On Sentiment of Online Fake News

1st Razieh Nokhbeh Zaeem
The Center for Identity
The University of Texas at Austin
razieh@identity.utexas.edu

2nd Chengjing Li
The Center for Identity
The University of Texas at Austin
chjl@utexas.edu

3rd K. Suzanne Barber
The Center for Identity
The University of Texas at Austin
sbarber@identity.utexas.edu

***Abstract*—The presence of disinformation and fake news on the Internet and especially social media has become a major concern. Prime examples of such fake news surged in the 2016 U.S. presidential election cycle and the COVID-19 pandemic. We quantify sentiment differences between true and fake news on social media using a diverse body of datasets from the literature that contains about 100K previously labeled true and fake news. We also experiment with a variety of sentiment analysis tools. We model the association between sentiment and veracity as conditional probability and also leverage statistical hypothesis testing to uncover the relationship between sentiment and veracity. With a significance level of 99.999%, we observe a statistically significant relationship between negative sentiment and fake news and between positive sentiment and true news. The degree of association, as measured by Goodman and Kruskal's gamma, ranges between .037 to .475. Finally, we make our data and code publicly available to support reproducibility. Our results assist in the development of automatic fake news detectors.**

***Index Terms*—disinformation, misinformation, fake news, sentiment analysis, social networks, veracity**

I. Introduction

The presence of fake news and disinformation has risen to one of the paramount issues on social media. Fake news includes news articles that are intentionally false and deceptive [1]–[3]. Numerous examples exist that demonstrate how fake news creates tangible threats to the society, let alone the political and social discourse [1], [4]. For instance, in the 2016 U.S. presidential election, disinformation and social media campaigns caused great social discord and were the subjects of a Special Counsel investigation. With the 2020 elections looming, similar fake news campaigns are feared in the U.S. As another example, amid the COVID-19 pandemic, we observed a surge of fake and unreliable news on a wide range of related topics: from what caused this new coronavirus outbreak (e.g., 5G mobile networks are to blame) to how to prevent it (e.g., eating garlic prevents infection with the new coronavirus). World Health Organization (WHO)[1] and prominent news outlets[2] have dedicated efforts to actively debunk such fake news. This is how WHO director-general phrased the fast spread of COVID-19 fake news in February 2020: "Fake news spreads faster and more easily than this virus and is just as dangerous."[3]

[1]https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

[2]https://www.bbc.com/news/reality_check

[3]https://www.who.int/dg/speeches/detail/munich-security-conference

FOSINT-SI 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

Over the past decade, many researchers have sought to study fake news on social media. They consider features of the news content (e.g., the headline) or the social context (e.g., the social media profile of the posting user) that may signal the presence of fake news. (For surveys of such features, see [2], [3].) These features not only help with automatic detection of fake news (e.g., [5]–[7]), but also with educating social media users: studies suggest that human beings are surprisingly weak at detecting lies, as low as 4% better than chance [8].

In this paper, we aim to quantitatively measure the correlation between the sentiment of a social media post—whether it is a positive, negative, or neutral statement—and its veracity (i.e., whether it is fake news). While many have investigated the relationship between sentiment and veracity in social media, we were not able to find a study that *quantifies* this relationship with a *large and diverse body of data*.

Psychological research on linguistic traits of lies suggest [9], [10] that liars *use more negative emotion terms*, often attributed to nonverbal leakage cues [11]. Most research on social media agrees that the same holds true for fake news. For instance, Kwon et al. [12] demonstrated that rumors were *significantly less likely to have a positive sentiment*. Rubin et al. [13] found *more negative sentiment in fake news* in their analysis of satire/fake and real news articles. Finally, Horne et al. [14] concluded that fake news on Facebook has *more negative sentiment*.

Yet, other researchers discover different traits in online fake news. For example, Hu et al. [15] sought to find out whether sentiment differences exist between spammers and other Twitter users. They reported that spammers tend to have *more positive sentiment* compared to normal users. Furthermore, Castillo et al., in their widely cited work [16], showed that fake and non-credible news tend to exhibit more sentiment, *both positive and negative, but particularly more positive sentiment*. Even though there are minute differences, e.g., between veracity and credibility [16] or between fake news and spam [15], the above findings in related work demonstrate that the relationship between sentiment and fake news is a hard and important research question.

In this work, we dedicate our attention to quantifying what sentiment differences exist between true and fake news on social media. We seek to address multiple threats that might arise to the validity of this study, so that the results are reproducible, general, and universally applicable. We make the following contributions:

1) We obtained **seven annotated corpora** of social media/Internet posts from the literature. These corpora include real news, fake news, and satire—**totaling 95,638 posts** from platforms such as Facebook, Twitter, The Onion, various news agencies, etc. Some of these corpora were annotated by human experts (at different professional fact-checking organizations such as PolitiFact) to add the veracity of posts. Others had articles labeled with the trustworthiness of their publishers. We present the first work that investigates fake news sentiment with such a diverse and large body of corpora.
2) We experiment with both **commercial and open source sentiment analysis tools**. We consider a wide range of sentiment analysis tools, from **dictionary- to machine-learning-based** to obtain conclusions that are independent of the sentiment analysis tools.
3) We are the first to formalize the relationship between sentiment and truth as **conditional probability**, a research question that has been asked [12]–[15], but has not been formalized, before.
4) We combine statistical hypothesis testing and a variety of datasets and sentiment analysis tools to demonstrate that sentiment has a statistically significant correlation with fake news in our datasets. In particular, we leverage Goodman and Kruskal's gamma, a symmetric **measure of association** between two ordinal variables of *sentiment* and *veracity*. Gamma is a measure of association that ranges between -1 (perfect inversion) to 1 (perfect agreement), where 0 indicates no association. We make our data and code available to support reproducibility.

Using the seven annotated corpora and three sentiment analysis tools (MeaningCloud, TextBlob, and AFINN) we conclude that negative news are more likely to be false and news with positive sentiment are more likely to be true, with a significance level of 0.99999. In all of the experiments where statistical significance is achieved, **we observe a positive association between veracity and sentiment, with gamma ranging from .037 to .475**.

The rest of this paper is organized as follows. Sections II and III explain the datasets and sentiment analysis tools. Section IV details the experiments including the formalization of the problem as conditional probability and statistical hypothesis testing. Section V addresses potential threats to the validity of this study and our strategies to address them. Section VI covers related work and Section VII concludes the paper.

II. Datasets

Through an extensive review of related work, we selected the following datasets for fake news analysis:

- Liar [17]
- Two datasets from Fakenewsnet (FNN) [18], namely FNN-Gossip Cop (FNN-G) and FNN-PolitiFact (FNN-P)
- Two datasets from the paper "This Just In" (TJI) [14], namely TJI-Buzzfeed (TJI-B) and TJI-Random (TJI-R)
- Two datasets from the paper "Truth of Varying Shades" (TVS) [19], namely TVS-PolitiFact (TVS-P) and TVS-Unreliable (TVS-U)

Tables I and II show these datasets and some details about each. In Table I, the second column lists the source of the datasets as reported by their authors. In order to assign truth labels, each dataset was fact-checked by human experts (column 3) before it was released by dataset authors. The fact-checkers are professional journalists, e.g., PolitiFact (https://www.politifact.com), or other experts in the corresponding filed of the dataset, e.g., Gossip Cop (https://www.gossipcop.com). The Unreliable News dataset of TVS was not fact-checked, but it was directly collected from unreliable news sources. **In order to reproduce our study**, one can download the datasets directly from their authors, through the links given in the last column of Table I. Alternatively, one can download a cleaned version from our repository at https://github.com/UTCID/SentimentOfFakeNewsData.

Table II reports the size and distribution of truth labels in the datasets. Some of the datasets (Liar and TVS-PolitiFact) leverage the six-level truth labeling system of PolitiFact, ranging from True to Pant on Fire (totally false). Even though PolitiFact always fact-checks news on a six-grade scale, some dataset authors have combined all fake news into one label of False/Fake and all partially true stories into one label of True. Consequently, some datasets fact-checked by PolitiFact, such as FNN-PolitiFact, have only two truth labels in them. Table II also displays the distribution of the number of Satire, Hoax, Propaganda, and True news for the TVS-Unreliable dataset. The seven datasets combined have 95,638 labeled news stories from a variety of social media (e.g., Facebook and Twitter), news agencies, unreliable web pages (e.g., The Natural News and American News), and satire (The Onion, The Borowitz Report, and Clickhole).

III. Sentiment Analysis Tools

This section explains the commercial and open source sentiment analysis tools that we used to research the datasets gathered in Section II. The idea is to investigate the results generated from *different* tools to make conclusions independent of the sentiment analysis tools leveraged.

We employ the following sentiment analysis tools:

1) MeaningCloud[4]
2) TextBlob[5]
3) AFINN[6]

Table III reports these tools and some information about each. We selected a combination of open source and commercial tools, aiming to cover both types of sentiment analysis techniques: knowledge-based and machine learning.

MeaningCloud is a commercial, knowledge-based sentiment analysis tool. It leverages a dictionary to identify the local polarity of different sentences or grammatical structures in the

[4]https://www.meaningcloud.com/developer/sentiment-analysis/doc/2.1
[5]https://textblob.readthedocs.io/en/dev/index.html
[6]https://github.com/fnielsen/afinn

TABLE I: Datasets of fake, real, and satire news: Liar, Fakenewsnet (FNN), This Just In (TJI), Truth of Varying Shades (TVS).

Dataset Name	Social Media Source	Ground Truth By	Dataset Download Link and Comments
Liar	TV ads, Facebook, Twitter, etc.	PolitiFact	https://www.cs.ucsb.edu/~william/data/liar_dataset.zip We combined training, test, and validation datasets.
FNN-Gossip Cop FNN-PolitiFact	Twitter Twitter	Gossip Cop, etc. PolitiFact	https://github.com/KaiDMML/FakeNewsNet
TJI-Buzzfeed TJI-Random	Facebook Various Websites	Buzzfeed snopes.com, etc.	https://github.com/rpitrust/fakenewsdata1 We used only news titles for both TJI datasets.
TVS-PolitiFact TVS-Unreliable	PolitiFact Various Websites	PolitiFact Not Performed	https://homes.cs.washington.edu/~hrashkin/factcheck.html Unreliable news collected from Satire and Hoax websites.

TABLE II: Truth label distribution in datasets: Liar, Fakenewsnet (FNN), This Just In (TJI), Truth of Varying Shades (TVS).

	Unreliable Sources, More False			Fact-Checked, More False			Fact-Checked, More True			Total
	Satire	Hoax	Propaganda	Pants on Fire	False	Barely True	Half True	Mostly True	True	
Liar				1,047	2,507	2,103	2,627	2,454	2,053	12,791
FNN-Gossip Cop					5,323				16,805	22,128
FNN-PolitiFact					432				624	1,056
TJI-Buzzfeed					48				53	101
TJI-Random	75				75				75	225
TVS-PolitiFact				867	1,964	1,717	2,152	2,003	1,780	10,483
TVS-Unreliable	14,047	6,942	17,870						9,995	48,854

TABLE III: Sentiment analysis tools.

Tool Name	Availability	Type	Toolkit Based on	GitHub or Homepage Link
MeaningCloud	Commercial	Knowledge-Based	Synthetic Dictionary	https://www.meaningcloud.com
TextBlob	Open Source	Machine Learning, Knowledge-Based	NLTK, Pattern.en	https://github.com/sloria/TextBlob
AFINN	Open Source	Knowledge-Based	Synthetic Dictionary	https://github.com/fnielsen/afinn

text and evaluates the relationships between them, resulting in a global polarity value for the whole text. We used the free plan of MeaningCloud. **To reproduce our results** one can download and install the Excel add-in of MeaningCloud at https://www.meaningcloud.com/developer/excel-addin and run it to obtain the sentiment of each sentence in our datasets.

TextBlob is an open source Python API built on NLTK (Natural Language Tool Kit)[7] and pattern[8]. It is suitable for common Natural Language Processing (NLP) tasks such as sentiment analysis. When analyzing a sentiment, TextBlob returns a polarity score in the range $[-1.0, 1.0]$, where -1.0 is the most negative and 1.0 is the most positive. We experimented with both the machine learning module (NaiveBayesAnalyzer, an NLTK classifier trained on movie reviews) and the knowledge-based module (PatternAnalyzer, based on the pattern library). However, we found NaiveBayesAnalyzer prohibitively slow: it took an average of 18s to run to find the sentiment of each statement, for the first 50 statements of the Liar dataset. Consequently, we found it impractical to run on our tens of thousands of statements. We report the sentiments extracted by the PatternAnalyzer of TextBlob.

AFINN is an open source dictionary-based sentiment analysis tool we used in Python. It returns a polarity score for each input text snippet but the score is not in a standardized range as it is based on sum of the values of the words. AFINN is suitable for sentiment analysis tasks on text especially those targeted to get sentiment strength sum.

We applied the above tools on the datasets with identical evaluation rules for each tool. The output of MeaningCloud is a sentiment label: Very Negative (N+), Negative (N), Neutral (NEU), Positive (P), and Very Positive (P+). MeaningCloud might also fail to detect the sentiment of a sentence, reporting NONE. We treat NONE as missing value throughout the paper. For TextBlob, after getting the polarity score for each statement, we assign one of the following five categories: Very Negative (N+) for polarity score in [-1, -0.5), Negative (N) for [-0.5, 0), Neutral (NEU) for 0, Positive (P) for (0, 0.5], and Very Positive (P+) for (0.5, 1]. For the AFINN tool, the sentiment is categorized into three types based on score: Negative for (-∞, 0), Neutral for 0, and Positive for (0, ∞). The categories are slightly different than TextBlob because the AFINN scores do not have a maximum or minimum. **To reproduce our results for TextBlob and AFINN** one can obtain our Python scripts, documented in Jupiter notebooks, from our repository and run them on our cleaned datasets.

IV. Experimental Results

By applying the tools we discussed in Section III on the datasets of Section II, we augment the seven datasets

[7]https://www.nltk.org

[8]https://www.clips.uantwerpen.be/pages/pattern-en

of about 100K statements, marking each statement with its automatically assigned sentiment. Each tool might produce a slightly different sentiment for a given statement. Also, recall that each statement already has its truth label. In this section we measure the association between sentiment and veracity.

A. First Research Question (RQ)

If S is the sentiment and T is the truth label of a statement: **RQ1:** What is $P(T|S)$—i.e., given a sentiment (e.g., Very Positive) what is the probability of a truth label (e.g., Barely True), for example $P(T = \text{“}BarelyTrue\text{”}|S = \text{“}VeryPositive\text{”})$?

We are the first to formalize this question with conditional probability, even though multiple previous work [12]–[15] has asked this research question. Answering this research question reveals any possible statistical relationship between sentiment and veracity, and also it can benefit the design of new tools for automatic detection of fake news [2], [5]–[7], [20].

We calculate and chart $P(T|S)$ for the combination of four datasets and three tools. The datasets we report in these experiments are Liar, FNN-G, FNN-P, and TVS-U. Throughout this paper, to determine **statistical significance**, we set $\alpha = 0.00001$. For the TJI-B and TJI-R datasets, statistical significance was not achieved when leveraging any of the sentiment analysis tools (even with $\alpha = 0.01$), because of their small dataset sizes of 101 and 205 statements, respectively. As a result, we do not report our experiments with TJI-B and TJI-R. Finally, upon further manual investigation, we realized that the TVS-P dataset is in fact a subset of the Liar dataset, with differences in punctuation. Therefore, we do not report separate results for TVS-P as they are largely the same as Liar.

B. Results: FNN-G

We first investigate the FNN-G and FNN-P datasets, each containing only True and False labels. For each dataset and each tool, we cross tabulate the truth labels versus sentiment. Tables IV, V, and VI display the FNN-G results for the sentiment analysis tools MeaningCloud, TextBlob, and AFINN, respectively. A chi-square test of independence is performed as shown in the last row.

In order to chart the conditional probability, we calculate:

$$P(T = T_i|S = S_j) = \frac{P(T = T_i \cap S = S_j)}{\Sigma_i P(T = T_i \cap S = S_j)} \quad (1)$$

Effectively, for each row of the table, the number of statements with a given truth label and a given sentiment should be divided by the (marginal) total number of statements that have that sentiment (i.e., the row total).

Figures 1, 2, and 3 plot $P(T|S)$ for the FNN-G dataset with MeaningCloud, TextBlob, and AFINN sentiments, in order. We furthermore plot $P(T)$ in red with square markers (ALL sentiments). It is in comparison with $P(T)$ that we observe if a particular sentiment S shows a higher or lower ratio of $P(T|S)$. All three figures reveal a correlation between sentiment and veracity. Positive sentiments (darker blue) are more true when compared with the average $P(T)$ (red), where negative sentiments (lighter blue) are more false.

TABLE IV: Truth labels vs. MeaningCloud sentiments for the FNN-G dataset.

	False	True	Total
N+	208	757	965
N	1237	3127	4364
NONE	2562	7093	9655
NEU	135	471	606
P	1005	4151	5156
P+	176	1206	1382
$\chi^2(4, N = 22128) = 189.60, p < .00001.$			22128

TABLE V: Truth labels vs. TextBlob sentiments for the FNN-G dataset.

	False	True	Total
N+	115	241	356
N	644	1882	2526
NEU	3118	8722	11840
P	1293	4998	6291
P+	153	962	1115
$\chi^2(4, N = 22128) = 157.19, p < .00001.$			22128

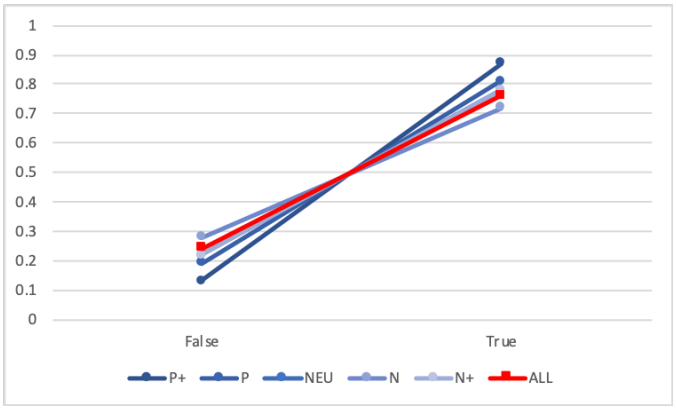

Fig. 1: $P(T|S)$ for the FNN-G dataset using MeaningCloud.

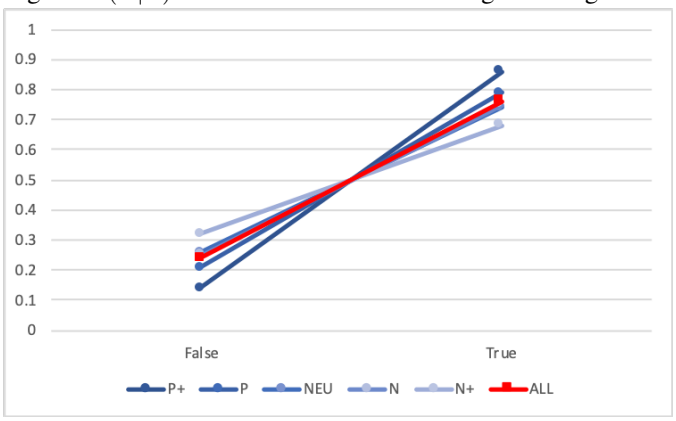

Fig. 2: $P(T|S)$ for the FNN-G dataset using TextBlob.

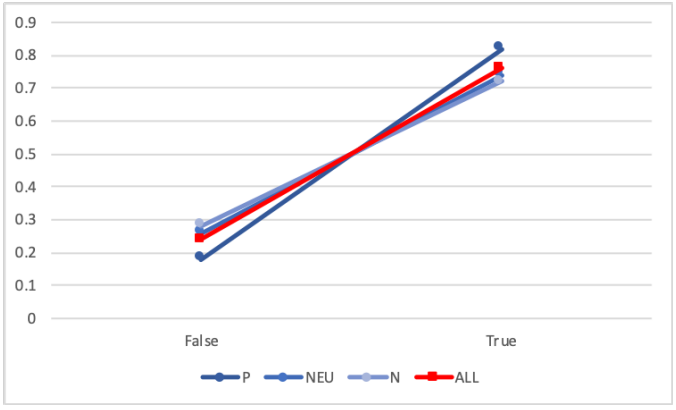

Fig. 3: $P(T|S)$ for the FNN-G dataset using AFINN.

TABLE VI: Truth labels vs. AFINN sentiments for the FNN-G dataset.

	False	True	Total
N	1405	3609	5014
NEU	2689	7604	10293
P	1229	5592	6821
$\chi^2(2, N = 22128) = 203.39, p < .00001.$			22128

TABLE VII: Truth labels vs. MeaningCloud sentiments for the FNN-P dataset.

	False	True	Total
N+	47	14	61
N	171	87	258
NONE	127	428	555
NEU	20	8	28
P	60	75	135
P+	7	12	19
$\chi^2(4, N = 1056) = 31.08, p < .00001.$			1056

TABLE VIII: Truth labels vs. TextBlob sentiments for the FNN-P dataset.

	False	True	Total
N+	10	2	12
N	86	41	127
NEU	240	462	702
P	89	111	200
P+	7	8	15
$\chi^2(4, N = 1056) = 61.08, p < .00001.$			1056

TABLE IX: Truth labels vs. AFINN sentiments for the FNN-P dataset.

	False	True	Total
N	211	96	307
NEU	165	431	596
P	56	97	153
$\chi^2(2, N = 1056) = 142.59, p < .00001.$			1056

C. Results: FNN-P

Tables VII, VIII, and IX include the results for FNN-P analyzed with MeaningCloud, TextBlob, and AFINN. Figures 4, 5, and 6 are the corresponding figures. These figures suggest a clearer association between positive sentiment and true news as well as between negative sentiment and fake news. In other words, in this dataset, knowing the sentiment of a statement (e.g., negative) dramatically increases the probability of the corresponding truth label (e.g., false). Consider Figure 5: the sentiments extracted by TextBlob for FNN-P. If the sentiment of a statement is very negative (N+), it is three times less likely, when compared to the average, to be True ($P(T|S = N+) \approx 0.2$ vs. $P(T) \approx 0.6$). The same association between negative sentiment and False labels as well as between positive sentiment and True labels holds for N, P, and P+ sentiments.

D. Results: Liar

Tables X, XI and XII cross-tabulate the truth labels vs. MeaningCloud, TextBlob and AFINN sentiments for the Liar dataset, respectively.

Figure 7 plots $P(T|S)$ for Liar and TextBlob. Figure 8 displays the same for Liar and AFINN. Sentiment has ordinal values (i.e., there is a natural order between sentiments going from N+ to P+). The application of MeaningCloud on Liar did not produce statistically significant results (even with the lower value of $\alpha = 0.01$).

We observe, in Figures 7 and 8, that even though the sentiment analysis tool used does have an effect on the $P(T|S)$ distribution, the trend is generally the same in these two figures. Both of these figures reflect the underlining truth distribution of the Liar dataset (shown with square markers in red). The relationship between sentiment and truth level is more nuance and requires a more specific statistical analysis as explained in Section IV-F.

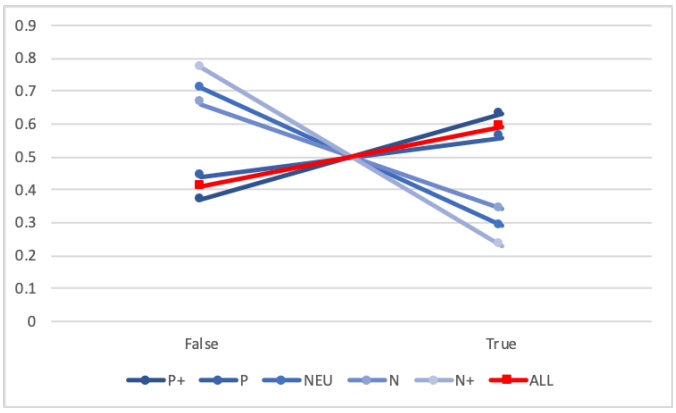

Fig. 4: $P(T|S)$ for the FNN-P dataset using MeaningCloud.

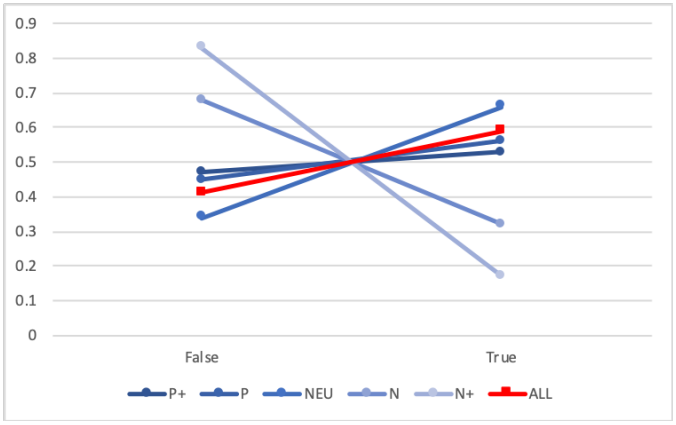

Fig. 5: $P(T|S)$ for the FNN-P dataset using TextBlob.

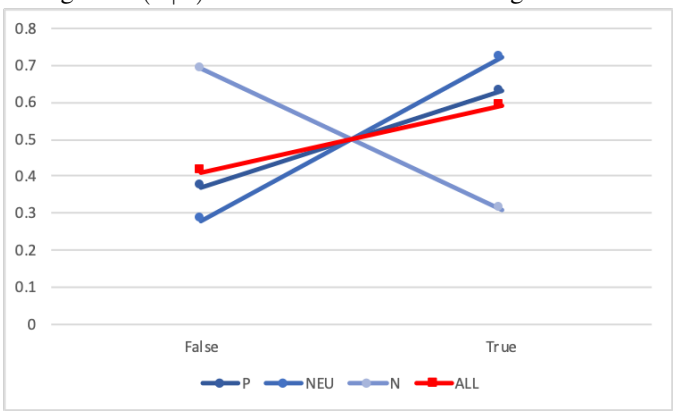

Fig. 6: $P(T|S)$ for the FNN-P dataset using AFINN.

E. Results: TVS-U

Finally, Tables XIII, XIV, and XV are the results of analyzing the TVS-U dataset with MeaningCloud, TextBlob, and AFINN, respectively. Since there is not, necessarily, an order between Satire, Hoax, and Propaganda in terms of veracity, we cannot chart $P(T|S)$ at the granularity of these veracity labels. In Section IV-F we combine Satire, Hoax, and Propaganda as fake news and perform further analysis. $P(T|S)$ for such combined truth labels shows association between negative

TABLE X: Truth labels vs. MeaningCloud sentiments for the Liar dataset.

	Pants on Fire	False	Barely True	Half True	Mostly True	True	Total
N+	77	170	152	212	212	165	988
N	275	599	529	685	561	469	3118
NONE	345	887	636	764	842	769	4243
NEU	41	104	106	122	103	68	544
P	267	641	574	674	591	491	3238
P+	42	106	106	170	145	91	660
$\chi^2(20, N = 12791) = 35.22, p = .019$. Not statistically significant.							12791

TABLE XI: Truth labels vs. TextBlob sentiments for the Liar dataset.

	Pants on Fire	False	Barely True	Half True	Mostly True	True	Total
N+	9	20	15	20	27	16	107
N	183	422	372	470	480	357	2284
NEU	523	1266	977	1113	1031	875	5785
P	310	757	698	982	880	768	4395
P+	22	42	41	42	36	37	220
$\chi^2(20, N = 12791) = 80.56, p < .00001$.							12791

TABLE XII: Truth labels vs. AFINN sentiments for the Liar dataset.

	Pants on Fire	False	Barely True	Half True	Mostly True	True	Total
N	375	783	701	960	861	658	4338
NEU	404	1065	782	961	958	862	5032
P	268	659	620	706	635	533	3421
$\chi^2(10, N = 12791) = 40.59, p < .00001$.							12791

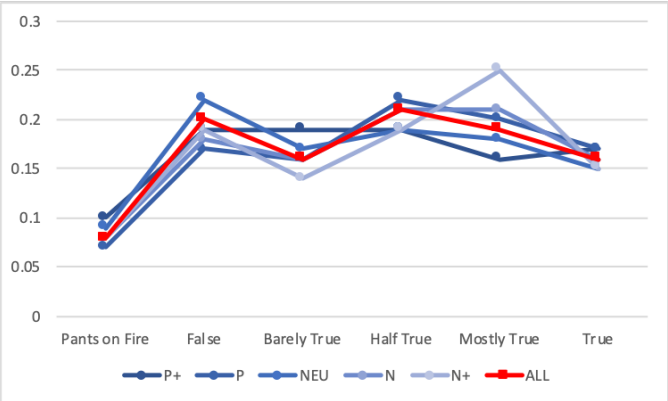

Fig. 7: $P(T|S)$ for the Liar dataset using TextBlob.

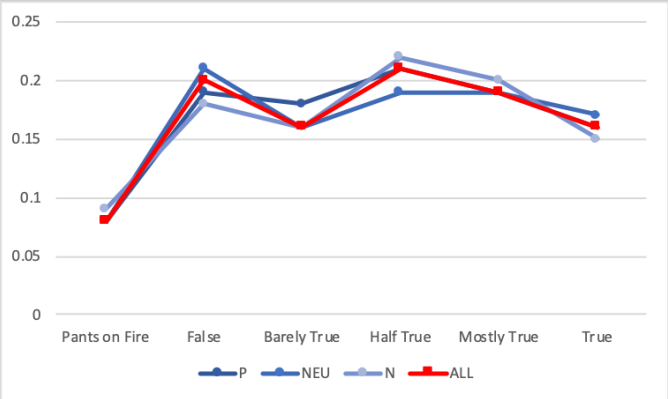

Fig. 8: $P(T|S)$ for the Liar dataset using AFINN.

TABLE XIII: Truth labels vs. MeaningCloud sent. for TVS-U.

	Satire	More False Hoax	More False Propaganda	More True Trusted	Total
N+	122	40	267	73	502
N	4109	3368	5730	2538	15745
NONE	57	10	1236	240	1543
NEU	3373	1662	5775	2192	13002
P	6208	1853	4743	4869	17673
P+	178	9	117	82	386
$\chi^2(12, N = 48851) = 2516.27, p < .00001$.					48851

TABLE XIV: Truth labels vs. TextBlob sentiments for TVS-U.

	Satire	More False Hoax	More False Propaganda	More True Trusted	Total
N+	4	9	62	2	77
N	2818	2404	1832	1661	8715
NEU	149	44	1953	322	2468
P	11037	4477	13928	7997	37439
P+	39	8	95	11	153
$\chi^2(12, N = 48852) = 3996.91, p < .00001$.					48852

sentiment and fake news and also between positive sentiment and true news. We do not chart $P(T|S)$ for the combined truth labels to preserve space.

To reproduce our charts and tables, the interested reader can import our cleaned datasets into the IBM SPSS software[9] and analyze them with Descriptive Statistics. We also provide the output of such analysis as .spv files in our repository.

[9]https://www.ibm.com/analytics/spss-statistics-software

TABLE XV: Truth labels vs. AFINN sentiments for TVS-U.

	More False			More True	
	Satire	Hoax	Propaganda	Trusted	Total
N	5828	4538	11073	4152	25591
NEU	584	226	1621	481	2912
P	7635	2178	5176	5360	20349
$\chi^2(6, N = 48852) = 3307.99, p < .00001.$					48852

F. Second Research Question

Inspired by our observation of $P(T|S)$, we form **RQ2:** *How much* is the correlation between sentiment and veracity?

Before we answer the second research question, in order to address the concern about the lack of an order between Satire, Hoax, and Propaganda in the TVS-U dataset, we put these three labels together to form the group of False statements. In Tables XIII to XV, the two new labels True and False are separated with lines. Note that in the data reported from MeaningCloud, we still ignore the NONE sentiments and treat them as missing values.

Recall chi-square tests of independence that were displayed in the last row of each table. We find that there is a statistically significant relationship between truth level and sentiment in eleven out of twelve combinations of datasets and sentiment analysis tools. In all but one combination (Liar with MeaningCloud) the results are statistically significant.

To quantify the relationship between sentiment and truth level, we recognize that they are two ordinal variables. We utilize Goodman and Kruskal's gamma, a symmetric measure of association between two ordinal variables. Gamma ranges between -1 and 1, with -1 showing perfect inversion and 1 meaning perfect agreement. The 0 value indicates no relationship. We pick gamma because it does not penalize for ties, but do report other measures such as tau-b and tau-c for the interested reader in our SPSS output files in our repository.

As Table XVI shows, gamma is always positive, except for the two cases where not statistically significant (Liar with MeaningCloud and Liar with AFINN). The positive value of gamma means positive association between being real news and having a positive sentiment. It also reveals positive association between fake news and negative sentiment. Gamma is between .037 and .475 for different dataset/tool combinations, with an average value of .220 and standard deviation of 0.130. **To reproduce these results,** the reader can utilize IBM SPSS and our datasets, or view the SPSS outputs reported in our repository.

V. Threats to Validity

The most important threat to the internal validity of this study is the applicability of the statistical measures and conditional probability. To address this threat, we carefully picked the most proper mathematical formulations we could find, e.g., gamma. In addition, we report other statistical measures in our publicly available repository. The major threat to the external validity of this study is whether the datasets and tools selected are representative of the entire body of social media/Internet posts and sentiment analysis tools. To mitigate this threat, we performed a wide study of all the available datasets in the literature we could find and selected seven diverse datasets with a total of about 100K statements. We further considered and experimented with various types of sentiment analysis tools, both commercial and open source.

VI. Related Work

Most of the literature on sentiment of social media fake news (e.g., [9]–[11]) agrees that there is a correlation between negative sentiment and fake news, and so does the psychological research on linguistic traits of lies. From this first group is the work of Kwon et al. [12] who experimented with a dictionary-based sentiment analysis tool applied on a set of about 2K tweets and demonstrated that rumors were *significantly less likely to have a positive sentiment*. Another example is from Rubin et al. [13] who found *more negative sentiment in fake news* in their analysis of 360 satire/fake and real news articles. Finally, Horne et al. [14] leveraged dictionary-based and academic tools for sentiment analysis. They concluded that fake news has *more negative sentiment*. This particular conclusion, however, was based on only one of their datasets with about 70 fake and real news Facebook posts. Our results are in line with this group. However, we utilize a bigger more diverse body of corpora and we quantitatively measure the correlation.

There is a second group, however, that comes to a different conclusion. For example, Castillo et al. [16] showed that fake and non-credible news tend to exhibit more sentiment, *both positive and negative, but particularly more positive sentiment*. They do not report the exact number of tweets in their dataset which was evaluated for credibility, or the sentiment analysis algorithm they use, as they cover a host of other features too. It is important to note that, credibility, as the perceived veracity value, is not the same as the actual veracity. Hu et al. [15] sought to find out whether sentiment differences exist between spammers and other social media users. Based on two Twitter datasets of 42K and 20K Twitter users, they reported that spammers tend to have *more positive sentiment* compared to normal users. They used a supervised method (linear regression) for sentiment analysis. Their work is close to ours, but they utilize sentiment for effective social *spammer detection*: according to later publications form the same authors [3], spammer detection most closely resembles finding social bots—non-human users that circulate fake news.

Besides analyzing the sentiment of fake news, other researchers have studied the sentiment of *comments* posted on fake news stories [6], [18], [21]–[23]. Our focus, however, is on the news content.

A. Datasets

We obtained seven datasets from the literature as explained in Section II. In addition, we downloaded the Credbank [24] dataset. This dataset, however, did not include the news text but rather links to the news. In an effort to download the news text, we realized that out of 2,282 links provided in this dataset, 291 were links to videos, 207 were links to photos,

TABLE XVI: Ordinal by ordinal symmetric measure, gamma. Asymptotic Standard Error calculated not assuming the null hypothesis. Approximate T calculated using the asymptotic standard error assuming the null hypothesis.

Dataset	Sentiment Analysis	Gamma	Asymptotic Standard Error	Approximate T	Approximate Significance
Liar	MeaningCloud	-.006	.012	-.491	.623
	TextBlob	.037	.010	3.695	.000
	AFINN	-.007	.010	-.741	.459
FNN-G	MeaningCloud	.200	.016	12.095	.000
	TextBlob	.146	.013	11.177	.000
	AFINN	.177	.013	13.724	.000
FNN-P	MeaningCloud	.380	.067	5.308	.000
	TextBlob	.177	.059	2.981	.003
	AFINN	.475	.047	9.314	.000
TVS-U	MeaningCloud	.241	.009	25.994	.000
	TextBlob	.097	.013	7.577	.000
	AFINN	.270	.010	25.979	.000

and 601 were broken links. As a result, about half of the links in this dataset were not useful to our study. Consequently, we did not use the Credbank dataset.

VII. Conclusion

We collected a diverse corpora of seven labeled true/fake news datasets, containing about 100K statements. Using various sentiment analysis tools, we studied the relationship between *sentiment* and *truth labels* (i.e., veracity), two ordinal variables. Leveraging conditional probability and statistical hypothesis testing, we found a statistically signification relationship between negative sentiment and fake news as well as between positive sentiment and true news, with $\alpha = 0.00001$. The degree of association, as measured by Goodman and Kruskal's gamma, ranges between .037 to .475, with an average value of .220. We make our code, datasets, and results publicly available, so that other researchers can fully reproduce this study. Our work paves the way for the design of better automatic fake news detectors as well as enlightens social media users on the characteristics of fake news.

References

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[2] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1–4, 2015.

[3] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[4] E. C. Tandoc Jr, Z. W. Lim, and R. Ling, "Defining "fake news" a typology of scholarly definitions," *Digital journalism*, vol. 6, no. 2, pp. 137–153, 2018.

[5] N. Hassan, F. Arslan, C. Li, and M. Tremayne, "Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1803–1812.

[6] M. D. Vicario, W. Quattrociocchi, A. Scala, and F. Zollo, "Polarization and fake news: Early warning of potential misinformation targets," *ACM Transactions on the Web (TWEB)*, vol. 13, no. 2, pp. 1–22, 2019.

[7] M. Hardalov, I. Koychev, and P. Nakov, "In search of credible news," in *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. Springer, 2016, pp. 172–180.

[8] C. F. Bond Jr and B. M. DePaulo, "Accuracy of deception judgments," *Personality and social psychology Review*, vol. 10, no. 3, pp. 214–234, 2006.

[9] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality and social psychology bulletin*, vol. 29, no. 5, pp. 665–675, 2003.

[10] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam," in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, 2013, pp. 497–501.

[11] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

[12] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *2013 IEEE 13th International Conference on Data Mining*. IEEE, 2013, pp. 1103–1108.

[13] V. L. Rubin, N. Conroy, Y. Chen, and S. 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*, 2016, pp. 7–17.

[14] B. D. Horne and S. 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 *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[15] X. Hu, J. Tang, H. Gao, and H. Liu, "Social spammer detection with sentiment information," in *2014 IEEE International Conference on Data Mining*. IEEE, 2014, pp. 180–189.

[16] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proceedings of the 20th international conference on world wide web*, 2011, pp. 675–684.

[17] W. Y. Wang, "liar, liar pants on fire: A new benchmark dataset for fake news detection," in *the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2017, pp. 422–426.

[18] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, pp. 171–188, 2020.

[19] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: Analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2931–2937.

[20] T.-C. Huang, R. N. Zaeem, and K. S. Barber, "It is an equal failing to trust everybody and to trust nobody: Stock price prediction using trust filters and enhanced user sentiment on Twitter," *ACM Transactions on Internet Technology (TOIT)*, vol. 19, no. 4, pp. 1–20, 2019.

[21] F. Zollo, P. K. Novak, M. Del Vicario, A. Bessi, I. Mozetič, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Emotional dynamics in the age of misinformation," *PloS one*, vol. 10, no. 9, 2015.

[22] L. Cui, S. Wang, and D. Lee, "Same: sentiment-aware multi-modal embedding for detecting fake news," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019, pp. 41–48.

[23] S. Vosoughi, M. Mohsenvand, and D. Roy, "Rumor gauge: Predicting the veracity of rumors on Twitter," *ACM transactions on knowledge discovery from data (TKDD)*, vol. 11, no. 4, pp. 1–36, 2017.

[24] T. Mitra and E. Gilbert, "Credbank: A large-scale social media corpus with associated credibility annotations," in *Ninth International AAAI Conference on Web and Social Media*, 2015.