Misleading or Falsification? Inferring Deceptive Strategies and Types in Online News and Social Media

Svitlana Volkova, Jin Yea Jang*
Pacific Northwest National Laboratory
Richland, Washington
svitlana.volkova@pnnl.gov,jinyea.jang@keti.re.kr

ABSTRACT

Deceptive information in online news and social media has had dramatic effect on our society in recent years. This study is the first to gain deeper insights into writers' intent behind digital misinformation by analyzing psycholinguistic signals: moral foundations and connotations extracted from different types of deceptive news ranging from strategic disinformation to propaganda and hoaxes. To ensure consistency of our findings and generalizability across domains, we experiment with data from: (1) confirmed cases of disinformation in news summaries, (2) propaganda, hoax, and disinformation news pages, and (3) social media news. We first contrast lexical markers of biased language, syntactic and stylistic signals, and connotations across deceptive news types including disinformation, propaganda, and hoaxes, and deceptive strategies including misleading or falsification. We then incorporate these insights to build machine learning and deep learning predictive models to infer deception strategies and deceptive news types. Our experimental results demonstrate that unlike earlier work on deception detection, content combined with biased language markers, moral foundations, and connotations leads to better predictive performance of deception strategies compared to syntactic and stylistic signals (as reported in earlier work on deceptive reviews). Falsification strategy is easier to identify than misleading strategy. Disinformation is more difficult to predict than to propaganda or hoaxes. Deceptive news types (disinformation, propaganda, and hoaxes), unlike deceptive strategies (falsification and misleading), are more salient, and thus easier to identify in tweets than in news reports. Finally, our novel connotation analysis across deception types provides deeper understanding of writers' perspectives and therefore reveals the intentions behind digital misinformation.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Neural networks**;

KEYWORDS

natural language processing; machine learning; deep learning; misinformation; deception; social media analysis; connotation analysis

*now at Korea Electronics Technology Institute (KETI)

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3188728

ACM Reference Format:
Svitlana Volkova, Jin Yea Jang. 2018. Misleading or Falsification? Inferring Deceptive Strategies and Types in Online News and Social Media. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3184558.3188728

1 INTRODUCTION

Information is power. Disinformation undermines this power. According to the World Economic Forum report [25], massive digital disinformation is listed as one of the main risks of the modern society. Different types of false information have been actively shared on the web and through social media. Regardless of the false information's type and purpose – misinterpretation by mistake or targeted propaganda campaigns [34], the consequences it has on people's lives are harmful [18, 32] and sometimes even fatal [21, 40].

Sparse availability of data annotated with deceptive news types, credibility levels, or checked facts limited research on analyzing and identifying false information in online communications [17]. The majority of earlier work on automatic deception detection relied on manually constructed small corpora to build predictive models to detect deceptive product reviews [36, 39]. Recent work has focused on annotated data from PolitiFact [50, 62], satirical news e.g., The Onion [54], and news articles [45], which typically targeted only one domain (generally politics), or focused on specific event types e.g., shooting events [56], natural disasters [57], or elections [16]. These studies have resulted in important findings about the effect of misinformation spread [33], influence campaigns [2, 7], and social bots [38] within specific newsworthy events.

Only [37] and [61] explored Twitter data to evaluate linguistic realizations in mainstream vs. alternative news sources, and built models to predict information credibility and deceptive news types e.g., propaganda, hoaxes, and clickbait vs. trusted news, respectively. However, to the best of our knowledge, limited prior work focused on analyzing deception strategies e.g., misleading vs. falsification [1], and analyzed linguistic realizations of disinformation. Thus, the unique source of public data that consists of confirmed cases of disinformation (https://euvsdisinfo.eu/, @EUvsDisinfo) annotated by the European Union's East Strategic Communications Task Force analyzed in this study, in addition to deceptive webpages and social media communications, will further advance understanding and improve predictive models for factuality assessment and information credibility online. We outline the main contributions of this study below.

First, we examine linguistic realizations across deceptive strategies: misleading vs. falsification, and types: disinformation, propaganda, and hoaxes across domains including disinformation reports,

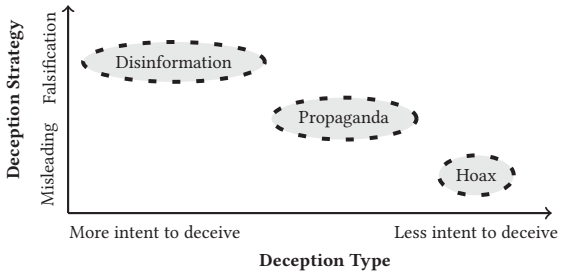


Figure 1: Deceptive news categorized based on deception type and strategy.

news pages, and social media posts. We report statistically significant differences in psycholinguistic cues, biased and subjective language, and moral foundations behind deceptive news content.

Second, we analyze connotations toward agents and targets of deceptive news across different types of deceptive content to provide deeper understanding of writers' perspectives and intentions behind strategic disinformation, propaganda, and hoaxes.

Finally, we incorporate our findings on differences in linguistic realizations across deception types and strategies to build predictive models for deception detection. We contrast machine learning and deep learning model performance trained on content, stylistic, syntactic, and psycholinguistic signals across domains to ensure generalizability of our models.

Deception Types

To study deception on a spectrum depending on writers intent, we focus on three popular types of deception–strategic disinformation, propaganda and hoaxes, and two deception strategies–misleading or falsification as shown in Figure 1. The example tweets of deceptive news along with definitions for three deceptive news types are presented below.

- **Hoax** is a type of misinformation that aims to deliberately deceive the reader [30]. The example hoax tweet: *BREAKING! Massive Volcano Eruption Only 32 Miles Away From MAJOR Nuclear Plant! Consciously Enlightened.*
- **Propaganda** is a form of persuasion that attempts to influence the emotions, attitudes, opinions, and actions of specified target audiences for political, ideological, and religious purposes through the controlled transmission of deceptive, selectively omitting, and one-sided messages [35]. The example propaganda tweet: *The UN Plans To Implement Universal Biometric Identification For All Of Humanity By 2030.*
- **Disinformation** denotes false facts that are conceived in order to deliberately deceive the audience [30]. The example disinformation tweet: *Soren Kern: European Union Declares War on Internet Free Speech.*

Deception Strategies

One of the contributions of this work focuses on analyzing disinformation strategies – misleading and falsification as defined below, and contrasting psycholinguistic realizations that distinguish them.

- **Misleading** strategy includes cases of topic changes, irrelevant information, and equivocations: *Austria and Slovenia are closing their borders with Serbia to stem the flow of refugees.*[1]
- **Falsification** strategy deals with contradictions or distortions: *Ukrainian engineers made a mistake constructing AN-178 plane, they have to fly with ballast.*[2]

Other types of deceptive strategies have been recently studied in [1], e.g., exaggeration and omission. We consider these strategies to be very specific cases of misleading strategies that require background knowledge to identify and exclude them from our analysis.

Unlike any previous work, we analyze and contrast moral foundations and connotations across deceptive news types, strategies, and data sources. We also incorporate our findings into predictive models that rely on machine learning and deep learning to automatically infer deception types and strategies across domains: summaries, webpages, and tweets. To evaluate model generalizability and consistency of our findings, we contrast our results across data sources (domains) and report the predictive power of different psycholinguistic signals extracted from news: content, syntax, style, connotations, and moral foundations behind deceptive content.

2 BACKGROUND

Previous work on deception detection primarily focused on spoken and written personal and criminal narratives and conference calls [4, 12, 13, 63]. Only recently researchers have proposed methods for deception detection in online communications: book and hotel reviews or essay data [14, 15]. The existing models primarily rely on shallow linguistic features e.g., n-grams, part-of-speech tags, readability, and syntactic complexity features combined with state-of-the-art machine learning models [36, 41, 46, 53]. Researchers found that textual clues to deception include self reference, negation statements, complaints, and generalizing items. More specifically, Linguistic Inquiry and Word Count (LIWC) features [43] were used to show that deceptive stories compared to true stories had lower cognitive complexity e.g., *cause, know*; use fewer exclusive words e.g., *but, except*; more negative emotion words e.g., *hate, worthless, sad* and more motion verbs e.g., *walk, go, carry*; fewer self-references, fewer first-person singular pronouns e.g., *I, me, and my* [39]. Recent work focused on building predictive models to distinguish between fake and verified news [45, 55], infer deceptive news types [50, 61], and assess information credibility [62].

Several papers analyzed the impact of false information spread on the web, focusing on hoaxes on encyclopedia [10, 30], satirical news [54], and misinformation propagation in social networks [6, 27, 48, 58]. Most of the work in this area has focused on engineered features, e.g., network structure and shallow semantic features [1, 3, 59]. More recent work took into account deeper semantic understanding of language for fact-checking and defined a statistical model to detect all mentions of events in news and then assess the degree of uncertainty around whether each mentioned event happened [9, 31]. The focus of this work is not to study misinformation contagion in social networks. Instead, we analyze linguistic realizations of misinformation that lead to misleading or falsified

[1] Disproof: Austria and Slovenia don't have common borders with Serbia.
[2] No proofs given.

Deception Types			
	Propaganda	Hoaxes	Disinfo
Webpages	17,872	5,297	166
Tweets	3,834	453	205
Deception Strategies			
	Misleading	Falsification	
Summaries	616	1,376	
Webpages	81	85	
Tweets	96	109	

Table 1: Dataset statistics: the number of news pages, tweets, and disinformation summaries annotated with deception types and strategies.

statements, and incorporate these psycholinguistic signals into predictive models to automatically infer deception types and strategies across domains.

3 DECEPTION DATASETS

This section presents three datasets used for our analysis: disinformation summaries, deceptive news pages and tweets, and data annotation and pre-processing details.

3.1 Disinformation Summaries

We rely on public data annotated with confirmed cases of disinformation (https://euvsdisinfo.eu/, @EUvsDisinfo) collected by the European Union's East Strategic Communications Task Force in 2015 – 2016. The total number of confirmed disinformation cases is 1,992 with 36 cases reported per week on average.

We annotated disinformation summaries as falsification and misleading using crowdsourcing. We first marked summaries that contained substrings *unprovable, no evidence, no proof, no supporting evidence,* etc. in the disproof as falsification. We then showed five annotators the remaining disinformation summaries with the URLs. Measured pairwise inter-annotator agreement on all responses was 64% (when at least four annotators agree 66%), and the kappa score on all responses was 0.43 (when at least four annotators agree, the score is 0.22). In total, we ended up with 1,376 (69%) summaries annotated as falsification and 616 (31%) as misleading.

We parsed all summaries, news pages, and tweets (described below) using the state-of-the-art dependency parser – SyntaxNet[3] [47]. We extracted grammar and syntax: subjects, verbs and objects, and the part-of-speech tags. This is an important step toward understanding agents and themes of deception and contrasting connotations across deception types.

3.2 Deceptive News Pages

We followed URLs in disinformation summaries to collect the original news pages. We propagated misleading and falsification labels from disinformation summaries to label the news pages. For our analysis we only focused on English webpages. In total we had 85 (51%) news pages in English marked as falsification and 81 (49%) as misleading. In addition, we downloaded 17,872 propaganda and 5,297 hoax news pages to contrast disinformation with other deceptive news types.

[3]https://github.com/tensorflow/models/tree/master/syntaxnet

3.3 Deceptive Tweets

We used subject, verb, and object tuples extracted from parsed disinformation summaries and the disinformation summary date field to query Twitter public API to extract disinformation tweets. We collected 7,969 disinformation tweets and retweets. After deduplication (4,457 tweets) and removing @mentions, URLs, and RTs (985 tweets), we removed tweets with edit distance and TFIDF cosine similarity above 0.8 to avoid overfitting. We ended up with a clean sample of 205 disinformation tweets annotated as misleading vs. falsification. We also collected tweets produced by example propaganda and hoax accounts produced in 2016.

In addition to annotating tweets with deceptive strategies e.g., misleading vs. falsification we also asked annotators to define targeted topics of disinformation following the annotation strategy proposed in [5]. We report top targeted topics of disinformation in summaries and tweets. We found that the most popular targets of disinformation are politics, security, and economics. Moreover, tweets and summaries are framed to mislead rather than falsify information about politics and external affairs. More summaries about security are falsified rather than misleading, but more tweets about security are misleading rather than falsified.

	Summaries		Tweets	
Frame	Mislead	Falsify	Mislead	Falsify
Political	**46.7**	41.7	**55.9**	37.0
Security	23.0	**31.1**	**37.2**	33.3
Economic	**9.6**	3.5	3.5	**10.2**
Crime	5.8	**6.3**	–	3.7
Cultural	2.7	**4.2**	–	–
Public	**2.7**	2.5	2.3	**2.8**
External	2.7	**3.9**	1.2	**9.3**

Table 2: Top targeted topics of disinformation summaries and tweets.

4 APPROACH

This section describes predictive models and different types of signals to infer deception types and strategies across domains.

4.1 Predictive Models

We use the state-of-the-art classifiers – MaxEntropy and RandomForest implemented in scikit-learn [42], Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN)-based models [61] implemented in keras[4] to predict deception types and strategies. We run experiments using 10-fold cross-validation and rely on the lexical, syntactic, stylistic, psycholinguistic, and connotation signals described below.

4.1.1 Content. For machine learning models, we rely on TFIDF features extracted from webpages, summaries, and tweets. We take advantage of StandardScalar[5] and dimensionality reduction[6] to avoid overfitting. For neural network models, we initialize the embedding layer with pre-trained Glove embeddings [44].

[4]https://keras.io/
[5]http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScalar.html
[6]http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

4.1.2 Style, complexity, and readability. We rely on language complexity and readability features designed to measure how difficult the text is to understand e.g., Automated Readability Index (ARI), Flesch-Kincaid readability tests, Coleman-Liau index, and the Gunning fog index etc.[7]

4.1.3 Syntax. We use syntactic signals demonstrated to be effective for deception detection in book and hotel reviews [36]. To evaluate the predictive power of syntactic signals across domains, we rely on part-of-speech tags extracted by SyntaxNet [47].

4.1.4 Biased language, moral foundations, and psycholinguistic signals. Recent work demonstrated that biased lexicons [52] and moral foundations [61] are predictive of suspicious news in social media. We outline biased language lexicons used to form our analysis and predictive models below.

- *Factive verbs* [29]: presuppose the truth of their complement clause e.g., *realize, know, regret.*
- *Assertive verbs* [24]: a complement clause requires a preposition; the level of certainty depends on the asserting verb e.g., *point out, claim.*
- *Report verbs* [52]: include verbs e.g., *admit, blame, criticize.*
- *Hedges* [26]: an expression of "tentativeness and possibility" or language corresponding to the "writer withholding full commitments to statements" [9] e.g., *may, possibly, seems.*
- *Implicative verbs* [28]: imply the truth or untruth of their complement, depending on the polarity of the main predicate e.g., *decline, hesitate, avoid, neglect.*
- *Intensifiers and dramatic adverbs*: include superlatives and comparatives e.g., *nicest, damper*; action, manner, modal adverbs e.g., *accidentally, freely, truly.*
- *Moral foundations* [20, 22, 23] Basic moral values emerge from cultural and evolutionary factors that people support – *care and harm, fairness and cheating, loyalty and betrayal, authority and subversion, and purity and degradation.* People differ in the way they endorse these values; thus, writers of different types of deceptive news might appeal to specific moral foundations of their readers.
- *Psycholinguistic cues* [43] Linguistic Inquiry Word Count (LIWC) cues include *imperative commands, personal pronouns, emotional language, quotations, and inclusions.*

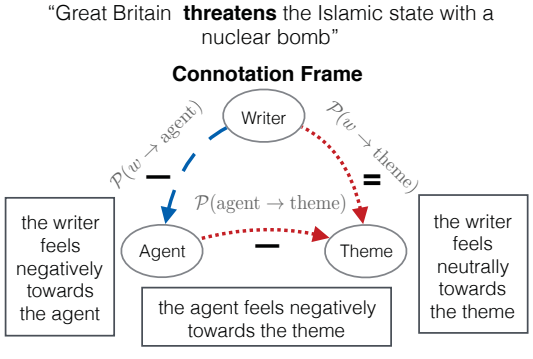


Figure 2: Example connotation frame.

[7] https://github.com/nltk/nltk_contrib/tree/master/nltk_contrib/readability

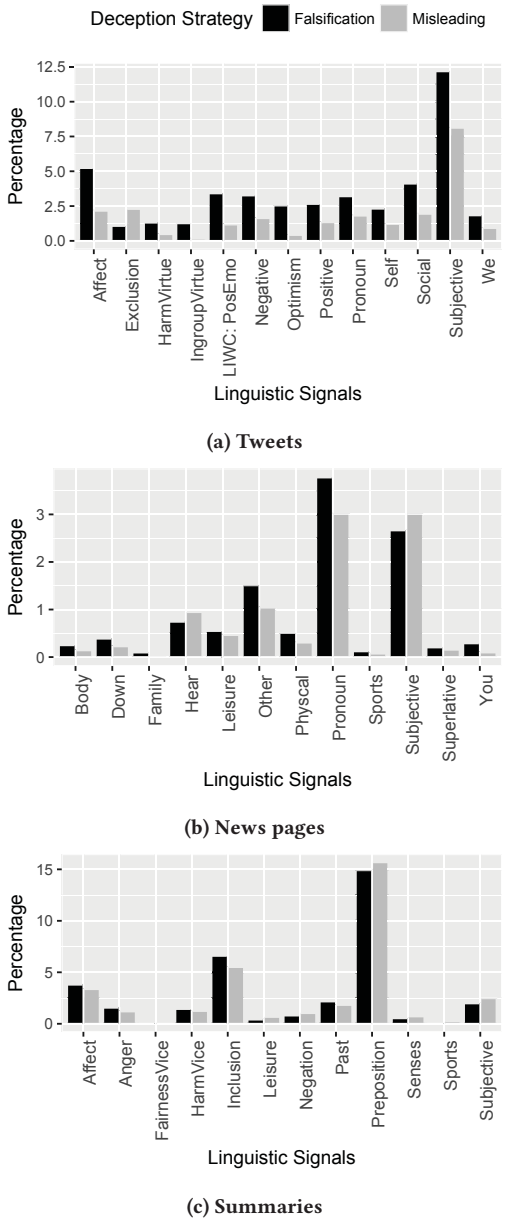
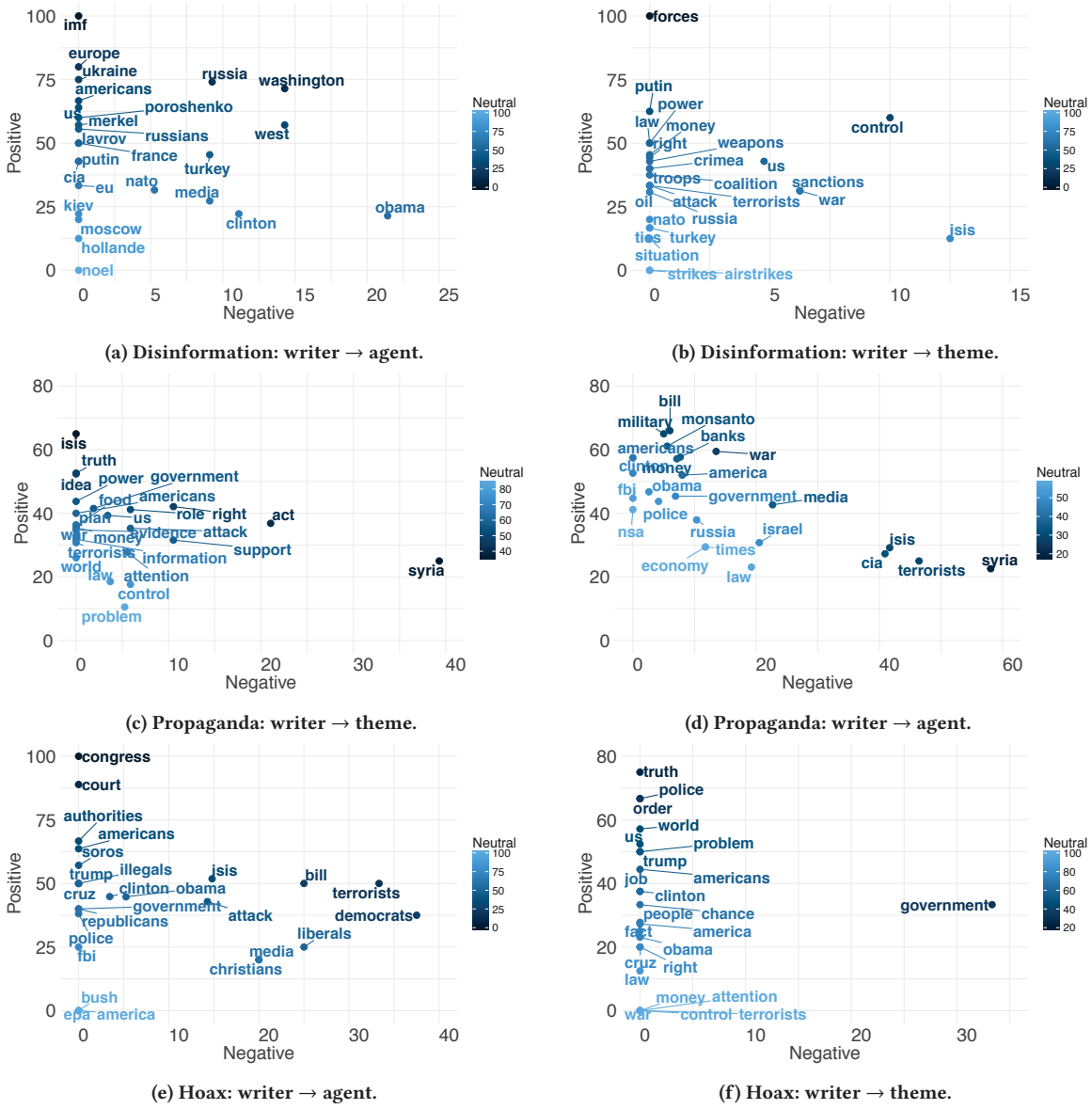


Figure 3: Psycholinguistic markers of biased and subjective language and moral foundations in misleading vs. falsification statements across domains (% of summaries, % of sentences in news pages, % of tweets). Only statistically significant results are reported ($p \leq 0.05$).

4.1.5 Connotations. Connotation frames allow the reader to estimate the author's perspective: positive, negative, or neutral toward the subject and object of the sentence, as well as the perspective of the subject toward the object [51]. Connotations provide insights about feelings that a word invokes in addition to its literal or primary meaning. The example connotations for the disinformation summary *Great Britain* **threatens** *the Islamic state with a nuclear bomb* is shown in Figure 2 include writer $\overset{-}{\rightarrow}$ Great Britain, writer $\overset{=}{\rightarrow}$ Islamic state, Great Britain (subj) $\overset{-}{\rightarrow}$ Islamic state (obj).



(a) Disinformation: writer → agent.

(b) Disinformation: writer → theme.

(c) Propaganda: writer → theme.

(d) Propaganda: writer → agent.

(e) Hoax: writer → agent.

(f) Hoax: writer → theme.

Figure 4: Connotation frame analysis in disinformation, propaganda, and hoax tweets and news headlines. We plot the writer's perspective (positive – Y axis, negative – X axis) toward agents and themes. Subjectivity (the opposite of neutrality or objectivity) of agents and themes are shown using color gradient – the darker the color, the more subjective (less neutral) the perspective is.

5 RESULTS

5.1 Linguistic Analysis of Deceptive News

Figures 3 presents differences in linguistic realizations between misleading and falsification statements across domains: tweets, news pages, and summaries. First, we observe that misleading vs. falsification deceptive strategies are realized in linguistically different ways across domains. The only shared linguistic signals are subjective language, harm, moral foundation, and negation.

Second, we found that subjective language and affect signals are the most represented in tweets, pronouns, and subjective language in news pages and prepositions and inclusion in summaries. Interestingly, differences in HarmVirtue and IngroupVirtue moral foundations are significant in tweets but HarmVice and FairnessVice in summaries. Finally, we estimated that language is more subjective in misleading statements than falsified content in summaries and news pages but not in tweets; there are more prepositions in summaries and fewer pronouns in news pages in falsified content than in misleading content.

		Summaries				News Pages				Tweets			
Signals	**Model**	**ROC**	**F1**	**F1:M**	**F1:F**	**ROC**	**F1**	**F1:M**	**F1:F**	**ROC**	**F1**	**F1: M**	**F1:F**
Content	ME	0.60	**0.58**	**0.49**	0.67	0.55	**0.55**	**0.41**	0.69	0.81	0.81	0.80	0.81
	RF	0.57	0.57	0.36	**0.78**	0.51	0.47	0.25	0.69	0.71	0.69	0.69	0.69
	LSTM	**0.61**	0.57	0.40	0.75	**0.57**	0.52	0.34	0.70	**0.92**	**0.81**	0.80	0.82
	CNN	0.60	0.57	0.40	0.74	0.50	0.44	0.16	0.73	0.90	0.80	0.79	0.80
Syntax	ME	0.54	0.53	0.40	0.65	0.58	**0.56**	0.39	0.73	0.78	0.76	0.74	0.77
	RF	0.55	0.54	0.33	0.75	0.51	0.49	0.27	0.70	0.73	0.73	0.73	0.73
	LSTM	**0.61**	0.57	0.39	0.65	**0.60**	0.46	0.20	0.72	**0.93**	**0.82**	0.80	**0.84**
	CNN	0.59	**0.62**	0.38	0.73	0.56	0.51	0.24	0.78	0.86	0.73	0.72	0.75
Style	ME	0.50	0.48	0.40	0.54	0.54	**0.51**	0.37	0.65	0.61	0.58	0.58	0.59
	RF	0.54	0.63	0.33	0.75	0.48	0.48	0.32	0.63	0.59	0.58	0.56	0.60
	LSTM	0.56	0.48	0.19	0.76	**0.56**	0.48	0.32	0.64	**0.93**	**0.80**	0.80	0.81
	CNN	**0.61**	**0.64**	0.39	0.74	0.54	0.41	0.05	0.76	0.91	0.77	0.74	0.79
Connotations	ME	0.50	0.53	0.36	0.62	0.47	0.43	0.32	0.55	0.56	0.51	0.36	0.67
	RF	0.52	0.61	0.31	0.72	0.48	0.47	0.28	0.66	0.58	0.53	0.37	0.70
	LSTM	0.62	0.57	0.38	0.75	**0.60**	**0.55**	0.38	0.72	**0.93**	**0.82**	0.81	0.83
	CNN	**0.63**	**0.66**	0.42	0.76	0.53	0.41	0.04	0.78	0.90	0.77	0.74	0.80
Lexicons	ME	0.56	0.57	0.44	0.65	0.56	0.52	0.39	0.65	0.65	0.63	0.62	0.63
	RF	0.54	0.65	0.32	0.76	0.54	0.53	0.34	0.71	0.65	0.63	0.63	0.63
	LSTM	0.61	0.56	0.39	0.74	**0.63**	**0.57**	0.43	0.72	**0.92**	**0.81**	0.80	0.83
	CNN	0.60	0.63	0.38	0.74	0.42	0.39	0.04	**0.78**	0.88	0.78	0.76	0.80
ALL	ME	0.58	0.64	0.43	0.74	0.46	0.42	0.24	0.61	0.83	0.82	0.80	0.83
	RF	0.56	**0.66**	0.35	0.77	0.45	0.44	0.23	0.64	0.71	0.69	0.67	0.70
	LSTM	**0.61**	0.57	0.38	0.75	**0.54**	**0.48**	0.27	0.69	**0.92**	**0.82**	**0.82**	0.82
	CNN	0.60	0.65	0.39	0.75	0.50	0.44	0.19	0.69	0.89	0.77	0.75	0.79

Table 3: Results for predicting deception strategies – misleading vs. falsification across domains (summaries, webpage, and tweets) using different signals – content, syntactic, stylistic (readability and complexity), affect (connotations), and lexicon features. Baseline Random Forest (RF) and MaxEntropy (ME) models rely on individual signals. Neural network models LSTM and CNN combine content and individual signals. F1 stands for macro F1, F1:M – misleading class, F1:F – falsification. ROC and F1 for top predictive models across different signals and domains are highlighted in bold.

Implications of Linguistic Analysis. Unlike [45, 50, 61, 62] that analyzed linguistic differences across disinformation types–propaganda, hoaxes, clickbaits, and trusted news in web pages, PolitiFact statements and tweets, respectively–we focused on differences in linguistic realizations between misleading and falsified statements. Interestingly, compared to earlier work, we found that only a small portion of linguistic signals is useful to distinguish between misleading and falsified statements. Our results not only show differences across different domains–news pages, tweets, and summaries–but also identified linguistic realizations useful to build predictive models for factuality assessment of the statement without relying on the external knowledge. Our findings will allow us to improve fact-checking systems by going beyond fake news classification [45, 62].

5.2 Connotation Analysis of Deceptive News

The purpose of identifying perspectives toward agents and themes of deceptive content is to capture the hidden agenda behind strategic misinformation and disinformation. In Figure 4 we contrast connotations for top agents and themes across deception types: disinformation, propaganda, and hoaxes. We plot writers' positive and negative perspectives about agents and themes and outline our key observations below.

Writer → Agent. Authors of disinformation express positive perspectives about *Europe, Ukraine*, and negative perspectives about *Obama and Clinton* agents; mixed perspectives (both positive and negative) are expressed toward *Russia, Washington,* and *West.* In contrast, writers of propaganda express positive connotations toward *military, Monsanto, bill* agents but negative perspectives toward *terrorists, Syria* and *CIA*; mixed perspectives expressed toward *government, Israel.* Writers of hoaxes express positive connotations toward *congress, court, authorities* agents, and negative perspectives toward *democrats, liberals, terrorists.*

Writer → Theme. Authors of disinformation express positive perspectives about *forces, power, law,* and *money* themes, and negative perspectives about *Turkey* and *strikes*; mixed perspectives are expressed about *terror, sanctions.* Writers of propaganda express positive perspectives about *truth, idea, power* themes, and negative connotations toward *Syria.* Finally, hoax writers express positive connotations toward *truth, police, order* themes, and negative connotations toward *government.*

Implications of Connotation Analysis. Note, our connotation analysis combines quantitative and qualitative methods: first, we automatically parse tweets to extract agents, verbs, and themes;

		News Pages				Tweets			
Signals	**Model**	**F1**	**F1:P**	**F1:H**	**F1:D**	**F1**	**F1:P**	**F1:H**	**F1:D**
Content	ME	0.48 ± 0.02	0.54 ± 0.06	0.56 ± 0.03	0.34 ± 0.08	0.65 ± 0.04	0.52 ± 0.08	0.70 ± 0.06	0.73 ± 0.05
	RF	0.79 ± 0.05	0.70 ± 0.06	0.72 ± 0.08	0.94 ± 0.03	0.71 ± 0.05	0.67 ± 0.07	0.70 ± 0.06	0.77 ± 0.04
	LSTM	**0.82 ± 0.01**	**0.76 ± 0.04**	**0.76 ± 0.02**	0.94 ± 0.01	**0.87 ± 0.02**	**0.83 ± 0.03**	**0.87 ± 0.03**	0.92 ± 0.02
	CNN	0.78 ± 0.05	0.71 ± 0.10	0.75 ± 0.06	0.88 ± 0.02	0.80 ± 0.02	0.75 ± 0.03	0.78 ± 0.05	0.88 ± 0.03
Syntax	ME	0.70 ± 0.04	0.68 ± 0.03	0.68 ± 0.07	0.73 ± 0.04	0.58 ± 0.04	0.55 ± 0.06	0.59 ± 0.05	0.60 ± 0.04
	RF	0.68 ± 0.04	0.65 ± 0.06	0.63 ± 0.05	0.77 ± 0.03	0.57 ± 0.03	0.54 ± 0.05	0.61 ± 0.05	0.57 ± 0.03
	LSTM	**0.75 ± 0.04**	0.72 ± 0.03	0.72 ± 0.08	0.82 ± 0.05	**0.85 ± 0.02**	0.79 ± 0.04	0.86 ± 0.03	0.89 ± 0.02
	CNN	0.72 ± 0.04	0.70 ± 0.06	0.69 ± 0.10	0.78 ± 0.02	0.77 ± 0.02	0.69 ± 0.03	0.76 ± 0.02	0.87 ± 0.02
Style	ME	0.50 ± 0.09	0.35 ± 0.15	0.63 ± 0.02	0.51 ± 0.16	0.57 ± 0.04	0.44 ± 0.08	0.62 ± 0.04	0.64 ± 0.02
	RF	0.53 ± 0.07	0.51 ± 0.09	0.58 ± 0.10	0.49 ± 0.03	0.61 ± 0.04	0.54 ± 0.06	0.67 ± 0.06	0.61 ± 0.06
	LSTM	**0.66 ± 0.06**	0.67 ± 0.06	0.65 ± 0.07	0.68 ± 0.08	**0.84 ± 0.01**	0.78 ± 0.03	0.84 ± 0.02	0.89 ± 0.02
	CNN	0.54 ± 0.06	0.48 ± 0.10	0.63 ± 0.05	0.52 ± 0.06	0.74 ± 0.05	0.66 ± 0.08	0.74 ± 0.04	0.83 ± 0.05
Connotations	ME	0.43 ± 0.03	0.36 ± 0.08	0.44 ± 0.07	0.49 ± 0.04	0.39 ± 0.04	0.57 ± 0.05	0.11 ± 0.12	0.49 ± 0.03
	RF	0.40 ± 0.08	0.39 ± 0.15	0.39 ± 0.10	0.42 ± 0.06	0.41 ± 0.05	0.57 ± 0.04	0.21 ± 0.11	0.44 ± 0.04
	LSTM	**0.67 ± 0.03**	0.67 ± 0.04	0.65 ± 0.05	0.69 ± 0.02	**0.85 ± 0.02**	0.79 ± 0.03	0.85 ± 0.03	0.90 ± 0.03
	CNN	0.61 ± 0.02	0.57 ± 0.05	0.57 ± 0.05	0.70 ± 0.06	0.75 ± 0.04	0.66 ± 0.06	0.74 ± 0.08	0.85 ± 0.01
Lexicons	ME	0.53 ± 0.06	0.34 ± 0.13	0.63 ± 0.04	0.62 ± 0.04	0.57 ± 0.03	0.51 ± 0.07	0.59 ± 0.05	0.61 ± 0.04
	RF	0.58 ± 0.03	0.54 ± 0.04	0.62 ± 0.02	0.59 ± 0.06	0.63 ± 0.04	0.58 ± 0.06	0.65 ± 0.06	0.66 ± 0.05
	LSTM	**0.67 ± 0.04**	0.66 ± 0.04	0.66 ± 0.05	0.70 ± 0.03	**0.85 ± 0.04**	0.80 ± 0.04	0.85 ± 0.03	0.90 ± 0.04
	CNN	0.64 ± 0.04	0.65 ± 0.03	0.57 ± 0.07	0.70 ± 0.07	0.76 ± 0.04	0.68 ± 0.06	0.76 ± 0.05	0.85 ± 0.02
ALL	ME	0.57 ± 0.01	0.64 ± 0.04	0.67 ± 0.03	0.40 ± 0.04	0.72 ± 0.04	0.62 ± 0.07	0.76 ± 0.05	0.79 ± 0.03
	RF	0.81 ± 0.03	0.74 ± 0.06	0.73 ± 0.04	**0.96 ± 0.03**	0.74 ± 0.04	0.68 ± 0.05	0.74 ± 0.06	0.80 ± 0.03
	LSTM	**0.81 ± 0.05**	0.76 ± 0.06	0.76 ± 0.08	0.91 ± 0.02	**0.86 ± 0.03**	0.81 ± 0.04	0.86 ± 0.06	**0.92 ± 0.01**
	CNN	0.74 ± 0.12	0.66 ± 0.14	0.70 ± 0.16	0.86 ± 0.05	0.78 ± 0.03	0.71 ± 0.05	0.78 ± 0.04	0.86 ± 0.01

Table 4: Results for predicting deception types: hoaxes, disinformation and propaganda across domains (webpage and tweets) using different signals: content, syntactic, stylistic (readability and complexity), connotations (targeted perspectives), and psycho-linguistic signals. Baseline Random Forest (RF) and MaxEntropy (ME) models rely on individual signals. Neural network models LSTM and CNN combine content and individual signals. F1 stands for macro F1, and individual F1:H for hoaxes class, F1:D – disinformation, F1:P – propaganda. F1 for top predictive models across different signals and domains are highlighted in bold. Confidence intervals obtained using 10-fold cross validation are omitted due to space constraints.

then we get a quantitative estimate of targeted perspectives (connotations) driven by the verb toward each agent and theme of deceptive statement; third, we qualitatively visualize positive vs. negative perspectives toward agents and themes across deception types. Such two-fold analysis allows us to demonstrate how agents and themes of strategic deception vary across deception types, and, as a result, qualitatively identify the hidden agenda of content from propaganda vs. disinformation vs. hoax tweets over the same time period. Such differences in hidden agenda may drive deeper insights about (a) intent behind deceptive content shared online and (b) targeted audiences influenced by such deceptive content.

Furthermore, our findings on connotations behind deceptive content online will contribute to research on misinformation propagation in social networks by identifying gatekeepers and persistent minorities, e.g., trolls and bots that potentially spread deceptive content [8, 38, 60].

5.3 Prediction Results

Tables 3 and 4 present prediction results for two disinformation strategies–misleading vs. falsification–and three deception types: propaganda, disinformation, and hoaxes, respectively. We report classification results obtained using the state-of-the-art models: MaxEntropy (ME), RandomForest (RF), Convolutional Neural Networks (CNN), and Long Short-Term Memory Networks (LSTM). We evaluate and contrast the predictive power of content, syntax, style, affect, and psycholinguistic signals across domains: summaries, webpages, and tweets. Content represents what is being discussed e.g., targeted topics, keywords; syntax and style represent how the content is being discussed; connotations and psycholinguistic signals represent how emotional the discussion is. We report model performance using the area under under the receiver operating characteristic (ROC) curve weighted, weighted F1 score, and F1 measures for each class.

Our best model for deception type prediction relies on content signals in combination with LSTM model yields F1 of .82 for webpages and F1 of .87 for tweets. Our best model for deception strategy detection relies on content combined with connotation signals incorporated into a CNN model yields F1 of .66 and ROC of .63 for summaries; content combined with lexicon signals incorporated into a LSTM model yields F1 of .57 and ROC of .63 for news pages; and content combined with either connotation or syntax signals incorporated into a LSTM model yields F1 of .82 and ROC of .93 for tweets. Our more detailed findings are shown below.

- *Psycholinguistic signals:* Unlike earlier work on deception detection in product reviews [36, 39], content combined with moral foundations and connotations are more predictive of deception strategies than syntactic and stylistic features. Interestingly, for predicting deception types (Table 4) content > syntax > connotations > style > lexicon signals. However, for predicting deception strategies (Table 3) connotations > lexicon > syntax > content > style.
- *Predictive models:* As expected, neural network models e.g., CNNs and LSTMs achieve higher performance compared to baseline machine learning models.
- *Deception types:* Disinformation (F1 is 0.96 for news pages, F1 is 0.92 for tweets) is less difficult to predict compared to propaganda (F1 is 0.76 for news pages, F1 is 0.83 for tweets) or hoaxes (F1 is 0.76 for news pages, F1 is 0.87 for tweets).
- *Deception strategies:* As expected, deception strategies are more difficult to predict than deception types, even though it is a binary classification compared to 3-way classification. Falsification strategy is easier to infer than misleading strategy in news pages (F1 is 0.78 and 0.41, respectively) and disinformation summaries (F1 is 0.78 and 0.49, respectively). However, falsification strategy is as easy to classify as misleading strategy in tweets (F1 is 0.84 and 0.82, respectively). Syntax is more predictive in tweets, lexicons in news pages, and content in summaries. Content is the most predictive of misleading strategy across all domains.
- *Domains:* Deception types, unlike deception strategies, are easier to identify in tweets (F1 is 0.87) than in news pages (F1 is 0.82).

6 SUMMARY AND DISCUSSION

To the best of our knowledge, this is the first work that focused on building predictive models to infer deception types and strategies across multiple domains, regardless the topic or event in the deceptive message, by analyzing and incorporating psycholinguistic signals e.g., connotations and moral foundations behind deceptive news content into predictive models.

Recent work on fake news detection online and in social media has focused on developing models to distinguish between fake vs. verified content (binary classification) [11, 45], or estimating credibility level of the tweet or PolitiFact statement (regression) [37, 50, 62]. These binary classification models achieve F1 between 78% for the news page domain [45] and as high as 95% for social media [61]. Regression models that predict the level of information credibility on Twitter [37] achieve 68%, and models that predict credibility of PolitiFact statements achieve 65% and 27%, respectively [50, 62].

Only [61] and [50] analyzed language differences to build models for *classifying types of deceptive content*, e.g., propaganda, hoaxes, satire, and clickbaits. Our work not only goes beyond that by incorporating disinformation into multi-class models but also develops models for inferring whether the *statement is misleading or falsification*. Unlike *misinformation* e.g., hoaxes [27, 32, 58], rumors [48], and clickbaits that have been a target of recent research, *disinformation* is more difficult to capture and study even though it is more harmful and influential on our society. *Misinformation* is conveyed in the honest but mistaken belief that the relayed incorrect facts are true, *disinformation* denotes false facts that are conceived in order to deliberately deceive the audience. Finally, our study targets a very important issue of model generalizability that is often overlooked, and demonstrates how predictive models, linguistic and connotation analysis generalize across domains such as online news, disinformation statements, and social media.

In the future, we plan to extend our models to make predictions (a) across languages by relying on a multilingual connotation framework recently developed by [49] and (b) across social media environments. Moreover, inspired by recent work [19] that incorporates images from social media in addition to text to predict the clickbaitness (score between 0 and 1) of tweets, we plan to incorporate image signals into our approach.

ACKNOWLEDGMENTS

The research described in this paper was conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy.

The authors would like to thank Hannah Rashkin from the University of Washington for her help with data collection and annotation, and anonymous reviewers for their helpful comments and suggestions.

REFERENCES

[1] Darren Scott Appling, Erica J. Briscoe, and Clayton J. Hutto. 2015. Discriminative Models for Predicting Deception Strategies. In *Proceedings of WWW*. 947–952.
[2] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of WSDM*. 65–74.
[3] Erica J Briscoe, D Scott Appling, and Heather Hayes. 2014. Cues to deception in social media communications. In *International Conference on System Sciences*. 1435–1443.
[4] David B Buller, Judee K Burgoon, JA Daly, and JM Wiemann. 1994. Deception: Strategic and nonstrategic communication. *Strategic interpersonal communication* (1994), 191–223.
[5] Dallas Card, Amber E Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of ACL*. 438–444.
[6] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of WWW*. 675–684.
[7] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. 2010. Measuring user influence in Twitter: The million follower fallacy. *Proceedings of ICWSM* 10, 10-17 (2010), 30.
[8] Justin Cheng, Cristian Danescu-Niculescu-Mizil, Jure Leskovec, and Michael Bernstein. 2017. Anyone Can Become a Troll. *American Scientist* 105, 3 (2017), 152.
[9] Eunsol Choi, Chenhao Tan, Lillian Lee, Cristian Danescu-Niculescu-Mizil, and Jennifer Spindel. 2012. Hedge detection as a lens on framing in the GMO debates: A position paper. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*. 70–79.
[10] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one* 10, 6 (2015), e0128193.
[11] Niall J Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.
[12] Bella M DePaulo, James J Lindsay, Brian E Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological bulletin* 129, 1 (2003), 74.
[13] Paul Ekman and Wallace V Friesen. 1969. Nonverbal leakage and clues to deception. *Psychiatry* 32, 1 (1969), 88–106.
[14] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of ACL*. 171–175.
[15] Song Feng, Longfei Xing, Anupam Gogar, and Yejin Choi. 2012. Distributional Footprints of Deceptive Product Reviews. *Proceedings of ICWSM* 12 (2012), 98–105.

[16] Emilio Ferrara. 2017. Disinformation and social bot operations in the run up to the 2017 French presidential election. (2017).

[17] Eileen Fitzpatrick and Joan Bachenko. 2012. Building a data collection for deception research. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*. 31–38.

[18] Fortune. 2016. Fake Bloomberg news report drives Twitter stock up 8%. http://fortune.com/2015/07/14/fake-twitter-bloomberg-report/. (2016). Accessed: 2016-12-12.

[19] Maria Glenski, Ellyn Ayton, Dustin Arendt, and Svitlana Volkova. 2017. Fishing for Clickbaits in Social Images and Texts with Linguistically-Infused Neural Network Models. *arXiv preprint arXiv:1710.06390* (2017).

[20] Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology* 96, 5 (2009), 1029.

[21] Guardian. 2015. Woman dies after taking "diet pills" bought over internet. https://www.theguardian.com/society/2015/apr/21/woman-dies-after-taking-diet-pills-bought-over-internet. (2015).

[22] Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research* 20, 1 (2007), 98–116.

[23] Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus* 133, 4 (2004), 55–66.

[24] Joan B. Hooper. 1975. On assertive predicates. In *Syntax and Semantics*, J. Kimball (Ed.), Vol. 4. 91–124.

[25] Lee Howell. 2013. Digital wildfires in a hyperconnected world. *WEF Report* (2013).

[26] Ken Hyland. 2005. *Metadiscourse*. Wiley Online Library.

[27] Fang Jin, Wei Wang, Liang Zhao, Edward Dougherty, Yang Cao, Chang-Tien Lu, and Naren Ramakrishnan. 2014. Misinformation propagation in the age of Twitter. *Computer* 47, 12 (2014), 90–94.

[28] Lauri Karttunen. 1971. Implicative verbs. *Language* (1971), 340–358.

[29] Paul Kiparsky and Carol Kiparsky. 1968. *Fact.* Indiana University.

[30] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *Proceedings of WWW*. 591–602.

[31] Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event Detection and Factuality Assessment with Non-Expert Supervision. *Proceedings of EMNLP* (2015).

[32] Newton Lee. 2014. Misinformation and Disinformation. In *Facebook Nation*. Springer, 169–188.

[33] Kristina Lerman and Rumi Ghosh. 2010. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. *Proceedings of ICWSM* (2010), 90–97.

[34] Cristian Lumezanu, Nick Feamster, and Hans Klein. 2012. # bias: Measuring the tweeting behavior of propagandists. (2012).

[35] Feamster Lumezanu and H Klein. 2012. Measuring the tweeting behavior of propagandists. In *Proceedings of ICWSM*.

[36] Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP*. 309–312.

[37] Tanushree Mitra, Graham P Wright, and Eric Gilbert. 2017. A parsimonious language model of social media credibility across disparate events. In *Proceedings of CSCW*. 126–145.

[38] Bjarke Mønsted, Piotr Sapieżyński, Emilio Ferrara, and Sune Lehmann. 2017. Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PLOS ONE* 12, 9 (09 2017), 1–12. https://doi.org/10.1371/journal.pone.0184148

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

[40] NYTimes. 2016. Google and Facebook Take Aim at Fake News Sites. http://www.nytimes.com/2016/11/15/technology/google-will-ban-websites-that-host-fake-news-from-using-its-ad-service.html. (2016).

[41] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of ACL-HLT*. 309–319.

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[43] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001), 2001.

[44] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings EMNLP*. 1532–1543.

[45] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic Detection of Fake News. *arXiv preprint arXiv:1708.07104* (2017).

[46] Verónica Pérez-Rosas and Rada Mihalcea. 2015. Experiments in Open Domain Deception Detection. *Proceedings of EMNLP* (2015), 1120–1125.

[47] Slav Petrov. 2016. Announcing SyntaxNet: The WorldâĂŹs Most Accurate Parser Goes Open Source. *Google Research Blog, May* 12 (2016), 2016.

[48] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of EMNLP*. 1589–1599.

[49] Hannah Rashkin, Eric Bell, Yejin Choi, and Svitlana Volkova. 2017. Multilingual Connotation Frames: A Case Study on Social Media for Targeted Sentiment Analysis and Forecast. In *Proceedings of ACL*, Vol. 2. 459–464.

[50] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of EMNLP*. 2921–2927.

[51] Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation Frames: A Data-Driven Investigation. In *Proceedings of ACL*. 311–321.

[52] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of ACL*. 1650–1659.

[53] Victoria L Rubin. 2010. On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–10.

[54] Victoria L Rubin, Yimin Chen, and Niall J Conroy. 2015. Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–4.

[55] Victoria L Rubin, Niall J Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In *Proceedings of NAACL-HLT*. 7–17.

[56] Kate Starbird. 2017. Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter. In *Proceedings of ICWSM*. 230–239.

[57] Bruno Takahashi, Edson C Tandoc, and Christine Carmichael. 2015. Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines. *Computers in Human Behavior* 50 (2015), 392–398.

[58] Marcella Tambuscio, Giancarlo Ruffo, Alessandro Flammini, and Filippo Menczer. 2015. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proceedings of WWW*. 977–982.

[59] Michail Tsikerdekis and Sherali Zeadally. 2014. Online deception in social media. *Commun. ACM* 57, 9 (2014), 72–80.

[60] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of ICWSM*. 280–289.

[61] Svitlana Volkova, Kyle Shaffer, Jin Yean Jang, and Nathan Hodas. 2017. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In *Proceedings of ACL*. 647–653.

[62] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of ACL*. 422–426.

[63] Miron Zuckerman, Richard Koestner, and Robert Driver. 1981. Beliefs about cues associated with deception. *Journal of Nonverbal Behavior* 6, 2 (1981), 105–114.