Satirical News Detection and Analysis using Attention Mechanism and Linguistic Features

Fan Yang and **Arjun Mukherjee**
Department of Computer Science
University of Houston
{fyang11,arjun}@uh.edu

Eduard Gragut
Computer and Information Sciences
Temple University
edragut@temple.edu

Abstract

Satirical news is considered to be entertainment, but it is potentially deceptive and harmful. Despite the embedded genre in the article, not everyone can recognize the satirical cues and therefore believe the news as true news. We observe that satirical cues are often reflected in certain paragraphs rather than the whole document. Existing works only consider document-level features to detect the satire, which could be limited. We consider paragraph-level linguistic features to unveil the satire by incorporating neural network and attention mechanism. We investigate the difference between paragraph-level features and document-level features, and analyze them on a large satirical news dataset. The evaluation shows that the proposed model detects satirical news effectively and reveals what features are important at which level.

1 Introduction

> *"When information is cheap, attention becomes expensive."* — James Gleick

Satirical news is considered to be entertainment. However, it is not easy to recognize the satire if the satirical cues are too subtle to be unmasked and the reader lacks the contextual or cultural background. The example illustrated in Table 1 is a piece of satirical news with subtle satirical cues.

Assuming readers interpret satirical news as true news, there is not much difference between satirical news and fake news in terms of the consequence, which may hurt the credibility of the media and the trust in the society. In fact, it is reported in *the Guardian* that people may believe satirical news and spread them to the public regardless of the ridiculous content[1]. It is also concluded that fake news is similar to satirical news via a thorough comparison among true news, fake news, and satirical news (Horne and Adali, 2017). This paper focuses on *satirical news detection* to ensure the trustworthiness of online news and prevent the spreading of potential misleading information.

...
"Kids these days are done with stories where things happen," said CBC consultant and world's oldest child psychologist Obadiah Sugarman. "We'll finally be giving them the stiff Victorian morality that I assume is in vogue. Not to mention, doing a period piece is a great way to make sure white people are adequately represented on television."
...

Table 1: A paragraph of satirical news

Some works tackling fake news and misleading information favor to discover the truth (Xiao et al., 2016; Wan et al., 2016) through knowledge base (Dong et al., 2015) and truthfulness estimation (Ge et al., 2013). These approaches may not be feasible for satirical news because there is no ground-truth in the stories. Another track of works analyze social network activities (Zhao et al., 2015) to evaluate the spreading information (Gupta et al., 2012; Castillo et al., 2011). This could be ineffective for both fake news and satirical news because once they are distributed on the social network, the damage has been done. Finally, works evaluating culture difference (Pérez-Rosas and Mihalcea, 2014), psycholinguistic features (Ott et al., 2011), and writing styles (Feng et al., 2012) for deception detection are suitable for satirical news detection. These works consider features at document level, while we observe that satirical cues are usually located in certain para-

[1] https://www.theguardian.com/media/2016/nov/17/facebook-fake-news-satire

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1979–1989
Copenhagen, Denmark, September 7–11, 2017.

graphs rather than the whole document. This indicates that many document level features may be superfluous and less effective.

To understand how paragraph-level features and document-level features are varied towards detection decision when only document level labels are available, we propose a 4-level neural network in a character-word-paragraph-document hierarchy and utilize attention mechanism (Bahdanau et al., 2014) to reveal their relative difference. We apply psycholinguistic features, writing stylistic features, structural features, and readability features to understand satire. The paragraph-level features are embedded into attention mechanism for selecting highly attended paragraphs, and the document-level features are incorporated for the final classification. This is the first work that unveils satirical cues between paragraph-level and document-level through neural networks to our knowledge.

We make the following contributions in our paper:

- We propose a 4-level hierarchical network for satirical news detection. The model detects satirical news effectively and incorporates attention mechanism to reveal paragraph-level satirical cues.
- We show that paragraph-level features are more important than document-level features in terms of the psycholinguistic feature, writing stylistic feature, and structural feature, while the readability feature is more important at the document level.
- We collect satirical news (16,000+) and true news (160,000+) from various sources and conduct extensive experiments on this corpus[2].

2 Related Work

We categorize related works into four categories: content-based detection for news genre, truth verification and truthfulness evaluation, deception detection, and identification of highly attended component using attention mechanism.

Content-based detection for news genre.Content-based methods are considerably effective to prevent satirical news from being recognized as true news and spreading through social media. Burfoot and Baldwin (2009) introduce headline features, profanity, and slang to embody satirical news. They consider absurdity as the major device in satirical news and model this feature by comparing entity combination in a given document with Google query results. Rubin et al. (2016) also consider absurdity but model it through unexpected new name entities. They introduce additional features including humor, grammar, negative affect, and punctuation to empower the detection. Besides satirical news, Chen et al. (2015) aim to detect click-baits, whose content exaggerates fact. Potthast et al. (2017) report a writing style analysis of hyperpartisan news. Barbieri et al. (2015) focus on multilingual tweets that advertise satirical news.

It is noteworthy that satirical news used for evaluation in above works are of limited quantity (around 200 articles). Diverse examples of satire may not be included as discussed by Rubin et al. (2016). This issue inspires us to collect more than 16,000 satirical news for our experiment.

Truth discovery and truthfulness evaluation. Although truth extraction from inconsistent sources (Ge et al., 2013; Wan et al., 2016; Li et al., 2016) and from conflicting sources (Yin et al., 2008; Li et al., 2014b), truth inference through knowledge base (Dong et al., 2015), and discovering evolving truth (Li et al., 2015) could help identify fact and detect fake news, they cannot favor much for satirical news as the story is entirely made up and the ground-truth is hardly found. Analyzing user activities (Farajtabar et al., 2017) and interactions (Castillo et al., 2011; Mukherjee and Weikum, 2015) to evaluate the credibility may not be appropriate for satirical news as it cannot prevent the spreading. Therefore, we utilize content-based features, including psycholinguistic features, writing stylistic features, structural features, and readability features, to address satirical news detection.

Deception detection. We believe satirical news and opinion spam share similar characteristics of writing fictitious and deceptive content, which can be identified via a psycholinguistic consideration (Mihalcea and Strapparava, 2009; Ott et al., 2011). Beyond that, both syntactic stylometry (Feng et al., 2012) and behavioral features (Mukherjee et al., 2013b) are effective for detecting deceptive reviews, while stylistic features are practical to deal with obfuscating and imitat-

[2]Please contact the first author to obtain the data

ing writings (Afroz et al., 2012). However, deceptive content varies among paragraphs in the same document, and so does satire. We focus on devising and evaluating paragraph-level features to reveal the satire in this work. We compare them with features at the document level, so we are able to tell what features are important at which level.

Identification of highly attended component using attention mechanism. Attention mechanism is widely applied in machine translation (Bahdanau et al., 2014), language inference (Rocktäschel et al., 2015), and question answering (Chen et al., 2016a). In addition, Yang et al. (2016b) propose hierarchical attention network to understand both attended words and sentences for sentiment classification. Chen et al. (2016b) enhance the attention with the support of user preference and product information to comprehend how user and product affect sentiment ratings. Due to the capability of attention mechanism, we employ the same strategy to show attended component for satirical news. Different from above works, we further evaluate linguistic features of highly attended paragraphs to analyze characteristics of satirical news, which has not been explored to our knowledge.

3 The Proposed Model

We first present our 4-level hierarchical neural network and explain how linguistic features can be embedded in the network to reveal the difference between paragraph level and document level. Then we describe the linguistic features.

3.1 The 4-Level Hierarchical Model

We build the model in a hierarchy of character-word-paragraph-document. The general overview of the model can be viewed in Figure 1 and the notations are listed in Table 2.

	Meaning
Superscript	Lowercase for notation purpose; $\top$ means matrix transpose.
Subscript	For index purpose.
Parameter	$\mathbf{W}$,$\mathbf{U}$,$\mathbf{w}^c$,$\mathbf{v}^a$: learnable weights; b: learnable bias.
Representation	$\mathbf{c}$: character; $\mathbf{x}$: word; $\mathbf{p}$: paragraph; $\mathbf{d}$: document; $\tilde{y}$: prediction $\mathbf{l}$: linguistic vector; y: label; $\mathbf{r}$: reset gate; $\mathbf{z}$: update gate; $\mathbf{h}$: hidden state for GRU; $\mathbf{u}$: hidden state for attention.

Table 2: Notations and meanings

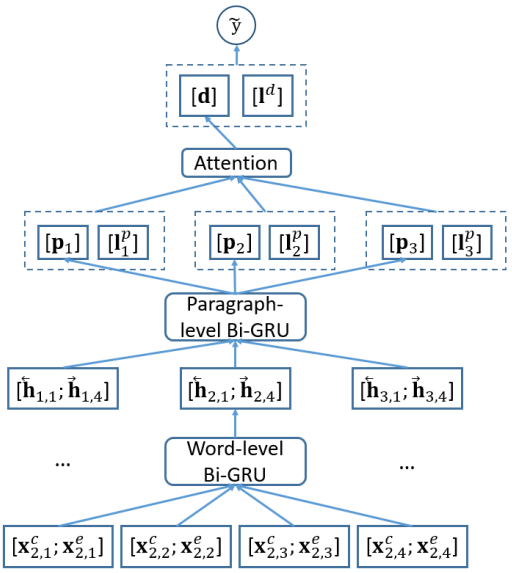

Figure 1: The overview of the proposed model. The document has 3 paragraphs and each paragraph contains 4 words. We omit character-level convolution neural network but leave $\mathbf{x}^c$ to symbolize the representation learned from it.

3.1.1 Character-Level Encoder

We use convolutional neural networks (CNN) to encode word representation from characters. CNN is effective in extracting morphological information and name entities (Ma and Hovy, 2016), both of which are common in news. Each word is presented as a sequence of n characters and each character is embedded into a low-dimension vector. The sequence of characters $\mathbf{c}$ is brought to the network. A convolution operation with a filter $\mathbf{w}^c$ is applied and moved along the sequence. Max pooling is performed to select the most important feature generated by the previous operation. The word representation $\mathbf{x}^c \in \mathbb{R}^f$ is generated with f filters.

3.1.2 Word-Level Encoder

Assume a sequence of words of paragraph i arrives at time t. The current word representation $\mathbf{x}_{i,t}$ concatenates $\mathbf{x}^c_{i,t}$ from character level with pre-trained word embedding $\mathbf{x}^e_{i,t}$, as $\mathbf{x}_{i,t} = [\mathbf{x}^c_{i,t}; \mathbf{x}^e_{i,t}]$. Examples are given in Figure 1. We implement Gated Recurrent Unit (GRU) (Cho et al., 2014) rather than LSTM (Hochreiter and Schmidhuber, 1997) to encode the sequence because GRU has fewer parameters. The GRU adopts reset gate $\mathbf{r}_{i,t}$ and update gate $\mathbf{z}_{i,t}$ to control the information flow between the input $\mathbf{x}_{i,t}$ and the candidate

state $\tilde{\mathbf{h}}_{i,t}$. The output hidden state $\mathbf{h}_{i,t}$ is computed by manipulating previous state $\mathbf{h}_{i,t-1}$ and the candidate state $\tilde{\mathbf{h}}_{i,t}$ regarding to $\mathbf{z}_{i,t}$ as in Equation 4, where $\odot$ denotes element-wise multiplication.

$$\mathbf{z}_{i,t} = \sigma(\mathbf{W}^z \mathbf{x}_{i,t} + \mathbf{U}^z \mathbf{h}_{i,t-1} + b^z) \quad (1)$$

$$\mathbf{r}_{i,t} = \sigma(\mathbf{W}^r \mathbf{x}_{i,t} + \mathbf{U}^r \mathbf{h}_{i,t-1} + b^r) \quad (2)$$

$$\tilde{\mathbf{h}}_{i,t} = \tanh(\mathbf{W}^h \mathbf{x}_{i,t} + \mathbf{r}_{i,t} \odot (\mathbf{U}^h \mathbf{h}_{i,t-1} + b^h)) \quad (3)$$

$$\mathbf{h}_{i,t} = (1 - \mathbf{z}_{i,t}) \odot \mathbf{h}_{i,t-1} + \mathbf{z}_{i,t} \odot \tilde{\mathbf{h}}_{i,t} \quad (4)$$

To learn a better representation from the past and the future, we use bidirectional-GRU (Bi-GRU) to read the sequence of words with forward $\overrightarrow{\text{GRU}}$ from $\mathbf{x}_{i,1}$ to $\mathbf{x}_{i,t}$, and backward $\overleftarrow{\text{GRU}}$ from $\mathbf{x}_{i,t}$ to $\mathbf{x}_{i,1}$. The final output of Bi-GRU concatenates the last state of $\overrightarrow{\text{GRU}}$ and $\overleftarrow{\text{GRU}}$, as $[\overrightarrow{\mathbf{h}}_{i,t}; \overleftarrow{\mathbf{h}}_{i,1}]$, to represent the ith paragraph.

3.1.3 Paragraph-Level Attention

We observe that not all paragraphs have satire and some of them are functional to make the article complete, so we incorporate attention mechanism to reveal which paragraphs contribute to decision making. Assuming a sequence of paragraph representations have been constructed from lower levels, another Bi-GRU is used to encode these representations to a series of new states $\mathbf{p}_{1:t}$, so the sequential orders are considered.

To decide how paragraphs should be attended, we calculate satirical degree α_i of paragraph i. We first convey $\mathbf{p}_i$ into hidden states $\mathbf{u}_i$ as in Equation 5. Then we product $\mathbf{u}_i$ with a learnable satire-aware vector $\mathbf{v}^a$ and feed the result into softmax function as in Equation 6. The final document representation $\mathbf{d}$ is computed as a weighted sum of α_i and $\mathbf{p}_i$.

$$\mathbf{u}_i = \tanh(\mathbf{W}^a \mathbf{p}_i + b^a) \quad (5)$$

$$\alpha_i = \frac{\exp(\mathbf{u}_i^\top \mathbf{v}^a)}{\sum_{j=0}^{t} \exp(\mathbf{u}_j^\top \mathbf{v}^a))} \quad (6)$$

$$\mathbf{d} = \sum_{i=0}^{t} \alpha_i \mathbf{p}_i \quad (7)$$

Linguistic features are leveraged to support attending satire paragraph. Besides $\mathbf{p}_i$, we represent paragraph i based on our linguistic feature set and transform it into a high-level feature vector $\mathbf{l}_i^p$ via multilayer perceptron (MLP). So $\mathbf{u}_i$ in Equation 5 is updated to:

$$\mathbf{u}_i = \tanh(\mathbf{W}^a \mathbf{p}_i + \mathbf{U}^a \mathbf{l}_i^p + b^a) \quad (8)$$

3.1.4 Document-Level Classification

Similar to the paragraph level, we represent document j based on our linguistic feature set and transform it into a high-level feature vector $\mathbf{l}_j^d$ via MLP. We concatenate $\mathbf{d}_j$ and $\mathbf{l}_j^d$ together for classification. Suppose $y_j \in (0, 1)$ is the label of the document j, the prediction $\tilde{y}_j$ and the loss function $\mathcal{L}$ over N documents are:

$$\tilde{y}_j = \text{sigmoid}(\mathbf{W}^d \mathbf{d}_j + \mathbf{U}^d \mathbf{l}_j^d + b^d) \quad (9)$$

$$\mathcal{L} = -\frac{1}{N} \sum_{j}^{N} y_j \log \tilde{y}_j + (1 - y_j) \log(1 - \tilde{y}_j) \quad (10)$$

3.2 Linguistic Features

Linguistic features have been successfully applied to expose differences between deceptive and genuine content, so we subsume most of the features in previous works. The idea of explaining fictitious content is extended here to reveal how satirical news differs from true news. We divide our linguistic features into four families and compute them separately for paragraph and document.

Psycholinguistic Features: Psychological differences are useful for our problem, because professional journalists tend to express opinion conservatively to avoid unnecessary arguments. On the contrary, satirical news includes aggressive language for the entertainment purpose. We additionally observe true news favors clarity and accuracy while satirical news is related to emotional cognition. To capture the above observations, we employ Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007) as our psycholinguistic dictionary. Each category of LIWC is one independent feature and valued by its frequency[3].

Writing Stylistic Features: The relative distribution of part-of-speech (POS) tags reflects informative vs. imaginative writing, which contributes to detecting deceptions (Li et al., 2014a; Mukherjee et al., 2013a). We argue that the stories covered by satirical news are based on imagination. In addition, POS tags are hints of the underlying

[3]Total counts divided by total words.

	#Train	#Validation	#Test	#Para	#Sent	#Words	# Capitals	#Punc	#Digits
True	101,268	33,756	33,756	20±7.8	32±24	734±301	118±58	28±26	93±49
Satire	9,538	3,103	3,608	12±4.4	25±12	587±246	87±44	11±13	86±43

Table 3: The split and the description (mean and standard deviation) of the dataset. Para denotes paragraphs, sent denotes sentences, and punc denotes punctuations.

humor (Reyes et al., 2012), which is common in satirical news. So we utilize POS tags (Toutanova et al., 2003) to apprehend satire. Each tag is regarded as one independent feature and valued by its frequency.

Readability Features: We consider readability of genuine news would differ from satirical news because the former is written by professional journalists and tend to be clearer and more accurate, while satirical news packs numerous clauses to enrich the made-up story as introduced by Rubin et al. (2016). Different from their work, we use readability metrics, including Flesch Reading Ease (Kincaid et al., 1975), Gunning Fog Index (Gunning, 1952), Automated Readability Index (Senter and Smith, 1967), ColemanLiau Index (Coleman and Liau, 1975), and syllable count per word, as features.

Structural Features: To further reflect the structure of news articles, we examine the following features: word count, log word count, number of punctuations, number of digits, number of capital letters, and number of sentences.

4 Experiment and Evaluation

We report satirical news detection results and show high weighted word features. Then, we provide a thorough analysis between paragraph-level and document-level features. Finally, we visualize an example of satirical news article to demonstrate the effectiveness of our work.

4.1 Dataset

The satirical news is collected from 14 websites that explicitly declare they are offering satire, so the correct label can be guaranteed. We also notice websites that mix true news, fake news, and satirical news. We exclude these websites in this work because it requires experts to annotate the news articles.

We maintain each satire source in only one of the train/validation/test sets[4] as the cross-domain setting in (Li et al., 2014a). Otherwise, the problem may become writing pattern recognition or news site classification. We also combined different sources together[5] as a similar setting of leveraging multiple domains (Yang et al., 2016a). The true news is collected from major news outlets[6] and Google News using FLORIN (Liu et al., 2015). The satirical news in the corpus is significantly less than true news, reflecting an impressionistic view of the reality. We omit headline, creation time, and author information so this work concentrates on the satire in the article body. We realize the corpus may contain different degree of satire. Without the annotation, we only consider binary classification in this work and leave the degree estimation for the future. The split and the description of the dataset can be found in Table 3.

4.2 Implementation Detail

For SVM, we use the sklearn implementation[7]. We find that using linear kernel and setting "class_weight" to "balanced" mostly boost the result. We search soft-margin penalty "C" and find high results occur in range $[10^{-1}, 10^{-4}]$. We use the validation set to tune the model so selecting hyper-parameters is consistent with neural network based model.

For neural network based models, we use the Theano package (Bastien et al., 2012) for implementation. The lengths of words, paragraphs, and documents are fixed at 24, 128, and 16 with necessary padding or truncating. Stochastic Gradient Descent is used with initial learning rate of 0.3 and decay rate of 0.9. The training is early stopped if the F1 drops 5 times continuously. Word embeddings are initialized with 100-dimension Glove embeddings (Pennington et al., 2014). Character embeddings are randomly initialized with 30 dimensions. Specifically for the proposed model, the following hyper-parameters are estimated based on the validation set and used

[4]Train: Onion, the Spoof. Test: SatireWorld, Beaverton, Ossurworld. Validation: DailyCurrent, DailyReport, EnduringVision, Gomerblog, NationalReport, SatireTribune, SatireWire, Syruptrap, and UnconfirmedSource.

[5]The combination is chosen to ensure enough training examples and balanced validation/test sets.

[6]CNN, DailyMail, WashingtonPost, NYTimes, TheGuardian, and Fox.

[7]sklearn.svm.SVC

Model	Validation				Test			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
SVM word n-grams	97.69	87.45	84.66	86.03	97.46	89.59	83.45	86.41
SVM word n-grams + LF	97.73	86.06	**87.14**	86.60	97.52	88.44	85.48	86.93
SVM word + char n-grams	97.43	87.10	81.57	84.24	97.64	90.76	84.12	87.31
SVM word + char n-grams + LF	97.76	90.13	82.44	86.11	97.93	92.71	85.31	88.86
SVM Rubin et al. (2016)	97.73	90.21	81.92	85.86	97.79	**93.47**	82.95	87.90
SVM Rubin et al. (2016) + char tf-idf + LF	**97.93**	**90.99**	83.69	**87.19**	**98.09**	92.98	**86.72**	**89.75**
Bi-GRU	97.67	89.17	82.28	85.58	97.58	93.11	80.96	86.61
SVM Doc2Vec Le and Mikolov (2014)	92.48	58.48	71.66	64.40	90.48	50.52	67.88	57.92
HAN Yang et al. (2016b)	97.91	92.06	82.24	86.88	97.83	90.85	86.17	88.45
4LHN	98.44	92.82	88.33	90.52	98.36	94.61	88.00	91.18
4LHNP	98.46	93.54	87.75	90.56	98.39	94.63	88.33	91.37
4LHND	98.36	**94.73**	85.24	89.74	98.18	**95.35**	85.31	90.05
4LHNPD	**98.54**	93.31	**89.01**	**91.11**	**98.39**	93.51	**89.50**	**91.46**

Table 4: Satirical news detection results.

in the final test set. The dropout is applied with probability of 0.5. The size of the hidden states is set at 60. We use 30 filters with window size of 3 for convolution.

4.3 Performance of Satirical News Detection

We report accuracy, precision, recall, and F1 on the validation set and the test set. All metrics take satirical news as the positive class. Both paragraph-level and document-level linguistic features are scaled to have zero mean and unit variance, respectively. The compared methods include:

SVM word n-grams: Unigram and bigrams of the words as the baseline. We report 1,2-grams because it performs better than other n-grams.

SVM word n-grams + LF: 1,2-word grams plus linguistic features. We omit comparison with similar work (Ott et al., 2011) as their features are subsumed in ours.

SVM word + char n-grams: 1,2-word grams plus bigrams and trigrams of the characters.

SVM word + char n-grams + LF: All the proposed features are considered.

SVM Rubin et al. (2016): Unigram and bigrams tf-idf with satirical features as proposed in (Rubin et al., 2016). We compare with (Rubin et al., 2016) rather than (Burfoot and Baldwin, 2009) as the former claims a better result.

SVM Rubin et al. (2016) + char tf-idf + LF: Include all possible features.

Bi-GRU: Bi-GRU for document classification. The document representation is the average of the hidden state at every time-step.

SVM Doc2Vec: Unsupervised method learning distributed representation for documents (Le and Mikolov, 2014). The implementation is based on Gensim (Řehůřek and Sojka, 2010).

HAN: Hierarchical Attention Network (Yang et al., 2016b) for document classification with both word-level and sentence-level attention.

4LHN: 4-Level Hierarchical Network without any linguistic features.

4LHNP: 4-Level Hierarchical Network with Paragraph-level linguistic features.

4LHND: 4-Level Hierarchical Network with Document-level linguistic features.

4LHNPD: 4-Level Hierarchical Network with both Paragraph-level and Document-level linguistic features.

In Table 4, the performances on the test set are generally better than on the validation set due to the cross-domain setting. We also explored word-level attention (Yang et al., 2016b), but it performed 2% worse than 4LHN. The result of Doc2Vec is limited. We suspect the reason could be the high imbalanced dataset, as an unsupervised learning method for document representation heavily relies on the distribution of the document.

4.4 Word Level Analysis

True		Satire	
:	day	"	stated
video	said the	sources	press
but the	twitter	continued	reporter
in statement	told the	added	resident
com	pictured	washington dc	said that

Table 5: High weighted word-level features

We report high weighted word-grams in Table 5 based on the SVM model as incorporating word-level attention in our neural hierarchy model reduces the detection performance. According

Psycholinguistic Feature					Writing Stylistic Feature					Readability Feature				
Name	S.m	S.std	T.m	T.std	Name	S.m	S.std	T.m	T.std	Name	S.m	S.std	T.m	T.std
Human.P	.011	.021	.009	.023	**JJ.P**	.061	.045	.058	.046	**FRE.D**	58.4	12.2	56.0	10.1
Past.P	.034	.035	.040	.042	**PRP.P**	.054	.047	.044	.047	**CLI.D**	9.08	1.66	9.48	1.61
Self.P	.017	.032	.010	.027	**RB.P**	.051	.048	.045	.054	**FOG.D**	13.71	3.25	14.00	2.89
Funct.D	.453	.045	.437	.049	**VBN.P**	.021	.026	.024	.031	Structural Feature				
Social.P	.097	.067	.091	.073	**NN.D**	.273	.038	.300	.043	**Punc.P**	7.69	5.35	4.69	3.83
Leisure.P	.017	.027	.018	.032	**VBZ.P**	.019	.026	.021	.029	**Cap.P**	7.44	6.08	5.75	4.8
Hear.P	.011	.019	.012	.021	**CC.P**	.023	.024	.024	.026	**Digit.P**	0.97	2.40	1.39	3.00
Bio.P	.026	.035	.023	.036	**CD.P**	.013	.027	.024	.043	**LogWc.P**	3.69	0.71	3.39	0.53

Table 6: Comparing feature values within each category. P stands for paragraph level. D stands for document level. S stands for satirical news. T stands for true news. m stands for mean and std stands for standard deviation. FRE: Flesch Reading Ease, the lower the harder. CLI: ColemanLiau Index. FOG: Gunning Fog Index. Punc: punctuation. Cap: Capital letters. LogWc: Log Word count

to Table 5, we conclude satirical news mimics true news by using news related words, such as "stated" and "reporter". However, these words may be over used so they can be detected. True news may use other evidence to support the credibility, which explains "twitter", "com", "video", and "pictured". High weight of " : " indicates that true news uses colon to list items for clarity. High weight of " " " indicates that satirical news involves more conversation, which is consistent with our observation. The final interesting note is satirical news favors "washington dc". We suspect that satirical news mostly covers politic topics, or satire writers do not spend efforts on changing locations.

4.5 Analysis of Weighted Linguistic Features

We use 4LHNPD to compare paragraph-level and document-level features, as 4LHNPD leverages the two-level features into the same framework and yields the best result.

Because all linguistic features are leveraged into MLP with non-linear functions, it is hard to check which feature indicates satire. Alternatively, we define the importance of linguistic features by summing the absolute value of the weights if directly connected to the feature. For example, the importance I of feature k is given by $\mathrm{I}_k = \frac{1}{M}\sum_{m=0}^{M}|\mathbf{w}_{k,m}|$, where $\mathbf{w} \in \mathbb{R}^{K\times M}$ is the directly connected weight, K is the number of features, and M is the dimension of the output. This metric gives a general idea about how much does a feature contribute to the decision making.

We first report the scaled importance of the four linguistic feature sets by averaging the importance of individual linguistic features. Then we report individual important features within each set.

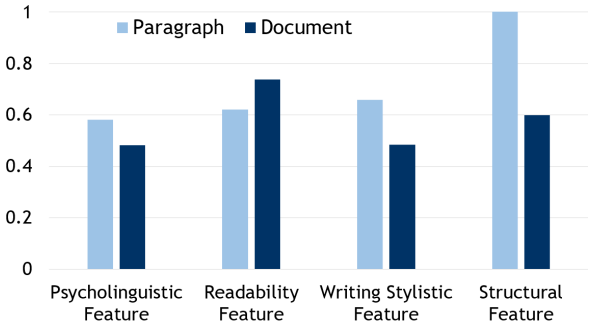

Figure 2: Comparing the importance of the four feature sets at paragraph level and document level.

4.5.1 Comparing the Four Feature Sets

According to Figure 2, the importance of paragraph-level features is greater than document-level features except for the readability feature set. It is reasonable to use readability at the document level because readability features evaluate the understandability of a given text, which depends on the content and the presentation. The structural feature set is highly weighted for selecting attended paragraph, which inspires us to focus on individual features inside the structural feature set.

4.5.2 Comparing Individual Features

Within each set, we rank features based on the importance score and report their mean and standard deviation before being scaled in Table 6. At paragraph level, we use top three attended paragraphs for calculating. The respective p-values of all features in the table are less than 0.01 based on the t-test, indicating satirical news is statistically significantly different from true news.

Comparing Table 6 and Table 3, we find that the word count, capital letters, and punctuations in true news are larger than in satirical news at the document level, while at paragraph level these

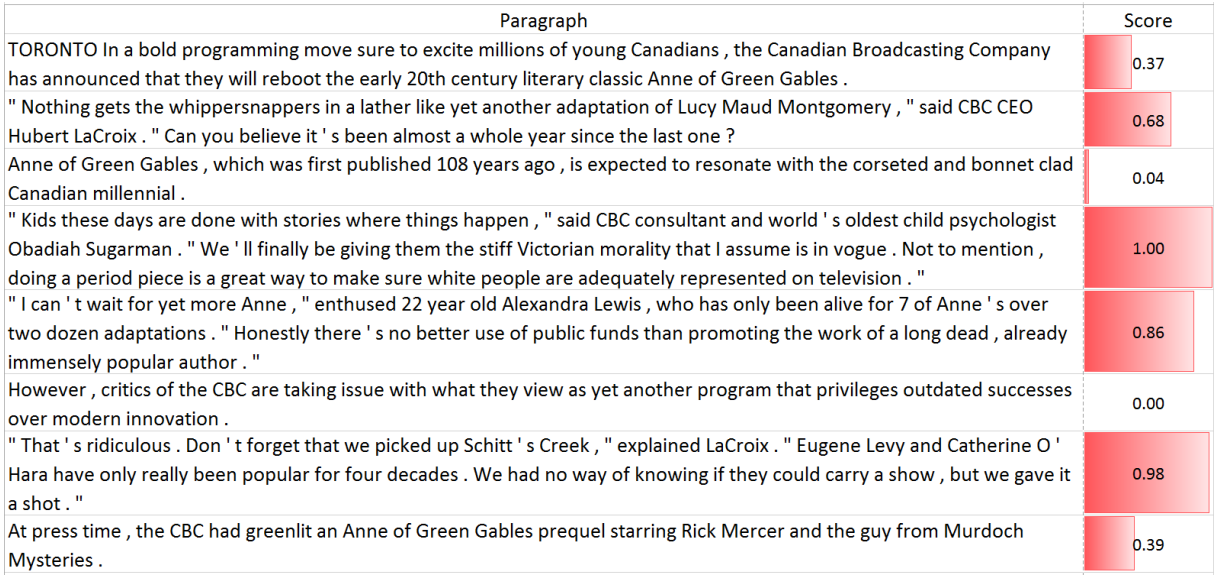

Figure 3: An example of attended paragraphs.

features in true news are less than in satirical news. This indicates satire paragraph could be more complex locally. It also could be referred as "sentence complexity", that *"satirical articles tend to pack a great number of clauses into a sentence for comedic effect"* (Rubin et al., 2016). Accordingly, we hypothesize top complex paragraphs could represent the entire satire document for classification, which we leave for future examination.

In Table 6, psycholinguistic feature "Humans" is more related to emotional writing than control writing (Pennebaker et al., 2007), which indicates satirical news is emotional and unprofessional compared to true news. The same reason also applies to "Social" and "Leisure", where the former implies emotional and the latter implies control writing. The "Past" and "VBN" both have higher frequencies in true news, which can be explained by the fact that true news covers what happened. A similar reason that true news reports what happened to others explains a low "Self" and a high "VBZ" in true news.

For writing stylistic features, it is suggested that informative writing has more nouns, adjectives, prepositions and coordinating conjunctions, while imaginative writing has more verbs, adverbs, pronouns, and pre-determiners (Rayson et al., 2001). This explains higher frequencies of "RB" and "PRP" in satirical news, and higher frequency of "NN" and "CC" in true news. One exception is "JJ", adjectives, which receives the highest weight in this feature set and indicates a higher frequency in satirical news. We suspect adjective could also be related to emotional writing, but more experiments are required.

Readability suggests satirical news is easier to be understood. Considering satirical news is also deceptive (as the story is not true), this is consistent with works (Frank et al., 2008; Afroz et al., 2012) showing deceptive writings are more easily comprehended than genuine writings. Finally, true news has more digits and a higher "CD"(Cardinal number) frequency, even at the paragraph level, because they tend to be clear and accurate.

4.6 Visualization of Attended Paragraph

To explore the attention, we sample one example in the validation set and present it in Figure 3. The value at the right represents the scaled attention score. The high attended paragraphs are longer and have more capital letters as they are referring different entities. They have more double quotes, as multiple conversations are involved.

Moreover, we subjectively feel the attended paragraph with score 0.98 has a sense of humor while the paragraph with score 0.86 has a sense of sarcasm, which are common in satire. The paragraph with score 1.0 presents controversial topics, which could be misleading if the reader cannot understand the satire. This is what we expect from the attention mechanism. Based on the visualization, we also feel this work could be generalized to detect figurative languages.

5 Conclusion

In this paper, we proposed a 4-level hierarchical network and utilized attention mechanism to understand satire at both paragraph level and document level. The evaluation suggests readability features support the final classification while psycholinguistic features, writing stylistic features, and structural features are beneficial at the paragraph level. In addition, although satirical news is shorter than true news at the document level, we find satirical news generally contain paragraphs which are more complex than true news at the paragraph level. The analysis of individual features reveals that the writing of satirical news tends to be emotional and imaginative.

We will investigate efforts to model satire at the paragraph level following our conclusion and theoretical backgrounds, such as (Ermida, 2012). We plan to go beyond the binary classification and explore satire degree estimation. We will generalize our approach to reveal characteristics of figurative language (Joshi et al., 2016), where different paragraphs or sentences may reflect different degrees of sarcasm, irony, and humor.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments. This work was support in part by the U.S. NSF grants 1546480 and 1527364.

References

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2015. Do we criticise (and laugh) in the same way? automatic detection of multi-lingual satirical news in twitter. In *IJCAI*, pages 1215–1221.

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. 2012. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*.

Clint Burfoot and Timothy Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 161–164. Association for Computational Linguistics.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016a. A thorough examination of the cnn/daily mail reading comprehension task. *arXiv preprint arXiv:1606.02858*.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016b. Neural sentiment classification with user and product attention. In *Proceedings of EMNLP*.

Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 15–19. ACM.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment*, 8(9):938–949.

Isabel Ermida. 2012. News satire in the press: Linguistic construction of humour inspoof news articles. *Language and humour in the media*, page 185.

Mehrdad Farajtabar, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha. 2017. Fake news mitigation via point process based intervention. *arXiv preprint arXiv:1703.07823*.

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

Mark G Frank, Melissa A Menasco, and Maureen O'Sullivan. 2008. Human behavior and deception detection. *Wiley Handbook of Science and Technology for Homeland Security*.

Liang Ge, Jing Gao, Xiaoyi Li, and Aidong Zhang. 2013. Multi-source deep learning for information trustworthiness estimation. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 766–774. ACM.

Robert Gunning. 1952. The technique of clear writing.

Manish Gupta, Peixiang Zhao, and Jiawei Han. 2012. Evaluating event credibility on twitter. In *SDM*, pages 153–164. SIAM.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Benjamin D Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *arXiv preprint arXiv:1703.09398*.

Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. 2016. Automatic sarcasm detection: A survey. *arXiv preprint arXiv:1602.03426*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196.

Jiwei Li, Myle Ott, Claire Cardie, and Eduard H Hovy. 2014a. Towards a general rule for identifying deceptive opinion spam. In *ACL (1)*, pages 1566–1576. Citeseer.

Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. 2014b. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1187–1198. ACM.

Xian Li, Weiyi Meng, and Yu Clement. 2016. Verification of fact statements with multiple truthful alternatives. In *12th International Conference on Web Information Systems and Technologies*.

Yaliang Li, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2015. On the discovery of evolving truth. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 675–684. ACM.

Qingyuan Liu, Eduard C Dragut, Arjun Mukherjee, and Weiyi Meng. 2015. Florin: a system to support (near) real-time applications on user generated content on daily news. *Proceedings of the VLDB Endowment*, 8(12):1944–1947.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312. Association for Computational Linguistics.

Arjun Mukherjee, Abhinav Kumar, Bing Liu, Junhui Wang, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013a. Spotting opinion spammers using behavioral footprints. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 632–640. ACM.

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie S Glance. 2013b. What yelp fake review filter might be doing? In *ICWSM*.

Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 353–362. ACM.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319. Association for Computational Linguistics.

James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. 2007. The development and psychometric properties of liwc2007. austin, tx, liwc. net.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Verónica Pérez-Rosas and Rada Mihalcea. 2014. Cross-cultural deception detection. In *ACL (2)*, pages 440–445.

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

Paul Rayson, Andrew Wilson, and Geoffrey Leech. 2001. Grammatical word class variation within the british national corpus sampler. *Language and Computers*, 36(1):295–306.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. 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, San Diego, California. Association for Computational Linguistics.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, DTIC Document.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Mengting Wan, Xiangyu Chen, Lance Kaplan, Jiawei Han, Jing Gao, and Bo Zhao. 2016. From truth discovery to trustworthy opinion discovery: An uncertainty-aware quantitative modeling approach. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1885–1894. ACM.

Houping Xiao, Jing Gao, Qi Li, Fenglong Ma, Lu Su, Yunlong Feng, and Aidong Zhang. 2016. Towards confidence in the truth: A bootstrapping based truth discovery approach. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1935–1944. ACM.

Fan Yang, Arjun Mukherjee, and Yifan Zhang. 2016a. Leveraging multiple domains for sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2978–2988, Osaka, Japan. The COLING 2016 Organizing Committee.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016b. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Xiaoxin Yin, Jiawei Han, and S Yu Philip. 2008. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1395–1405. ACM.