.743	0.755	0.741	0.664	0.700
Horne_RandomPoliNews	0.733	0.667	**0.857**	0.667	0.750
McIntire	0.665	0.652	0.674	0.647	0.660
Perez-Rosas_Celebrity	0.580	0.500	0.628	0.509	0.562
Perez-Rosas_AMT	0.542	0.547	0.566	0.588	0.577
FA-KES	0.468	0.493	0.575	0.479	0.523

Table 6: Single-dataset performance summary of readability model

Readability Feature Unlike the prior two features, the readability feature model (Table 6) performs the best in Horne and Adali's Buzzfeed dataset, while *FA-KES* [39] and Pérez-Rosas et al.'s two datasets [34] continue to stay on the lower end. Particularly (Appendix B. Figure 3), *FA-KES* and Pérez-Rosas et al.'s Celebrity set have their readability model performing worse than the no-skill classifier. The crowdsourced dataset of Pérez-Rosas et al. also positions quite close to the no-skill precision-recall curve. Horne and Adali's set of self-collected political news [17] does well in precision but not so much in the recall.

In the cross-comparison table (Table 7), the Buzzfeed dataset is found with low generalizability, given the POS model trained on this set predicts other datasets with much lower accuracy. The readability model trained on *FNN*'s PolitiFact set [41] predicts the Buzzfeed dataset with decent accuracy of 0.723. Similarly, the *McIntire*-trained [24] model also predicts the Buzzfeed set with an accuracy of 0.743, but these two relationships are not mutual, the Buzzfeed-trained model predicts *FNN*'s PolitiFact set with only 0.594 accuracy and *McIntire*'s set with only 0.635 accuracy. The "reciprocal" relationship does not continue for *FNN*'s GossipCop set and

	Train Data				
Test Data	FNN_PolitiFact	FNN_GossipCop	FA-KES	ISOT	McIntire
FNN_PolitiFact	**0.747**	0.551	0.634	0.412	0.542
FNN_GossipCop	0.424	**0.753**	0.506	0.353	0.365
FA-KES	0.542	0.471	**0.468**	0.483	0.502
ISOT	0.557	0.546	0.466	**0.743**	0.468
McIntire	0.573	0.498	0.485	0.510	**0.665**
Horne_Buzzfeed	0.723	0.495	0.505	0.455	0.743
Horne_Random	0.547	0.480	0.533	0.600	0.633
Perez-Rosas_crowd	0.483	0.488	0.483	0.485	0.508
Perez-Rosas_celebrity	0.546	0.520	0.512	0.522	0.498

	Train Data			
Test Data	Horne_Buzzfeed	Horne_Random	Perez-Rosas_AMT	Perez-Rosas_Celebrity
FNN_PolitiFact	0.594	0.456	0.577	0.537
FNN_GossipCop	0.405	0.325	0.433	0.550
FA-KES	0.515	0.512	0.531	0.484
ISOT	0.486	0.514	0.491	0.571
McIntire	0.635	0.588	0.447	0.521
Horne_Buzzfeed	**0.810**	0.663	0.436	0.624
Horne_Random	0.593	**0.733**	0.507	0.600
Perez-Rosas_crowd	0.494	0.500	**0.542**	0.467
Perez-Rosas_celebrity	0.522	0.514	0.496	**0.580**

Table 7: Cross comparison: accuracy results of single-dataset-trained readability classifier tested on other datasets

Pérez-Rosas et al.'s *Celebrity* set in readability model. Most readability models are predicting other non-train datasets with accuracy around 0.5, implying that this kind of model might not generalize well overall.

The lack of generalizability might be explained by the analysis of the top features (Appendix A. Table 13) of the readability model. Specifically, in Horne and Adali's Buzzfeed dataset the best feature, Flesch Reading Ease Index (FREI), has a negative coefficient. Given that a high FREI score notes easier text and a low FREI score notes professional text, the negative coefficient indicates that the easier the text, the

higher the FREI score, the more likely the article is real (coded as 0). This result does not align with the previous findings that fake news leans toward heuristics (i.e. less cognitive processing) rather than arguments [17]. Similarly, Pérez-Rosas et al.'s Celebrity dataset, *FNN*'s PolitiFact dataset, and *ISOT* [3] have a negative coefficient, while all other datasets have a positive coefficient. Such conflicting results indicate that the readability model is highly tailored for each dataset and that no general readability pattern is observed within the study's nine datasets. In another way, this result might signify that the readability feature model is just not a good model for fake news detection, given the lower average accuracy (compared to previous features) and the muddled prediction pattern.

To further examine the mixed patterns, each dataset received a two-sample independent t-test with pooled variance scored by each of the readability features (Table 8). From Table 8, I found that 6 out of 9 datasets have a negative t-value for FREI, which means that, generally, their real news articles (group 1) score lower than fake news articles on FREI. Such a result confirms the perception that fake news is less complicated to read and understand. However, this study observes that Horne and Adali's Buzzfeed dataset, despite having a negative coefficient in the readability model, has a negative t-value, indicating that overall fake news in this dataset is easier to read than real news. This conflict reveals that the composition of fake news articles might be very polarized. While more than half of fake articles hold better reading ease, the rest might read to be surprisingly sophisticated for fake news. These sophisticated cases might have skewed the classifier so that the model finds fake news articles to be sophisticated overall, and hence the coefficient for FREI becomes negative. (The other readability scores share the same logic: high FKGL/ARI/GFI implies complex

	FREI	FKGL	ARI	GFI	CLI
McIntire	-10.656⋆⋆	10.668⋆⋆	10.664⋆⋆	10.673⋆⋆	-1.432
FakeNewsNet	-8.402⋆⋆	8.398⋆⋆	8.399⋆⋆	8.397⋆⋆	-0.800
Horne_Buzzfeed	-4.477⋆⋆	4.480⋆⋆	4.480⋆⋆	4.482⋆⋆	-0.006
Horne_Random	-3.878⋆⋆	3.836⋆⋆	3.834⋆⋆	3.844⋆⋆	2.787⋆⋆
FNN_GossipCop	-3.174⋆	3.161⋆	3.152⋆	3.162⋆	-0.249
Perez-Rosas_celebrity	-2.526⋆	2.512⋆	2.510⋆	2.509⋆	1.819
FA-KES	0.786	-0.814	-0.820	-0.819	0.983
Perez-Rosas_AMT	1.935	-2.197⋆	-2.232⋆	-2.185⋆	0.802
ISOT	11.742⋆⋆	-13.553⋆⋆	-13.924⋆⋆	-13.440⋆⋆	62.980⋆⋆

	#words	#syllables	#characters	#polysyllables	#difficultwords
McIntire	10.674⋆⋆	10.392⋆⋆	10.539⋆⋆	10.288⋆⋆	9.902⋆⋆
FakeNewsNet_Politifact	8.396⋆⋆	8.348⋆⋆	8.352⋆⋆	8.036⋆⋆	8.061⋆⋆
Horne_Buzzfeed	4.482⋆⋆	4.355⋆⋆	4.392⋆⋆	4.247⋆⋆	4.240⋆⋆
Horne_Random	3.812⋆⋆	4.181⋆⋆	4.164⋆⋆	5.105⋆⋆	4.817⋆⋆
FakeNewsNet_GossipCop	3.153⋆	3.099⋆	3.001⋆	1.775	2.729⋆
Perez-Rosas_celebrity	2.504⋆	2.698⋆	2.657⋆	2.668⋆	3.035⋆
FA-KES	-0.830	-0.629	-0.699	0.132	-0.368
Perez-Rosas_AMT	-2.303⋆	-2.311⋆	-2.311⋆	-1.394	-1.506
ISOT	-14.583⋆⋆	-6.760⋆⋆	-9.036⋆⋆	1.514	6.802⋆⋆

⋆⋆ stands for $p < 0.001$, ⋆ stands for $p < 0.05$

Table 8: Two-sample t-test value of readability features between fake/real news

	Test Accuracy	Training Accuracy	Precision	Recall	F1 Score
ISOT	**0.849**	**0.849**	**0.858**	**0.799**	**0.828**
FakeNewsNet_GossipCop	0.754	0.765	0.519	0.049	0.089
McIntire	0.706	0.703	0.706	0.713	0.709
FakeNewsNet	0.681	0.716	0.629	0.579	0.603
Horne_RandomPoliNews	0.633	0.742	0.684	0.722	0.703
Perez-Rosas_Celebrity	0.600	0.617	0.651	0.528	0.583
Perez-Rosas_AMT	0.542	0.552	0.581	0.490	0.532
Horne_Buzzfeed	0.524	0.600	0.714	0.385	0.500
FA-KES	0.519	0.504	0.609	0.583	0.596

Table 9: Performance Summary of Psychological Cues Model (LIWC)

text and high CLI implies easier text.) Besides the datasets that produce better-performing models, the t-test result also demonstrates the consistency of fake and real articles in content and style for *FA-KES* and Pérez-Rosas et al.'s crowdsourced dataset. As mentioned before, these datasets are constructed to be very similar in topics and style to add challenge to the field of fake news detection. These results further reveal their balancedness. Surprisingly, *ISOT* does not have a significant difference in the number of polysyllables between its fake set and real set. Compared with the significant differences in the number of words/syllables/characters, this relatively low t-value of the number of polysyllables possibly implies that *ISOT*'s real news does not use many long words, but its fake news does more often employ shorter words.

Psychological Cues Feature (LIWC) Finally, the results of the LIWC model conform with that of the BOW and POS model, with *ISOT*'s model [3] as the best-performing one and *FA-KES* [39] as the worst-performing one (Table 9). *FA-KES*, the Pérez-Rosas et al.'s AMT dataset [34] and Horne and Adali's Buzzfeed dataset

[17] continues to perform worse than the no-skill classifier (Appendix B. Figure 4). In the cross-comparison table (Table 10), however, the *ISOT*-trained LIWC model does not generalize well, only the *FNN*'s GossipCop dataset is predicted with an accuracy of 0.76; the prediction of other datasets only yields accuracy around 0.5. On the other direction, the model trained by *FNN*'s GossipCop set does not predict *ISOT* well (0.569 accuracy), thus not establishing a reciprocal relationship seen in earlier models. The celebrity dataset pair shows some reciprocity, with 0.626 accuracy predicting *Celebrity* from *FNN*'s GossipCop model (self-test 0.773) and 0.622 predicting *FNN*'s GossipCop set from the *Celebrity* model (self-test 0.610). Another reciprocal relationship is observed between *FNN*'s PolitiFact set and the Buzzfeed set, where the *FNN*'s PolitiFact model predicts the Buzzfeed set with 0.772 accuracy (self-test 0.736) and the Buzzfeed model predicts the *FNN*'s PolitiFact set with 0.729 accuracy (self-test 0.667). Although not as significant, *FNN*'s PolitiFact and the Buzzfeed set also display reciprocity with Horne and Adali's Random Political News. These four political datasets place predicts each other better (higher accuracy) than the other datasets in the study. Such a "reciprocity bubble" demonstrates that political news datasets are alike in the psychological cues they carry.

According to the top features of the LIWC model (Appendix A. Table 14), informal word types are pervasive in the nine individually trained models. Specifically, the swear word feature gives a positive coefficient in the McIntire dataset, the Buzzfeed set, Horne and Adali's Random Political News set, which implies that political fake news tends to be more swear-infused. In fact, the use of informal language persistently comes to be the top 10 feature among 8 out of 9 datasets, although their $+/-$ sign shifts back and forth. The left-alone Pérez-Rosas et al.'s crowdsourced dataset [34],

	Train Data				
Test Data	FNN_PolitiFact	FNN_GossipCop	FA-KES	ISOT	McIntire
FNN_PolitiFact	**0.681**	0.576	0.437	0.576	0.424
FNN_GossipCop	0.461	**0.754**	0.315	0.760	0.240
FA-KES	0.531	0.470	**0.519**	0.470	0.530
ISOT	0.602	0.548	0.458	**0.849**	0.452
McIntire	0.567	0.493	0.479	0.493	**0.706**
Horne_Buzzfeed	0.495	0.525	0.485	0.525	0.475
Horne_Random	0.633	0.500	0.420	0.500	0.500
Perez-Rosas_AMT	0.498	0.500	0.498	0.500	0.500
Perez-Rosas_Celebrity	0.562	0.500	0.444	0.500	0.500

	Train Data			
Test Data	Horne_Buzzfeed	Horne_Random	Perez-Rosas_AMT	Perez-Rosas_Celebrity
FNN_PolitiFact	0.617	0.576	0.435	0.544
FNN_GossipCop	0.401	0.760	0.480	0.626
FA-KES	0.483	0.470	0.477	0.529
ISOT	0.494	0.548	0.697	0.579
McIntire	0.557	0.493	0.522	0.539
Horne_Buzzfeed	**0.524**	0.525	0.515	0.455
Horne_Random	0.507	**0.633**	0.613	0.613
Perez-Rosas_AMT	0.506	0.500	**0.542**	0.523
Perez-Rosas_Celebrity	0.488	0.500	0.546	**0.600**

Table 10: Cross comparison: accuracy results of single-dataset-trained LIWC model tested on other datasets

however, has the count of fillers words as its best predictor. We know that the fake articles of this dataset are not real-world occurring, but made up by Amazon Turkers. The filler words might be a special strategy for the AMT workers to write their copy of fake news and finish the paid task. The rest of the features, for example, negative emotion words and discrepancy words, appear frequently in the top list but have contrasting directions of prediction across the datasets. Perhaps similarly to the top readability features, the internal structures of these datasets are so complicated that an average of features fails to describe the dataset properly. Unfortunately, because of the vast feature selection in the LIWC model, I could not conduct a t-test study as in the readability model. Further studies will be required to fathom the detailed differences between these datasets.

5 Discussion

In this study, I investigated the internal differences of 9 well-known and readily available binary fake news datasets. Four common stylistic features were constructed to evaluate the differences. This study analyzes the features under the datasets through performance comparisons, top feature analysis, and an independent two-sample t-test in the readability feature. In the cross-comparison of test accuracy (Table 3, 5, 7, 10), the feature models predicts other datasets much worse than the dataset it was trained on, unless the train dataset and the test dataset are from the same news category. This result indicates that datasets vary in content (wording/semantics) and structure (syntax) to a large extent and feature models trained on one dataset can hardly be brought to a more generalized content with multiple news categories. Given

these datasets are retrieved with popular methods in the study of fake news, I can conclude that the observed internal differences exist across most, if not all, fake news datasets available. For example, with the readability feature model, I found that some datasets differ very little in reading ease between their real article set and fake article set (Table 8). *ISOT* [3] even poses a reading ease difference of opposite direction, contrasting most datasets. Thus, I can say that no standardized consistency exists between fake and real news for reading ease in the current field of fake news detection, as well as for grammar and psychologically (e.g. emotionally) charged content. Indeed, fake news might be naturally easier to read according to some, but the study's mixed observation of readability score t-values and feature coefficient says otherwise. Therefore, this study proposes that more attention should be given to fake/news data consistency in the further construction and use of fake news datasets, not only for academic rigor but also for better application to real-world fake news problems. Combining with this study's results, I present the following suggestions for future design and use of fake news datasets.

More variety in the news category should be introduced. As this study has shown, it is hard for fake news models trained on political news to predict the veracity labels of a celebrity news dataset. Different news categories, or even different outlets, can disagree in writing styles quite a lot, not mentioning that different news topics bring out dissimilar collections of commonly used words. To tackle this problem, I suggest that more than one category of news should be introduced when constructing a fake news dataset so that the detection model developed can adapt to a better variety of fake news and find more common traits of fake news across news categories.

Similar topics should be maintained for fake and real news. While employing news from more categories, a dataset should also minimize the topical gap between fake and real articles. The fake articles should strive to report similar events as real articles so that topical words (e.g."Trump", "election", "affair") are consistent across the fake set and the real set. Otherwise, if fake and real articles are taken from on different topics, the detection model would be picking up the news themes instead of the deceptive styles of the fake news. In addition, I find that BOW, POS, and LIWC models are generally stronger among political datasets (Table 2), possibly because political fake news is more often hyperpartisan rather than neutral. In contrast to political news, war news and celebrity news has a consistent style, where celebrity news sticks to exposing privacy and war news sticks to reporting attacks and discussing the effect and responsibility of attacks. Such consistency perhaps contributed to the poor performance of stylistic features in *FA-KES* [39] and Pérez-Rosas et al.'s Celebrity set [34]. Therefore, extra caution is needed when hyperpartisan sources are exploited for fake news datasets, as hyperpartisan journalism focuses on very different topics than neutral political journalism.

Concept of *fake news* should be further specified. Many fake news studies aim to tackle the fake news problem as in the 2016 Presidential Election. Nevertheless, this kind of fake news is more than just intentional false content; it is politically charged and possibly highly partisan. Hyperpartisanship is known to motivate different writing styles [23] so that fake news detection models might very likely be picking up the hyperpartisan styles instead of the wanted deceptive styles. Thus, researchers should examine what writing styles or information their models are detecting exactly.

Indeed, the academic definition of fake news remains fuzzy and requires more research that explores the multifaceted occurrence of fake news. This definitional work can potentially be drawn from a massive field study of naturally occurring fake news. When fake news is properly classified by its type and intentions, better algorithms can be made to further tackle the problem of misinformation overall.

5.1 Limitations

This study only managed to examine the performance of four feature sets under nine datasets, thus giving rise to potential selection bias in feature and datasets. The full landscape of fake news detection studies requires more inclusive research with more features and datasets to consider. Also, I did not recreate the full models the nine datasets were applied in so that this study cannot conclude concretely about the effectiveness of published models in fake news detection. In addition, the analysis methods in this study are quite rudimentary. If possible, more statistical tests should be conducted to explore the differences between datasets and between fake news and real news within one dataset.

References

[1] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *2012 IEEE Symposium on Security and Privacy*, pages 461–475. IEEE, 2012.

[2] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on*

intelligent, secure, and dependable systems in distributed and cloud environments, pages 127–138. Springer, 2017.

[3] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9, 2018.

[4] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554, 2019.

[5] Gary D Bond, Rebecka D Holman, Jamie-Ann L Eggert, Lassiter F Speller, Olivia N Garcia, Sasha C Mejia, Kohlby W Mcinnes, Eleny C Ceniceros, and Rebecca Rustige. 'lyin'ted','crooked hillary', and 'deceptive donald': Language of lies in the 2016 us presidential debates. *Applied Cognitive Psychology*, 31(6):668–677, 2017.

[6] Charles F Bond Jr and Bella M DePaulo. Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234, 2006.

[7] Michael Robert Brennan and Rachel Greenstadt. Practical attacks against authorship recognition techniques. In *Twenty-First IAAI Conference*, 2009.

[8] Judee K Burgoon, J Pete Blair, Tiantian Qin, and Jay F Nunamaker. Detecting deception through linguistic analysis. In *International Conference on Intelligence and Security Informatics*, pages 91–101. Springer, 2003.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[10] Anshika Choudhary and Anuja Arora. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications*, 169:114171, 2021.

[11] Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. A novel approach for selecting hybrid features from online news textual metadata for fake news detection. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, pages 914–925. Springer, 2019.

[12] Stamatis Elntib, Graham F Wagstaff, and Jacqueline M Wheatcroft. The role of account length in detecting deception in written and orally produced autobiographical accounts using reality monitoring. *Journal of Investigative Psychology and Offender Profiling*, 12(2):185–198, 2015.

[13] Pamela Engel. Here are the most-and least-trusted news outlets in america. *Business Insider*, 21, 2014.

[14] Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 171–175, 2012.

[15] Christie M Fuller, David P Biros, and Rick L Wilson. Decision support for determining veracity via linguistic-based cues. *Decision Support Systems*, 46(3):695–703, 2009.

[16] Georgios Gravanis, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213, 2019.

[17] Benjamin Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 2017.

[18] Mayank Kumar Jain, Dinesh Gopalani, Yogesh Kumar Meena, and Rajesh Kumar. Machine learning based fake news detection using linguistic features and word vector features. In *2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pages 1–6. IEEE, 2020.

[19] Christian Janze and Marten Risius. Automatic detection of fake news on social media platforms. In *PACIS*, page 261, 2017.

[20] Marcia K Johnson and Carol L Raye. Reality monitoring. *Psychological review*, 88(1):67, 1981.

[21] KaiDMML. Fakenewsnet: This is a dataset for fake news detection research. Available at https://github.com/KaiDMML/FakeNewsNet.

[22] Nir Kshetri and Jeffrey Voas. The economics of "fake news". *IT Professional*, 19(6):8–12, 2017.

[23] Mihael Liskij, Dominik Prester, and Ivan Šego. Judging a book by its cover: Predicting partisanship using only article titles. *Text Analysis and Retrieval 2019 Course Project Reports*, page 40.

[24] G McIntire. Fake real news dataset. *GeorgeMcIntire's Github*, 2018. Available at https://github.com/lutzhamel/fake-news.

[25] George McIntire. Machine learning finds "fake news" with 88% accuracy, 2017. Available at https://www.kdnuggets.com/2017/04/machine-learning-fake-news-accuracy.html.

[26] Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 2015.

[27] Bhawna Narwal. Fake news in digital media. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 977–981. IEEE, 2018.

[28] Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007, 2021.

[29] Steven Nelson. Ap twitter account reports fake obama assassination attempt, Apr 2013. Available at https://www.usnews.com/news/newsgram/articles/2013/04/23/ap-twitter-account-tweets-fake-obama-assassination-attempt.

[30] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675, 2003.

[31] Che Eembi Normala, Iskandar Jamil, Fatimah Sidi Ishak, and Affendey Lilly Suriani. Fakeheader: A tool to detect deceptive online news based on

misleading news headlines and contents. *Turkish Journal of Computer and Mathematics Education Vol*, 12(3):2217–2223, 2021.

[32] Jeppe Nørregaard, Benjamin D Horne, and Sibel Adalı. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 630–638, 2019.

[33] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

[34] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. *International Conference on Computational Linguistics (COLING)*, 2018.

[35] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

[36] Harita Reddy, Namratha Raj, Manali Gala, and Annappa Basava. Text-mining-based fake news detection using ensemble methods. *International Journal of Automation and Computing*, 17(2):210–221, 2020.

[37] Meg Risdal. Getting real about fake news, Nov 2016. Available at https://www.kaggle.com/mrisdal/fake-news.

[38] Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings*

of the second workshop on computational approaches to deception detection, pages 7–17, 2016.

[39] Fatima K Abu Salem, Roaa Al Feel, Shady Elbassuoni, Mohamad Jaber, and May Farah. Fa-kes: A fake news dataset around the syrian war. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 573–582, 2019.

[40] selfagency. bs-detector, Nov 2016. Available at https://github.com/selfagency/bs-detector.

[41] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and spatial-temporal information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[42] Craig Silverman. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate, Oct 2016. Available at https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis.

[43] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE, 2019.

[44] Yla R Tausczik and James W Pennebaker. The psychological meaning of words:

Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[45] Udo Undeutsch. Beurteilung der glaubhaftigkeit von aussagen. *Handbuch der psychologie*, 11:26–181, 1967.

[46] Atro Voutilainen. Part-of-speech tagging. *The Oxford handbook of computational linguistics*, pages 219–232, 2003.

[47] Aldert Vrij. *Detecting lies and deceit: The psychology of lying and implications for professional practice*. Wiley, 2000.

[48] Aldert Vrij. *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons, 2008.

[49] Aldert Vrij, Lucy Akehurst, Stavroula Soukara, and Ray Bull. Detecting deceit via analyses of verbal and nonverbal behavior in children and adults. *Human communication research*, 30(1):8–41, 2004.

[50] William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

[51] Xichen Zhang and Ali A Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2):102025, 2020.

[52] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. A framework for authorship identification of online messages: Writing-style features and classifi-

cation techniques. *Journal of the American society for information science and technology*, 57(3):378–393, 2006.

[53] Lina Zhou, Judee K Burgoon, Jay F Nunamaker, and Doug Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation*, 13(1):81–106, 2004.

[54] Xinyi Zhou, Atishay Jain, Vir V Phoha, and Reza Zafarani. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2):1–25, 2020.

[55] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

[56] Melissa Zimdars. False, misleading, clickbait-y, and satirical "news" sources. *Google Docs*, 2016.

[57] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.

A Top Feature Analysis

This section presents the top feature analysis tables for each of the four features.

FNN_PolitiFact		McIntire		ISOT	
coef	feature	coef	feature	coef	feature
-1.105	hotline	2.541	how	4.969	are
-0.956	use supported	1.504	000	1.496	need to
0.952	scans	0.905	not	1.458	care
0.739	scene	0.863	type	1.357	000
-0.616	leadership in	0.820	policy in	-1.192	said
-0.602	lexisnexis	0.816	policy of	0.936	law
-0.599	that any	0.747	nonsense	0.931	needs
0.592	not even	0.729	10	0.884	tweeted
-0.579	00 et	0.703	however	0.853	homeland
0.561	for sale	0.686	around	0.821	degree

Horne_Buzzfeed		Horne_RandomPoliNews		Perez-Rosas_AMT	
coef	feature	coef	feature	coef	feature
1.678	000	-1.344	typically	1.031	are
1.326	area	1.180	the clinton	0.928	purchase
1.252	probably get	1.117	in america	0.925	the future
1.244	probably good	1.016	problem with	0.922	the flight
1.218	argue	0.964	problems	0.755	00
1.184	fraud illicit	0.919	archive	-0.733	000
-1.178	000 and	0.908	000 by	0.719	improved
1.163	disputed	-0.907	disagreed	0.692	of several
1.161	free ride	0.905	central	0.690	found it
1.158	of famed	0.869	century	-0.667	four

FA-KES		FNN_GossipCop		Perez-Rosas_Celebrity	
coef	feature	coef	feature	coef	feature
0.797	04	2.273	business	3.614	now
0.753	2016 sana	2.155	000	1.327	career
0.716	killed three	1.900	twitter	1.084	unhappy
0.695	01	1.766	but	1.077	care of
-0.611	2016 10	1.472	days	1.055	she wants
0.585	saying	1.431	know that	0.982	problems
0.585	correspondent	1.386	herself	0.980	problem
-0.583	correspondent added	1.295	day or	0.936	she was
-0.574	fateh and	1.241	move	0.931	like she
-0.572	preparing for	1.180	stuff	0.909	career and

Table 11: Top 10 features in bag-of-words (BOW) model

FNN_PolitiFact		McIntire		ISOT	
coef	feature	coef	feature	coef	feature
-6891040.0	VBZ	17613300.0	EX	1.047300e+09	VBN
-3215250.0	VBG	8083310.0	WP	-7.350670e+08	RB
-2435200.0	JJ	4616930.0	VBP	7.126910e+08	IN
-1955770.0	RBS	1568480.0	JJR	5.356740e+08	VBD
-1123340.0	TO	1559490.0	IN	-4.864810e+08	PRP
1069070.0	VBN	-1086670.0	NN	-3.527530e+08	VBZ
69876.4	WP	-988284.0	TO	-3.270700e+08	DT
61352.5	VBP	-943506.0	CD	-2.154940e+08	VB
-34763.2	CD	-860062.0	VBD	-1.851400e+08	NNS
32754.1	NNP	807263.0	MD	-1.205330e+08	TO

Horne_Buzzfeed		Horne_RandomPoliNews		Perez-Rosas_AMT	
coef	feature	coef	feature	coef	feature
185638.0	WP	853430.0	WRB	0.589	DT
174741.0	VBP	496667.0	VBP	-0.487	CD
-148646.0	NNP	484608.0	NN	0.486	VBN
126189.0	VBZ	443991.0	VB	-0.416	NN
-70692.8	JJR	-396423.0	TO	0.381	MD
54012.2	MD	-277830.0	VBD	-0.337	WDT
-27651.2	NN	-244918.0	CD	-0.320	WP
-27527.9	WDT	-189071.0	MD	-0.234	EX
26383.1	WP$	163559.0	PRP$	-0.214	WRB
25522.1	DT	-158585.0	RB	0.213	TO

FA-KES		FNN_GossipCop		Perez-Rosas_Celebrity	
coef	feature	coef	feature	coef	feature
0.848	NNPS	-315244000.0	RP	-4745570.00	RBR
0.314	PRP	-57114000.0	VBP	-776075.00	PDT
0.309	RB	-51347000.0	PRP$	14956.90	UH
-0.305	VBD	51271600.0	CD	-5414.53	RP
0.250	VBN	38188900.0	EX	-5003.84	WRB
-0.232	CC	-27999000.0	WRB	-4604.77	WP
-0.229	NNS	-25632100.0	IN	-4551.44	CD
0.200	CD	-24593400.0	POS	4488.35	VBD
0.198	DT	-21877500.0	VBZ	-3504.47	VBP
0.194	JJ	20752500.0	TO	-3256.32	FW

Table 12: Top 10 features in shallow syntax model (POS)

FNN_PolitiFact		McIntire		ISOT	
coef	feature	coef	feature	coef	feature
17724.300	polysyllable_count	-1.812170e+07	polysyllable_count	3643500.000	difficult_words
14738.900	difficult_words	-1.535360e+07	difficult_words	2644960.000	polysyllable_count
12660.600	word_count	-9.925120e+06	syllable_count	940752.000	ARI
12117.500	syllable_count	-9.821320e+06	character_count	916214.000	syllable_count
11740.100	character_count	-8.775590e+06	ARI	820943.000	FKGL
10683.000	ARI	-8.255080e+06	word_count	774213.000	character_count
9380.690	FKGL	-6.009450e+06	FKGL	517251.000	word_count
6137.850	GFI	-2.496710e+06	GFI	241085.000	GFI
-1286.600	FREI	3.474830e+05	FREI	-9385.770	FREI
0.935	CLI	-4.788470e+02	CLI	260.518	CLI

Horne_Buzzfeed		Horne_RandomPoliNews		Perez-Rosas_AMT	
coef	feature	coef	feature	coef	feature
-17565.500	FREI	51.591	FREI	-9162.570	polysyllable_count
6814.190	polysyllable_count	-46.666	polysyllable_count	-6783.320	difficult_words
6675.690	difficult_words	-41.256	difficult_words	-1232.320	syllable_count
4785.680	syllable_count	-36.362	syllable_count	-1204.550	ARI
4618.330	character_count	-35.457	character_count	-1107.900	character_count
4109.540	ARI	-34.029	word_count	-1049.770	FKGL
4020.530	word_count	-33.746	ARI	-912.019	word_count
3951.750	FKGL	-32.735	FKGL	-660.085	GFI
3511.320	GFI	-31.308	GFI	41.005	FREI
20.583	CLI	-1.154	CLI	-26.963	CLI

FA-KES		FNN_GossipCop		Perez-Rosas_Celebrity	
coef	feature	coef	feature	coef	feature
11914.500	FREI	-1.247200e+08	FKGL	-115.866	FREI
-5365.830	polysyllable_count	-6.199300e+07	syllable_count	50.565	ARI
-5205.290	character_count	-2.098620e+07	ARI	50.104	word_count
-5110.150	syllable_count	-1.769580e+07	character_count	49.594	FKGL
-4854.770	word_count	-1.492870e+07	word_count	46.002	GFI
-4598.940	ARI	-1.154460e+07	GFI	44.564	character_count
-4542.910	difficult_words	4.096490e+05	FREI	44.374	syllable_count
-4008.420	FKGL	3.668640e+03	polysyllable_count	38.859	polysyllable_count
-2915.100	GFI	2.864330e+03	difficult_words	35.966	difficult_words
-20.428	CLI	-2.020950e+03	CLI	0.526	CLI

Table 13: Top 10 features in readability model

FNN_PolitiFact		McIntire		ISOT	
coef	feature	coef	feature	coef	feature
0.727468	informal	938987.00	swear	-16978500.0	swear
-0.627558	differ	-391528.00	assent	-3165070.0	assent
-0.599308	discrep	-188118.00	affect	-1132380.0	netspeak
0.535900	negemo	99251.20	posemo	791236.0	insight
0.496724	see	-71473.60	percept	131582.0	affect
-0.404556	nonflu	60307.60	differ	-119001.0	posemo
-0.399345	posemo	-34083.90	tentat	-113012.0	informal
-0.385030	netspeak	-18684.10	hear	82111.4	cause
-0.344078	assent	-15146.80	cogproc	51049.5	discrep
0.313370	tentat	8270.45	informal	-46012.0	nonflu

Horne_Buzzfeed		Horne_RandomPoliNews		Perez-Rosas_AMT	
coef	feature	coef	feature	coef	feature
-1.029020	differ	3.367640	swear	-0.826089	filler
-0.624734	anger	1.961420	filler	-0.462564	hear
0.609142	swear	-1.624290	informal	-0.275894	negemo
-0.495870	informal	1.557940	see	0.247163	function
0.420480	insight	1.261010	negemo	0.239871	affect
0.389275	discrep	-1.063960	hear	0.186804	discrep
0.344304	feel	1.052030	certain	0.172454	cogproc
0.324589	negemo	0.777674	netspeak	0.166332	percept
0.311667	netspeak	-0.773018	percept	0.152752	anger
-0.291646	posemo	0.622542	nonflu	-0.116707	cause

FA-KES		FNN_GossipCop		Perez-Rosas_Celebrity	
coef	feature	coef	feature	coef	feature
0.292149	informal	32082000.00	informal	0.442779	negemo
-0.214492	swear	-82176.80	swear	-0.418214	netspeak
-0.203353	nonflu	-22853.20	netspeak	0.404794	informal
0.182484	tentat	-20658.90	negemo	-0.388374	posemo
-0.172925	netspeak	14495.80	assent	-0.315837	hear
-0.150966	percept	6447.35	feel	-0.271712	nonflu
-0.138705	anger	4365.09	hear	0.250321	cause
-0.130189	cause	-4171.63	tentat	0.245630	discrep
-0.116473	discrep	3586.28	anger	0.219763	cogproc
-0.101402	differ	-3174.27	sad	0.190716	anx

Table 14: Top 10 features in psychological cues model (LIWC)

B Precision-Recall Curves

This section presents the precision-recall curves from the self-test accuracy of each dataset for each of the four features.

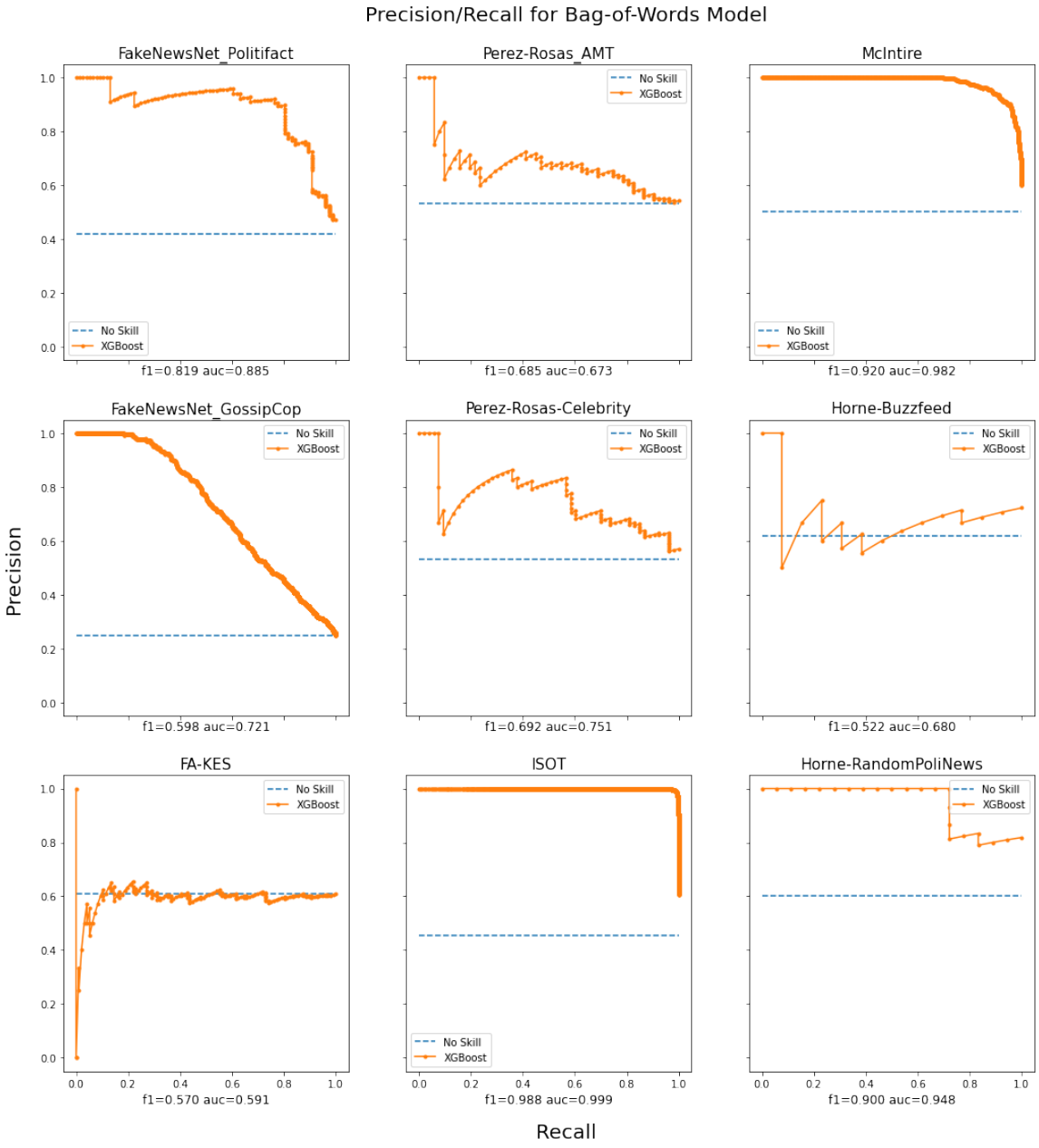

Figure 1: Precision/Recall plot for single-dataset performances on BOW model

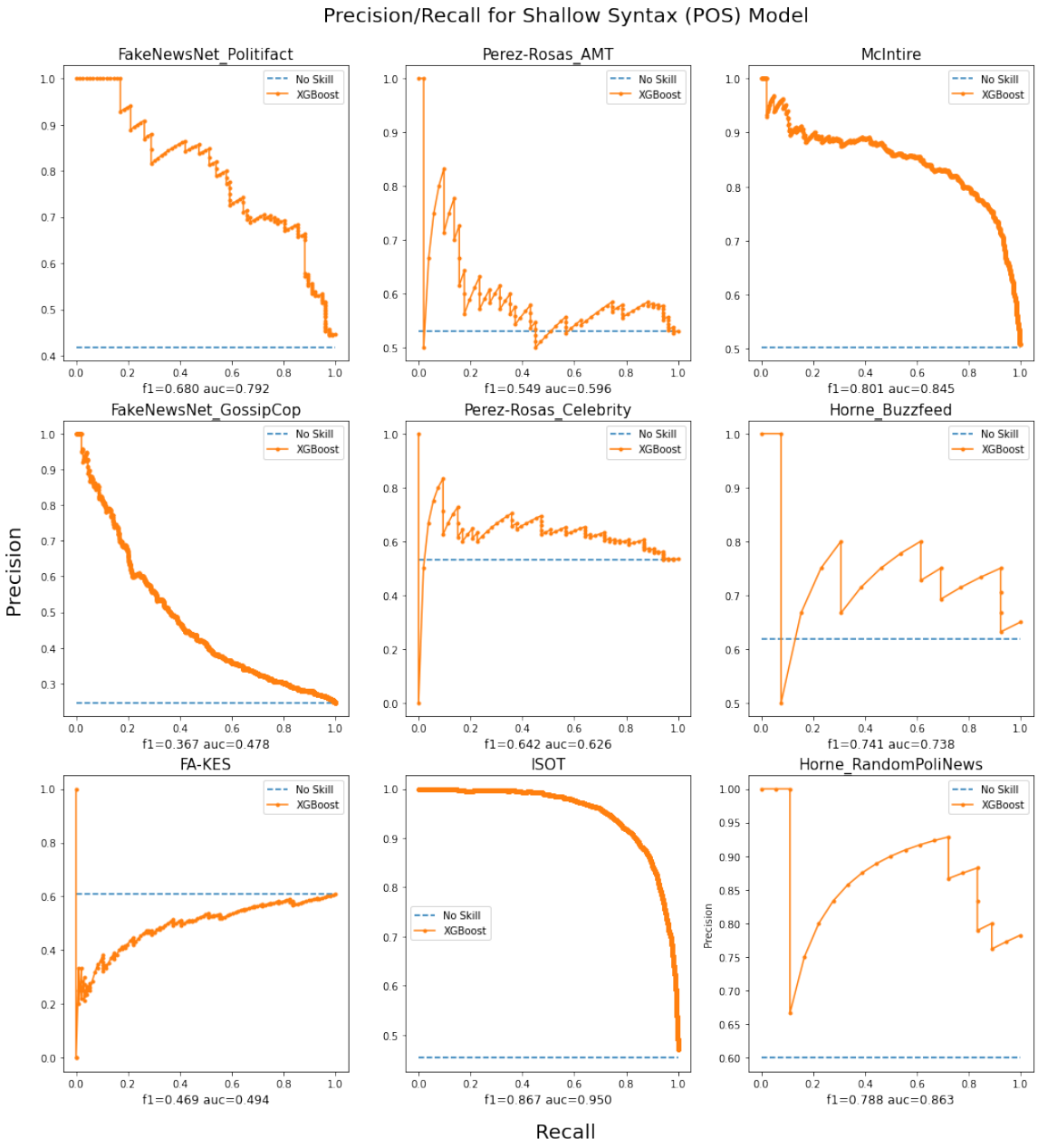

Figure 2: Precision/Recall plot for single-dataset performances on shallow syntax model (POS)

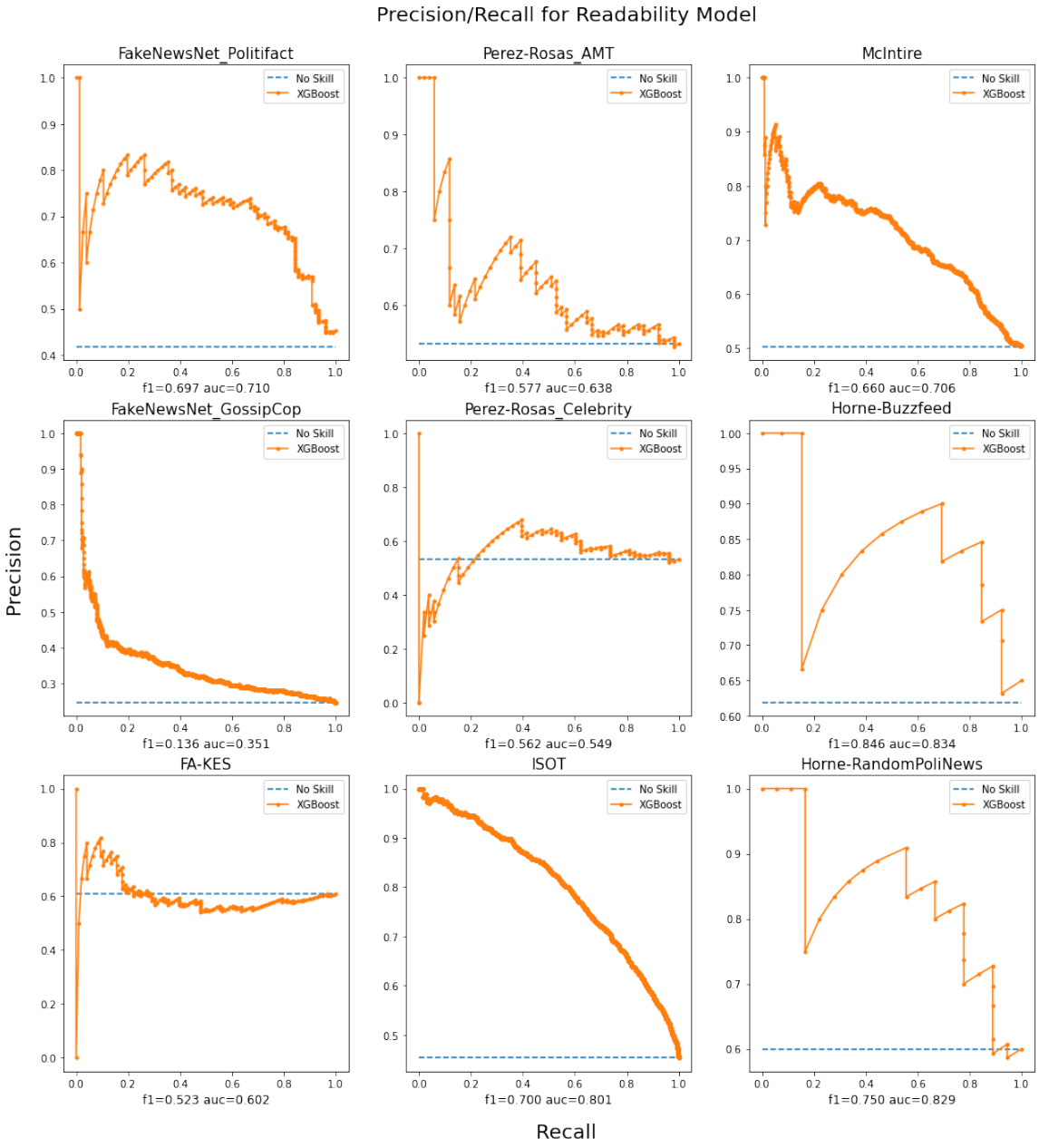

Figure 3: Precision/Recall plot for single-dataset performances on readability model

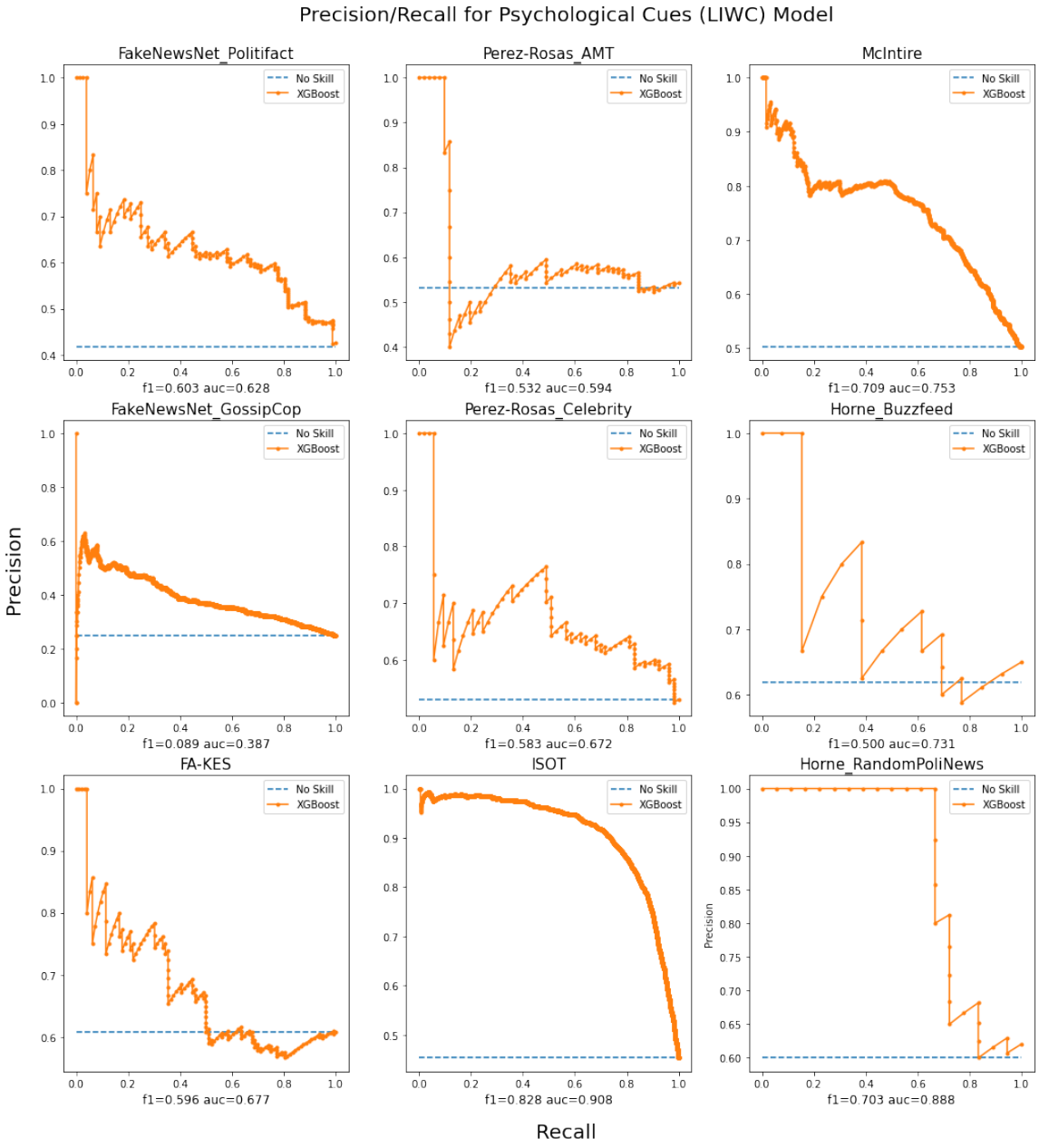

Figure 4: Precision/Recall plot for performances on psychological cues model (LIWC)