Rubin, V.R., Conroy, N. J., Chen, Y. & Cornwell, S. (2016) Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In the Proceedings of the Workshop on Computational Approaches to Deception Detection at the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-CADD2016), San Diego, California, June 17, 2016.

Fake News or Truth?
Using Satirical Cues to Detect Potentially Misleading News.

Victoria L. Rubin, Niall J. Conroy, Yimin Chen, and Sarah Cornwell
Language and Information Technology Research Lab (LIT.RL)
Faculty of Information and Media Studies
University of Western Ontario, London, Ontario, CANADA
vrubin@uwo.ca, nconroy1@uwo.ca, ychen582@uwo.ca, scornwel@uwo.ca

Abstract

Satire is an attractive subject in deception detection research: it is a type of deception that intentionally incorporates cues revealing its own deceptiveness. Whereas other types of fabrications aim to instill a false sense of truth in the reader, a successful satirical hoax must eventually be exposed as a jest. This paper provides a conceptual overview of satire and humor, elaborating and illustrating the unique features of satirical news, which mimics the format and style of journalistic reporting. Satirical news stories were carefully matched and examined in contrast with their legitimate news counterparts in 12 contemporary news topics in 4 domains (civics, science, business, and "soft" news). Building on previous work in satire detection, we proposed an SVM-based algorithm, enriched with 5 predictive features (*Absurdity, Humor, Grammar, Negative Affect,* and *Punctuation*) and tested their combinations on 360 news articles. Our best predicting feature combination (*Absurdity, Grammar* and *Punctuation*) detects satirical news with a 90% precision and 84% recall (F-score=87%). Our work in algorithmically identifying satirical news pieces can aid in minimizing the potential deceptive impact of satire.

1. Introduction

In the course of news production, dissemination, and consumption, there are ample opportunities to deceive and be deceived. Direct falsifications such as journalistic fraud or social media hoaxes pose obvious predicaments. While fake or satirical news may be less malicious, they may still mislead inattentive readers. Taken at face value, satirical news can intentionally create a false belief in the readers' minds, per classical definitions of deception (Buller & Burgoon, 1996; Zhou, Burgoon, Nuna-maker, & Twitchell, 2004). The falsehoods are intentionally poorly concealed, and beg to be unveiled. Yet some readers simply miss the joke, and the fake news is further propagated, with often costly consequences (Rubin, Chen, & Conroy, 2015).

1.1. The News Context

In recent years, there has been a trend towards decreasing confidence in the mainstream media. According to Gallup polls, only 40% of Americans trust their mass media sources to report the news "fully, accurately and fairly" (Riffkin, 2015) and a similar survey in the UK has shown that the most-read newspapers were also the least-trusted (Reilly & Nye, 2012). One effect of this trend has been to drive news readers to rely more heavily on alternative information sources, including blogs and social media, as a means to escape the perceived bias and unreliability of mainstream news (Tsfati, 2010). Ironically, this may leave the readers even more susceptible to incomplete, false, or misleading information (Mocanu, Rossi, Zhang, Karsai, & Quattrociocchi, 2015).

In general, humans are fairly ineffective at recognizing deception (DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997; Rubin & Conroy, 2012; Vrij, Mann, & Leal, 2012). A number of factors may explain why. First, most people show an inherent truth-bias (Van Swol, 2014): they tend to assume that the information they receive is true and reliable. Second, some people seem to exhibit a "general gullibility" (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2015) and are inordinately receptive to ideas that they do not fully understand. Third, confirmation bias can cause people to simply see only what they want to see – conservative viewers of the news satire program *the Colbert Report*, for example, tend to believe that the comedian's

statements are sincere, while liberal viewers tend to recognize the satirical elements (LaMarre, Landreville, & Beam, 2009).

1.2. Problem Statement

High rates of media consumption and low trust in news institutions create an optimal environment for the "rapid viral spread of information that is either intentionally or unintentionally misleading or provocative" (Howell, 2013). Journalists and other content producers are incentivized towards speed and spectacle over accuracy (Chen, Conroy, & Rubin, 2015) and content consumers often lack the literacy skills required to interpret news critically (Hango, 2014). What is needed for both content producers and consumers is an automated assistive tool that can save time and cognitive effort by flagging/filtering inaccurate or false information.

In developing such a tool, we have chosen news satire as a starting point for the investigation of deliberate deception in news. Unlike subtler forms of deception, satire may feature more obvious cues that reveal its disassociation from truth because the objective of satire is for at least some subset of readers to recognize the joke (Pfaff & Gibbs, 1997). And yet, articles from *The Onion* and other satirical news sources are often shared and even reprinted in newspapers as if the stories were true[1]. In other words, satirical news may mislead readers who are unaware of the satirical nature of news, or lacking in the contextual or cultural background to interpret the fake as such. In this paper, we examine the textual features of satire and test for the most reliable cues in differentiating satirical news from legitimate news.

2. Literature Review

2.1. Satire

As a concept, satire has been remarkably hard to pin down in the scholarly literature (Condren, 2012). One framework for humor, proposed by Ziv (1988), suggests five discrete categories of humor: aggressive, sexual, social, defensive, and intellectual. Satire, according to Simpson (2003), is complicated because it occupies more than one place in the framework: it clearly has an aggressive and social function, and often expresses an intellectual aspect as well. From this, satire can be conceptualized as "a rhetorical strategy (in any medium) that seeks wittily to provoke an emotional and intellectual reaction in an audience on a matter of public … significance" (Phiddian, 2013).

On the attack, satirical works use a variety of rhetorical devices, such as hyperbole, absurdity, and obscenity, in order to shock or unease readers. Traditionally, satire has been divided into two main styles: Juvenalian, the more overtly hostile of the two, and Horatian, the more playful (Condren, 2014). Juvenalian satire is characterized by contempt and sarcasm and an often aggressively pessimistic worldview (e.g., Swift's *A Modest Proposal*). Horatian satire, by contrast, tends more towards teasing, mockery, and black humor (e.g., Kubrick's *Dr. Strangelove*). In each, there is a mix of both laughter and scorn – though Juvenalian tends towards scorn, some hint of laughter must be present and vice versa for Horatian.

Satire must also serve a purpose beyond simple spectacle: it must aspire "to cure folly and to punish evil" (Highet, 1972: 156). It is not enough to simply mock a target; some form of critique or call to action is required. This "element of censoriousness" or "ethically critical edge" (Condren, 2012: 378) supplies the social commentary that separates satire from mere invective. However, the receptiveness of an audience to satire's message depends upon a level of "common agreement" (Frye, 1944: 76) between the writer and the reader that the target is worthy of both disapproval and ridicule. This is one way that satire may miss its mark with some readers: they might recognize the satirist's critique, but simply disagree with his or her position.

As a further confounding factor, satire does not speak its message plainly, and hides its critique in irony and double-meanings. Though satire aims to attack the folly and evil of others, it also serves to highlight the intelligence and wit of its author. Satire makes use of opposing scripts, text that is compatible with two different readings (Simpson, 2003: 30), to achieve this effect. The incongruity between the two scripts is part of what makes satire funny (e.g., when Stephen Colbert, states "I give people the truth, unfiltered by rational argument."[2]), but readers who fail to grasp the humor become, themselves, part of the joke. In this way, we consider satire a type of deception, but one that is intended to be found out by at least some subset of the audience.

1 See: https://www.washingtonpost.com/news/worldviews/wp/2015/06/02/7-times-the-onion-was-lost-in-translation

2 2006 White House Correspondents Dinner: https://youtu.be/2X93u3anTco

Although "the author mostly assumes that readers will recover the absurdity of the created text, which hopefully will prompt the readers to consider issues beyond the text" (Pfaff & Gibbs, 1997), what one reader considers absurd might be perfectly reasonable to another.

In his *Anatomy of Satire*, Highet distinguishes satire from other forms of "lies and exaggerations intended to deceive" (1972: 92). Whereas other types of fabrications and swindles aim to instill a false sense of truth in the reader, which benefits the deceiver for as long as it remains undiscovered, a successful satirical hoax must eventually be exposed. After all, an author of satire cannot be appreciated for his or her wit if no one recognizes the joke. This interpretation of satire is in line with Hopper & Bell's (1984) category of "benign fabrications," which include make believe, jokes, and teasing – types of deception that are generally both socially acceptable and fairly easy to uncover.

2.2. Satire in News

News satire is a genre of satire that mimics the format and style of journalistic reporting. The fake news stories are typically inspired by real ones, and cover the same range of subject matter: from politics to weather to crime. The satirical aspect arises when the factual basis of the story is "comically extended to a fictitious construction where it becomes incongruous or even absurd, in a way that intersects entertainment with criticism" (Ermida, 2012: 187). News satire is most often presented in the Horatian style, where humor softens the impact of harshness of the critique – the spoonful of sugar that helps the medicine go down. More than mere lampoon, fake news stories aim to "arouse the readers' attention, amuse them, and at the same time awaken their capacity to judge contemporary society" (Ermida, 2012: 188).

With the rise of the internet, news satire sites such as *The Onion* have become a prolific part of the media ecosystem. Stories from satirical sources are frequently shared over social media, where they deceive at least some of their readers. Indeed, people are fooled often enough that internet sites such as Literally Unbelievable (Hongo, 2016) have sprung up to document these instances.

Several factors contribute to the believability of fake news online. Recent polls have found that only 60% of Americans read beyond the headline (The Media Insight Project, 2014). Furthermore, on social media platforms like Facebook and Twitter, stories which are "liked" or "shared" all appear in a common visual format. Unless a user looks specifically for the source attribution, an article from *The Onion* looks just like an article from a credible source, like *The New York Times*. In an effort to counteract this trend, we propose the creation of an automatic satire detection system.

So, how can satirical news stories be identified? Ermida (2012: 194-195) proposes the model of parodic news satire in Figure 1.

COMPONENTS OF PARODIC NEWS SATIRE

I	II	III
INTERTEXTUAL COMPONENT	CRITICAL COMPONENT	COMIC COMPONENT
I.a) Structural component		III.a) Lexical component
I.b) Stylistic component		III.b) Pragmatic component
		III.c) Rhetorical component

Figure 1: Ermida's (2012) model of satirical news.

For the purposes of this research, we focus on component III to inform our investigation into cues to differentiate news satire from legitimate news reporting. The next section overviews satirical news detection efforts to date and positions them as beneficial in the deception detection research.

3. Detection Methodology Review

The methods described in this review demonstrate promising results for satire and humor detection. The goal of screening legitimate news content is achieved based on the assumption that successful identification of satire is independent from both the originating source of the piece, and its provenance as a news document. Instead, Natural Language Processing (NLP) methods in combination with machine learning deal with content directly by detecting language patterns, topicality, sentiment, rhetorical devices and word occurrences which are common to satire and irony. There is a need for a unified approach that combines best practices for a comprehensive NLP satire detection system.

3.1. Word Level Features

Burfoot & Baldwin (2009) attempted to determine whether or not newswire articles can be automatically classified as satirical. The approach relied on

lexical and semantic features such as headlines, profanity, or slang, as well as support vector machines (SVMs) on simple bag-of-words features which were supplemented with feature weighting. Similar attempts have used corpus-based relatedness which uses cosine similarity and tf*idf weighting. At the granularity of individual words and *n*-grams, text cues can point to the presence of satire, for example counterfactual words ("nevertheless", "yet"), and temporal compression words ("now", "abruptly") (Mihalcea, Strapparava, & Pulman, 2010). As a base measure, tf*idf on bi-grams provided F-score of roughly 65%. The use of additional features derived from semantic analysis heuristics improved classifier performance on joke detection. Using combined joke specific features led to 84% precision, demonstrating synergistic effect of feature combination.

We hypothesize expanding the possibilities of word-level analysis by measuring the utility of features like part of speech frequency, and semantic categories such as generalizing terms, time/space references, positive and negative polarity. In addition, we incorporate the use of exaggerated language (e.g., profanity, slang, grammar markers) and frequent run-on sentences. We take these features as indicators of satirical rhetoric component (III.c, per Ermida, 2012).

3.2. Semantic Validity

Satirical news contains a higher likelihood of imbalances between concepts expressing temporal consistency, as well as contextual imbalances (Reyes, Rosso, & Buscaldi, 2012), for example, well known people in unexpected settings. A similar idea of unexpectedness was explored by Burfoot and Baldwin (2009) who measured the presence of absurdity by defining a notion of "semantic validity": true news stories found on the web will contain differences in the presence of co-occurring named entities than those entities found in satire. Shallow and deep linguistic layers may represent relevant information to identify figurative uses of language. For example, ontological semantics such as ambiguity and incongruity, and meta-linguistic devices, such as polarity and emotional scenarios can achieve precision accuracy of 80% for classifying humorous tweets (Reyes et al., 2012). Semantically disparate concepts can also influence absurdity, when judged based on semantic relatedness (distance in WordNet hierarchy), summed and normalized (Reyes & Rosso, 2014). In this study, we use this measure of absurdity as an indicator of the pragmatic component (II.b) of satirical news (Ermida, 2012).

3.3. Humor and Opposing Scripts

Sjöbergh and Araki (2007) presented a machine learning approach for classifying sentences as one-liner jokes or normal sentences. Instead of deep analysis, they rely on weighting derived from a combination of simple features like word overlap with a joke corpus, and ambiguity (number of *dictionary.com* word senses), achieving an 85% accuracy. The incongruity (resolution) theory is a theory of comprehension which depends on the introduction of "latent terms". Humor is found when two scripts overlap and oppose. As a joke narration evolves, some "latent" terms are gradually introduced, which set the joke on a train of thought. Yet because of ambiguous quality of the terms, the humorous input advances on two or more interpretation paths. The interpretation shifts suddenly from the starting point of the initial sequence:

> *"Is the doctor at home?" the patient asked in a whisper. "No", the doctor's pretty wife whispered back, "Come right in."* (Attardo, Hempelmann, & Mano, 2002: 35)

The latter path gains more importance as elements are added to the current interpretation of the reader, and eventually ends up forming the punch line of the joke (Hempelmann, Raskin, & Triezenberg, 2006). To detect opposing scripts, Polanyi and Zaenen (2006) suggest a theoretical framework in which the context of sentiment words shifts the valence of the expressed sentiment. Hempelmann et al. (2006) employed Text Meaning Representations (TMRs) which are data models aimed at preserving vagueness in language while using an ontology representation from a fact repository, a pre-processor, and an analyzer to transform text. They propose an identification method by checking the number of word senses of the last word that "make sense" in the current context, although no performance evaluation is provided.

Other means of joke classification rely on naive Bayes and SVM based on joke-specific features, including polysemy and Latent Semantic Analysis (LSA) trained on joke data, as well as semantic relatedness. Mihalcea and Liu (2006) and Mihalcea and Pulman (2007) presented such a system that relies on metrics including knowledge-based similar-

ity (path between concepts) and corpus based similarity (term co-occurrence from large corpus). They conclude that the most frequently observed semantic features are negative polarity and human-centeredness. A correct punch line, which generates surprise, has a minimum relatedness with respect to the set-up. The highest overall precision of 84% was obtained with models that rely on joke-specific features, with the LSA model trained on a jokes corpus (Mihalcea et al., 2010). These methods informed our investigation of Ermida's lexical component (III.a) (2012).

Past research in humor detection provides useful heuristics for NLP. A shortcoming is that best practices have yet to be combined in a comprehensive detection methodology. For example, punchline detection may be employed to longer, discourse layer content beyond mere one liners, through the comparison of constituent text segments. Absurdity, measured through the presence of atypical named entities, may be extended to other contexts. Our examination of satire sources hints at the tendency to introduce new, unfamiliar named entities at the end of news articles as a form of ironic non-sequitur.

3.4. Recognizing Sarcasm and Irony

Recognition of sarcasm can benefit many sentiment analysis applications and the identification of fake news. Sarcasm is defined as "verbal irony … the activity of saying or writing the opposite of what you mean" (Tsur, Davidov, & Rappoport, 2010). Utsumi (2000) introduced a cognitive computational framework that models the ironic environment from an axiomatic system depending heavily on "world knowledge" and expectations. It requires analysis of each utterance and its context to match predicates in a specific logical formalism. Davidov, Tsur, and Rappoport (2010) looked for elements to automatically detect sarcasm in online products reviews, achieving a 77% precision and 83% recall. They proposed surface features (information about the product, company, title, etc.), frequent words or punctuation marks, to represent sarcastic texts. Ironic expressions often use such markers to safely realize their communicative effects (e.g., '*'Trees died for this book?"* - book review; "*All the features you want. Too bad they don't work!*" - smart phone review). Beyond grammar and word polarity, emotional scenarios capture irony in terms of elements which symbolize abstractions such as overall sentiments and moods. Presence of humor may be correlated to polarity of positive/negative semantic orientation and emotiveness. Using Twitter content, models of irony detection were assessed along these linguistic characteristics, and positive results provide valuable insights into figurative speech in the task of sentiment analysis (Reyes & Rosso, 2014; Reyes, Rosso, & Veale, 2013).

4. Methodology

4.1. Dataset and Data Collection Methods

In this study we collected and analyzed a dataset of 360 news articles as a wide-ranging and diverse data sample, representative of the scope of US and Canadian national newspapers. The dataset was collected in 2 sets. The first set was collected from 2 satirical news sites (*The Onion* and *The Beaverton*) and 2 legitimate news sources (*The Toronto Star* and *The New York Times*) in 2015. The 240 articles were aggregated by a 2 x 2 x 4 x 3 design (US/Canadian; satirical/legitimate online news; varying across 4 domains (civics, science, business, and "soft" news) with 3 distinct topics within each of the 4 domains (see Table 1).

CIVICS	SCIENCE	BUSINESS	"SOFT" NEWS
Gun Violence	Environment	Tech	Celebrity
Immigration	Health	Finance	Sports
Elections	Other Sciences	Corporate Announcements	Local News

Table 1: Sample News Topicality. 5 Canadian and 5 American satirical and legitimate article pairs were collected on 12 topics across 4 domains.

For each of the 12 topics, 5 Canadian (from *The Beaverton*) and 5 American (from *the Onion*) satirical articles were collected. Each satirical piece was then matched to a legitimate news article that was published in the same country, and as closely related in subject matter as possible. For example, in the Environment topic, the *Beaverton* article "Invasive homo sapiens species meet at forestry conference to discuss pine beetles" was paired with a *Toronto Star* article on invasive species: "'Dangerous and inva-

sive' Khapra beetle intercepted at Pearson". See Figure 2 for the pairing of the articles about Hillary Clinton in the Elections topic.

Figure 2: Two news articles about Hillary Clinton: from the Onion and The New York Times.

An additional set of 120 articles was collected in 2016 to expand the inventory of sources and topics, and to serve as a reliability test for the manual findings within the first set. The second set, still evenly distributed between satirical and legitimate news, was drawn from 6 legitimate[3] and 6 satirical[4] North American online news sources.

Analysis: A trained linguist content-analyzed each pair (legitimate vs. satirical), looking for insights on similarities and differences as well as trends in language use and rhetorical devices. For machine learning we used the combined set of 360, reserving random 25% of the combined 2 sets data for testing, and performing 10-fold cross-validation on the training set. The complete dataset is available from the lab's public website[5].

5. Results

5.1. Data-Driven Linguistic Observations

Absurdity and Humor: Similar to Burfoot & Baldwin (2009), we found that the headlines were especially relevant to detecting satire. While legitimate news articles report new material in the first line, satirical articles tend to repeat material from the title. We also found the final line of each article relevant to satire detection. In particular, the final line was commonly a "punchline" that highlighted absurdities in the story or introduced a new element to the joke. This humorous function diverges sharply from the standard "inverted pyramid" article structure of legitimate news articles, and it proved useful in identifying satirical journalism.

These observations informed the selection of 2 features: *Absurdity* and *Humor.* For example, the final line of *The Beaverton*'s "Scientists at University of the Lord discover that Jesus is Lord" introduces new named entities (indicating absurdity) and is semantically dissimilar (indicating humor) from the remainder of the article:

> *"At press time, researchers from Christopher Hitchens Memorial University discovered that it was fun to drink a lot of Johnny Walker Red Label and call people sheep."*

We also observed a high frequency of slang and swear words in the satirical news pieces, but unlike Burfoot and Baldwin (2009), for our dataset these features did not add predictive powers.

Sentence Complexity: A noticeable syntactic difference between the satirical and legitimate articles was sentence length and complexity. Especially for quotations, the satirical articles tend to pack a greater number of clauses into a sentence for comedic effect. For example, compare these two excerpts - the first from *The Onion*, and the second from its paired article in *The New York Times*:

> *(1) "Not too long ago, these early people were alive and going about their normal daily lives, but sadly, by the time we scaled down the narrow 90-meter chute leading into the cave, they'd already been dead for at least 10,000 decades," said visibly upset University of the Witwatersrand paleoanthropologist Lee R. Berger, bemoaning the fact that they could have saved the group of human predecessors if they had just reached the Rising Star cave system during the Pleistocene epoch." - The Onion* "Tearful Anthropologists Discover Dead Ancestor of Humans 100,000 Years Too Late".
>
> *(2) "With almost every bone in the body represented multiple times, Homo naledi is already practically the best-known fossil member of our lineage," Dr. Berger said." - The New York Times* "Homo Naledi, New Species in Human Lineage, Is Found in South African Cave".

*The Onion'*s quotation is 3 times longer; it does not just quote Dr. Berger, but also describes his motive and emotional state. The greater number of clauses increases the number of punctuation marks found in satire, which informed the development of

3 The Globe and Mail, The Vancouver Sun, Calgary Herald, National Post, The Edmonton Journal, The Hamilton Spectator, USA Today, The Wall Street Journal, Los Angeles Times, The New York Post, Newsday, and The Denver Post.

4 www.cbc.ca/punchline, thelapine.ca, syruptrap.ca, www.thecodfish.com, urbananomie.com, www.newyorker.com/humor/borowitz-report, dailycurrant.com, www.thespoof.com, nationalreport.net, worldnewsdailyreport.com, and thepeoplescube.com.

5 http://victoriarubin.fims.uwo.ca/news-verification/

our successful *Punctuation* feature. Our *Grammar* feature set incorporates the complexity of phrasing.

5.2. Our Satirical Detection Approach and News Satire Features

Based on previous methodological advances in irony, humor, and satire detection, and our data-driven linguistic observations, we propose and test a set of 5 satirical news features: *Absurdity, Humor, Grammar, Negative Affect* and *Punctuation*. The method begins with performing a topic-based classification followed by sentiment-based classification, and feature selection based on absurdity and humor heuristics. The training and evaluation of our model uses a state-of-the-art method of support vector machines (SVMs) using a 75% training and 25% test split of the dataset, and 10-fold cross validation applied to the training vectors. We combined cross-validation with the holdout method in reporting overall model performance. Cross validation on our 75% training produced a performance prediction on incoming data. We then confirmed this prediction in a second stage using our 25% hold out testing set. This allowed us to investigate which records from the set were incorrectly predicted by the model.

Text Processing and Feature Weighting: The text classification pipeline was scripted in Python 2.7 and used the *scikit-learn* open source machine learning package (Pedregosa et al., 2011) as the primary SVM classification and model evaluation API. In our approach, we described news articles as sparse feature vectors using a topic-based classification methodology with the term frequency-inverse document frequency (tf*idf) weighting scheme. The baseline results for our news corpus was achieved using a linear kernel classifier (Pedregosa et al., 2011) that assigns positive instances to satire. First, news article text was pre-processed, transforming the raw text to a Pandas data structure for use in Python. Stop words were removed, and unigrams and bigrams were tokenized. Both the training and test data were converted to tf-idf feature vectors. Term frequency values were also normalized by article length to account for length variability between satirical and legitimate news articles. The process is implemented as a semi-automated pipeline, summarized in Figure 3.

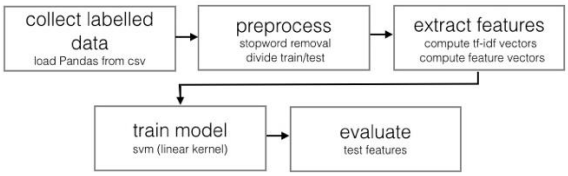

Figure 3: News satire detection pipeline for distinguishing satirical from legitimate news.

Feature Selection: ***The Absurdity Feature (Abs)*** was defined by the unexpected introduction of new named entities (people, places, locations) within the final sentence of satirical news. To implement *Absurdity* detection we used the Natural Language Toolkit[6] (NLTK) Part of Speech tagger and Named Entity Recognizer to detect the list of named entities. We defined the list as the non-empty set (LNE), and compared this with the set (NE) of named entities appearing in the remaining article. The article was deemed absurd when the intersection (LNE ∩ NE) was empty (0=non-absurd, 1=absurd).

Humor (Hum) detection was based on the premises of opposing scripts and maximizing semantic distance between two statements as method of punchline identification (Mihalcea et al., 2010). Similarly, in a humorous article, the lead and final sentence are minimally related. Our modification of the punchline detection method assigned the binary value (humor=1) when the relatedness between the first and last article sentences was the minimum with respect to the remaining sentences.

We used a knowledge-based metric to measure the relatedness between statement pairs in the article and sought to minimize relatedness of the lead and last line. Given a metric for word-to-word relatedness, we define the semantic relatedness of two text segments S1 and S2 using a metric that combines the semantic relatedness of each text segment in turn with respect to the other text segment. For each word *w* in the segment S1 we identified the word in the segment S2 that has the highest semantic relatedness, as per (Wu & Palmer, 1994) word-to-word similarity metric. The depth of two given concepts in the WordNet taxonomy, and the depth of the least common subsumer (LCS) were combined into a similarity score (Wu & Palmer, 1994). The same

6 http://www.nltk.org/

process was applied to find the most similar word in S1 starting with words in S2. The word similarities were weighted, summed, and normalized with the length of each text segment. The resulting relatedness scores were averaged.

Grammar (Gram) feature vector was the set of normalized term frequencies matched against the Linguistic Inquiry and Word Count (LIWC) 2015 dictionaries, which accounts for the percentage of words that reflect different linguistic categories (Pennebaker, Boyd, Jordan, & Blackburn, 2015). We counted the presence of parts of speech terms including adjectives, adverbs, pronouns, conjunctions, and prepositions, and assigned each normalized value as the element in a feature array representing grammar properties.

Negative Affect (Neg) and ***Punctuation (Pun)*** were assigned as feature weights representing normalized frequencies based on term-for-term comparisons with LIWC 2015 dictionaries. Values were assigned based on the presence of negative affect terms and punctuation (periods, comma, colon, semi-colon, question marks, exclamation, quotes) in the training and test set. Features representing *Absurdity, Humor, Grammar, Negative Affect* and *Punctuation* were introduced in succession to train the model, combined overall. The predictive performance was measured at the introduction of each new feature and the best performing features were combined and compared to the overall performance of all 5.

6. Evaluation

We conducted multiple experiments to identify the best-performing combination of features for satirical news identification. We used scikit-learn (Pedregosa et al., 2011) and the *tf***idf* F-measure as the baseline (***Base***, Table 2), with features added incrementally. Scikit-learn library contains several tools designed for machine learning applications in Python[7], and has been utilized in the supervised learning applications of real and fake news detection (Pisarevskaya, 2015; Rubin & Lukoianova, 2014). The *Sklearn.svm* package is a set of supervised learning methods used for classification, regression and outlier detection, capable of performing multi-class classification. We assigned two classes: satirical news (1) and legitimate news (0), and used *Sklearn.svm.SVC* (Support Vector Classification) for supervised training with a linear kernel algorithm, which is suitable for 2 class training data. Our model was trained on 270 and tested on a set of 90 news articles, with equal proportions of satirical and legitimate news. Table 2 presents the measures of precision, recall, and F-score with associated 10-fold cross validation confidence results for our satire detection model. The F-score was maximized in the case when *Grammar, Punctuation* and *Absurdity* features were used. Precision was highest when *Punctuation* and *Grammar* were included. *Absurdity* showed the highest recall performance.

FEATURES	PRECISION	RECALL	F-SCORE	CONFIDENCE
Base (*tf-idf*)	78	87	82	85
Base+Abs	85	**89**	87	83
Base +Hum	80	87	83	83
Base+Gram	**93**	82	87	82
Base+Neg	81	84	83	84
Base+Pun	**93**	82	87	87
Base+Gram+Pun+Abs	**90**	84	87	84
Base+All	88	82	87	87

Table 2: Satirical news detection evaluation results with 10-fold cross-validation. [**Legend: Base**= baseline tf-idf topic vector; **Abs**= Absurdity [0,1]; Hum= Humor [0,1]; **Gram**= Grammar features (pronoun, preposition, adjective, adverb, conjunction); Neg = Negative Affect; **Pun**= Punctuation; All= all features combined.]

7. Discussion

Our findings produced a set of observations about the benefits of detection methods for satire in news, as well as what methods were not useful and how they can be improved. We were able to integrate word level features using an established machine learning approach in text classification, SVM (Burfoot & Baldwin, 2009), to derive empirical cues indicative to deception. This included establishing a baseline model in the form of tf-idf feature weights, while measuring net effects of additional features against the baseline. Our baseline model performed with 82% accuracy, an improvement on Mihalcea et al. (2010) who achieved a 65% baseline score with tf-idf method on similar input data. Contrary to our expectations, we discovered that individual textual features of shallow syntax (parts of speech) and punctuation marks are highly indicative of the presence of satire, producing a detection improvement

7 http://scikit-learn.org

of 5% (87% F-score). This suggests that the rhetorical component of satire may provide reliable cues to its identification. Based on our manual analysis, this finding may be due to the presence of more complex sentence structures (prevalence of dependent clauses) in satirical content, and strategic use of run-on sentences for comedic affect. However, this pattern did not translate to longer sentences per se, since our average words-per-sentence feature did not increase predictive accuracy. Also contrary to our expectation, markers such as profanity and slang, word level features deemed significant by Burfoot and Baldwin (2009), produced no measurable improvements in our trials. Our dataset may not have included more extreme examples of satire.

One major contribution of the current research is that we were able to integrate a heuristic for *Absurdity* feature *(Abs*, Table 2*)* derived from the concept of *semantic validity*, the stark and strategic use of mismatched named entities in a story as a comic device. Burfoot and Baldwin's (2009) method of translating search results on co-occurring entities to an absurdity measure informed our approach, when we noticed that satirical sources often introduce previously non-occurring entities in the final sentence. Compared to Burfoot and Baldwin's 65% performance, our results show a 5% improvement (F-score 87%), when we added this feature to our baseline. Our intuition about named entities proved to be a defining empirical feature of satirical news.

Another concept derived from the theoretical literature on humor is the idea of shifting reference frames, and incongruity resolution based on the semantic derivation of textual components. We adapted methods based on identifying latent terms (Sjöbergh and Araki, 2007) in punchline detection against a setup statement. Instead of relying on more complex methods of representing longer text (through TMRs per Hempelmann et al. (2006), we modified the semantic distance approach between the lead sentence (setup) and the last sentence (punchline) which is hypothesized to be maximized in satire. The results of Mihalcea and Pulman (2007) showed a performance of 84% using word-wise semantic distance in WordNet classification. Our model showed a comparable performance of 83% when this feature was added to the baseline, opening up another avenue of extending opposing scripts to a news corpus.

The presence of polarity in satire has been noted in previous methods (Reyes et al., 2013); such as indicated by positive or negative semantic orientation, emotiveness, and emotional words. Our findings partially bolster this conclusion when we demonstrated that features representing negative affect improved the performance to 83% in the identification task. However, we found no similar improvement when we measured the contrasting effect of positive semantic orientation.

8. Conclusions & Future Work

In this paper, we have translated theories of humor, irony, and satire into a predictive method for satire detection that reaches relatively high accuracy rates (90% precision, 84% recall, 87% F-score). Since satirical news is at least part of the time deceptive, identifying satirical news pieces can aid in minimizing the potential deceptive impact of satirical news. By analyzing the current news production landscape captured within our dataset, we demonstrate the feasibility of satire detection methods even when divorced from attribution to a satirical source. Our conceptual contribution is in linking deception detection and computational satire, irony and humor research. Practically, this study frames fake news as a worthy target for filtering due to its potential to mislead news readers. Areas of further investigation can include ways to translate more complex characteristics of the anatomy of satire into linguistic cues. Critique and call to action, combined with mockery, is a key component of satire, but this critical component (II, Ermida, 2012) has not yet received much attention in the field of NLP. This feature could be subject to automated discourse-level quantification through the presence of imperative sentences. Also, our positive results of shallow syntax features showed us that more complex language patterns, for example deep syntax and the ordering of grammatical patterns might also be detectable in satire through techniques such as regular expression pattern matching against a grammatical parse of article content.

Acknowledgments

This research has been funded by the Government

of Canada Social Sciences and Humanities Research Council (SSHRC) Insight Grant (#435-2015-0065) awarded to Dr. Rubin for the project *Digital Deception Detection: Identifying Deliberate Misinformation in Online News.*

References

Attardo, S., Hempelmann, C. F., & Mano, S. D. (2002). Script oppositions and logical mechanisms: Modeling incongruities and their resolutions. *Humor, 15*(1), 3-46.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal Deception Theory. *Communication Theory, 6*(3), 203-242.

Burfoot, C., & Baldwin, T. (2009). *Automatic satire detection: are you having a laugh?* Paper presented at the Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Suntec, Singapore.

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). *News in an online world: the need for an automatic crap detector.* Paper presented at the Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community.

Condren, C. (2012). Satire and definition. *Humor, 25*(4), 375. doi:10.1515/humor-2012-0019

Condren, C. (2014). Satire. In S. Attardo (Ed.), *Encyclopedia of Humor Studies*. Thousand Oaks: Sage Publications, Inc.

Davidov, D., Tsur, O., & Rappoport, A. (2010). *Semi-supervised recognition of sarcastic sentences in twitter and amazon.* Paper presented at the Fourteenth Conference on Computational Natural Language Learning, Uppsala, Sweden.

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The Accuracy-Confidence Correlation in the Detection of Deception. *Personality and Social Psychology Review, 1*(4), 346-357.

Ermida, I. (2012). News Satire in the Press: Linguistic Construction of Humour in Spoof News Articles. In J. Chovanec & I. Ermida (Eds.), *Language and humour in the media*. Newcastle upon Tyne, UK: Cambridge Scholars Pub.

Frye, N. (1944). The Nature of Satire. *University of Toronto Quarterly, 14*(1), 75-89.

Hango, D. (2014). *University graduates with lower levels of literacy and numeracy skills*. Statistics Canada Retrieved from http://www.statcan. gc.ca/pub/75-006-x/2014001 /article/14094-eng.htm.

Hempelmann, C., Raskin, V., & Triezenberg, K. E. (2006). *Computer, Tell Me a Joke... but Please Make it Funny: Computational Humor with Ontological Semantics.* Paper presented at the FLAIRS Conference, Melbourne Beach, Florida, USA.

Highet, G. (1972). *The Anatomy of Satire*. Princeton, N.J: Princeton University Press.

Hongo, H. (2016). Literally Unbelievable: Stories from The Onion as interpreted by Facebook. Retrieved from http://literallyunbelievable.org/

Hopper, R., & Bell, R. A. (1984). Broadening the Deception Construct. *Quarterly Journal of Speech, 70*(3), 288-302.

Howell, L. (2013). *Global Risks 2013*. Retrieved from Cologny/Geneva, Switzerland: http://www3. wefo-rum.org/docs/WEF_GlobalRisks_Report_2013.pdf

LaMarre, H. L., Landreville, K. D., & Beam, M. A. (2009). The Irony of Satire: Political Ideology and the Motivation to See What You Want to See in The Colbert Report. *The International Journal of Press/Politics, 14*(2), 212-231. doi:10.1177/1940161208330904

Mihalcea, R., & Liu, H. (2006). *A corpus-based approach to finding happiness.* Paper presented at the AAAI Symposium on Computational Approaches to Analyzing Weblogs.

Mihalcea, R., & Pulman, S. (2007). Characterizing humour: An exploration of features in humorous texts. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (Vol. 4394, pp. 337-347). Berlin Heidelberg: Springer.

Mihalcea, R., Strapparava, C., & Pulman, S. (2010). *Computational models for incongruity detection in humour.* Paper presented at the Computational linguistics and intelligent text processing, Iasi, Romania.

Mocanu, D., Rossi, L., Zhang, Q., Karsai, M., & Quattrociocchi, W. (2015). Collective attention in the age of (mis)information. *Computers in Human Behavior, 51*, 1198-1204. doi:10.1016/j.chb.2015.01.024

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825-2830.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. from University of Texas at Austin http://hdl.handle.net/ 2152/31333

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and

detection of pseudo-profound bullshit. *Judgment and Decision Making, 10*(6), 549-563.

Pfaff, K. L., & Gibbs, R. W. (1997). Authorial intentions in understanding satirical texts. *Poetics, 25*(1), 45-70. doi:10.1016/s0304-422x(97)00006-5

Phiddian, R. (2013). Satire and the limits of literary theories. *Critical Quarterly, 55*(3), 44-58.

Pisarevskaya, D. (2015). *Rhetorical Structure Theory as a Feature for Deception Detection in News Reports in the Russian Language.* Paper presented at the Artificial Intelligence and Natural Language & Information Extraction, Social Media and Web Search (AINL-ISMW) FRUCT Conference, Saint-Petersburg, Russia.

Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. In J. G. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing attitude and affect in text: Theory and applications* (1 ed., pp. 1-10): Springer Nether-lands.

Reilly, R., & Nye, R. (2012). *Power, principles and the press.* Retrieved from http://www.theopen-road.com/wp-content/uploads/2012/09/Power-princi-ples-and-the-press-Open-Road-and-Populus1.pdf

Reyes, A., & Rosso, P. (2014). On the difficulty of automatically detecting irony: beyond a simple case of negation. *Knowledge and Information Systems, 40*(3), 595-614. doi:10.1007/s10115-013-0652-8

Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering, 74*, 1-12. doi:10.1016/j.datak.2012.02.005

Reyes, A., Rosso, P., & Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation, 47*(1), 239-268. doi:10.1007/s10579-012-9196-x

Riffkin, R. (2015, 2015/09/28/). Americans' Trust in Media Remains at Historical Low. *Gallup*. Retrieved from http://www.gallup.com/poll/185927/americans-trust-media-remains-historical-low.aspx

Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). *Deception detection for news: three types of fakes.* Paper presented at the Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community.

Rubin, V. L., & Conroy, N. (2012). Discerning truth from deception: Human judgments and automation efforts. *First Monday, 17*(3). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/3933/3170

Rubin, V. L., & Lukoianova, T. (2014). Truth and Deception at the Rhetorical Structure Level. *Journal of the Association for Information Science and Technology, 66*(5), 12. doi:10.1002/asi.23216

Simpson, P. (2003). *On the Discourse of Satire*: John Benjamins Publishing Company.

Sjöbergh, J., & Araki, K. (2007). *Recognizing humor without recognizing meaning.* Paper presented at the International Workshop on Fuzzy Logic and Applications, Camogli, Italy.

The Media Insight Project. (2014). *The rational and attentive news consumer*. Retrieved from https://www.americanpressinstitute.org/publications/reports/survey-research/rational-attentive-news-consumer/

Tsfati, Y. (2010). Online News Exposure and Trust in the Mainstream Media: Exploring Possible Associations. *American Behavioral Scientist, 54*(1), 22-42. doi:10.1177/0002764210376309

Tsur, O., Davidov, D., & Rappoport, A. (2010, May 23-26 2010). *ICWSM — A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews.* Paper presented at the Fourth International AAAI Conference on Web and Social Media, George Washington University, Washington, D.C. USA.

Utsumi, A. (2000). Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics, 32*(12), 1777-1806. doi:http://dx.doi.org/10.1016/S0378-2166(99)00116-2

Van Swol, L. (2014). Truth Bias. In T. Levine (Ed.), *Encyclopedia of Deception* (Vol. 1, pp. 904-906). Thousand Oaks, California: SAGE Publications.

Vrij, A., Mann, S., & Leal, S. (2012). Deception Traits in Psychological Interviewing. *Journal of Police and Criminal Psychology, 28*(2), 115-126.

Wu, Z., & Palmer, M. (1994). *Verbs semantics and lexical selection.* Paper presented at the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating Linguistics-Based Cues for Detecting Deception in Text-Based Asynchronous Computer-Mediated Communica-tions. *Group Decision and Negotiation, 13*(1), 81-106.

Ziv, A. (1988). Teaching and Learning with Humor. *The Journal of Experimental Education, 57*(1), 4-15. doi:10.1080/00220973.1988.10806492