SatiricLR: a Language Resource of Satirical News Articles

Alice Frain and Sander Wubben
Tilburg University
Tilburg, Netherlands
E-mail:A.Frain@tilburguniversity.edu, S.Wubben@tilburguniversity.edu

Abstract

In this paper we introduce the Satirical Language Resource: a dataset containing a balanced collection of satirical and non satirical news texts from various do- mains. This is the first dataset of this magnitude and scope in the domain of satire. We envision this dataset will facilitate studies on various aspects of of sat- ire in news articles. We test the viability of our data on the task of classification of satire.

Keywords: language resource, text categorisation, figurative language, irony, satire

1. Introduction

With the ever-growing amount of satire on web and the complex nature of Figurative Language Processing (FLP), studies in the field of satire are becoming more and more important. Whereas most NLP research centers on the modeling of natural language, FLP attempts to find elements in language that aid in the computational processing of figurative forms of language.

As outlined by Reyes, Rosso and Buscaldi (2012), figurative language differs from literal language as it represents meaning by using linguistic devices such as ambiguity, irony, metaphors etcetera. Uncovering the true meaning of these devices relies on our cognitive abilities, which help us reason beyond the mere syntax of a sentence. One can therefore imagine that automatically detecting satire is a challenging NLP task. In order to get to the real meaning of a text and to detect whether it is satire, a model will need to have access to contextual knowledge and has to be able to rely on various social and cognitive capacities which are difficult to computationally represent. Despite its challenging nature, a lot of work has been done in the field of FLP with promising results. Common areas of research include: the automatic detection of similes and metaphors (Veale & Hao 2007, 2010), detecting humorous texts (Mihalcea & Pulman 2007), irony detection (Carvalho, Sarmento, Silva & Oliveira 2009; Reyes & Rosso, 2011; Reyes, Rosso & Buscaldi, 2012) and sarcasm detection (Davidov, Tsur & Rappoport, 2010). Studies related to the automatic recognition of satire are scarce. Nonetheless, satire knows various definitions in academic literature. Claridge (2011) emphasizes the critical aspect of satire and reasons that satire accomplishes its satirical goal by using various figurative devices such as exaggeration, hyperbole and irony. Notice how different figurative devices are incorporated in the aforementioned definitions. This demonstrates that even though they are different devices, they are often used in combination with each other to achieve a satirical intention. This includes related concepts such as irony and humor. We therefore argue that automatic satire detection is closely related to humor, sarcasm and irony detection.

This paper builds upon the study of Burfoot & Baldwin (2009). To our knowledge, this has been the only study that focused on the detection of satire. This paper contributes to current literature by (i) using a larger dataset consisting of an equal distribution of satirical and non-satirical articles from the topical domains politics, entertainment and technology which allows investiga- tion how well models perform per topical do- main, (ii) by publishing this dataset so it can be used for future FLP related research and finally (iii) by expanding upon the findings of Burfoot and Baldwin by creating and applying a feature set that consists of textual features that have been used in related studies on humor, sarcasm and irony detection. To get some insight in the dataset, we consider the problem of satire detection, which can be cast as a binary classification task.

We approach the problem of satire detection as a binary classification task. Three types of models are tested: Bag of Words (BOW) models using unigrams or bigrams, a model based on 8 textual features derived from previous studies and a model that combines the BOW models with these 8 textual features.

2. Data

Data was collected by scraping news websites (Reuters, CNET, CBS News) and various satirical news websites (Daily Currant, DailyMash, DandyGoat, EmpireNews, The Lapine, Lightly Braised Turnip, Mouthfrog, NewsBiscuit, NewsThump, SatireWire, National Report and The Spoof). The articles on satirical news websites are often referred to as ‘fake news’. They mimic the style and format of real news agencies but incorporate humor and other aspects of figurative language. Unlike real news

agencies, the 'news' that is communicated on these websites is not true at all. They have humoristic qualities and mostly serve as entertainment. Usually they have a disclaimer somewhere on the website stating that their content is purely satirical and that it does not represent any factual information. This is often not explicitly communicated on the homepage or within their articles. As articles on satirical websites are scarcer i.e. they do not publish as many articles per day as regular news websites, we had to take articles from a wider variety of websites in order to get a substantial amount of data. The final dataset consists of articles from various topical domains ranging from 2015 to 2013. The websites were scraped using Google Chrome plugin Webscraper.IO. See Table 1 for an overview.

	Satire	*Non-satire*	*Total*
Politics	545	574	1119
Entertainment	557	578	1135
Technology	604	553	1157
Total	1706	1705	3411

Table 1: Data overview.

For each domain we aimed to gather a somewhat equally distributed amount of satire and non-satirical articles. By doing so, we work with a larger and more balanced dataset than the one used by Baldwin & Burfoot (2009). Our larger dataset is publicly available via the following link[1] and is free to use for future research projects. Data is cleaned using regular expressions to remove HTML and other markup such as website-specific headers or author signatures. This leaves us only with articles consisting of the article title and its content.

3. Method

We conduct three experiments. In experiment 1, we aim to classify using a BOW model using unigrams and bigrams. This means that no specific features are extracted, we merely look at n-gram representations of each article. Stop words are removed and for both models we use a minimum document frequency of 3 in order to remove n-grams that occur in less than 3 documents. Maximum document frequency is set to 1 in order to remove corpus-specific stop words based on intra corpus document frequencies. In experiment 2 we investigate the performance of a set of 8 textual features. Experiment 3 revolves around enhancing the unigram and bigram models from experiment 1 with our textual features. Both the unigram and the bigram model are enhanced by adding textual features from experiment 2.

Each experiment is repeated three times. First, we perform an experiment concerning all articles, so we do not make any topical distinguishment. Next, we investigate the usefulness of this model per topic. So within each experiment, we train three additional classifiers, one for each topic, using the same model. This way we can investigate how well these models perform when we classify on a topical level. Classifiers are trained on training data (70% of the data). In this phase, we perform grid searches in order to find the optimal parameters for our classifiers. This was done using 10-fold cross validation (CV) in order to account for over-fitting and to make optimal use of the amount training data we have. Finally, the model is evaluated on the test set (30% of the data), using the optimal parameters found during the training phase, in order to measure the generalizability of our model. This has been done using three classifiers: Naive Bayes (NB), Decision Trees (DT) and Support Vector Classifiers (SVC).

3.1. Feature extraction

Profanity: This is a measure of profanity words within each article. We count the number of profanity words within an article using a pre-defined library with profanity words and divide it by the article word length leaving us with a relative measure of profanity.

Punctuation: As research by Davidov, Tsur and Rappoport (2010) suggests, punctuation might be a relevant feature in detecting figurative language. It will include measures of '?', '!', and '...' within an article. This is calculated by counting the occurrences of '!', '?' and ' ...' and divide it by the word length of the article which leads to 3 features: relative exclamation marks, relative question marks and relative dots.

Negativity & Positivity: As indicated by Mihalcea and Pulman (2007), humorous articles are often characterized by negative words. This feature is extracted by measuring the amount of both negative and positive oriented words within an article divided by the total amount of words of the article. This leaves us with 2 features: one representing the relative positivity, and one representing the relative negativity of an article. We used a subjectivity lexicon consisting of positive and negative words compiled by Wilson, Wiebe and Hoffmann (2005).

Human-centeredness: As indicated by Mihalcea and Pulman (2007), humorous articles are often characterized by the use of personal pronouns and human-centered words. They operationalize human-centered words as words that belong to specific WordNet (Fellbaum, 1998) synsets. Words that belong to the synsets 'person', 'human', 'individual', 'someone', 'somebody', 'mortal', 'soul', 'relative', 'relation', 'relationship', and 'human relationship' are considered to be human-centered words. For each article, we count the amount of words that belong to these synsets and divide them by the total amount of words within the article. Regarding the measure of personal pronouns, we count the personal pronouns, which are then divided by the total amount of

[1] https://github.com/swubb/SatiricLR

words of the article which also leaves us with a relative measure of personal pronouns.

4. Results

This section presents the F-scores of the best performing classifier for each experiment on the test sets (all articles *n*=1024, politics *n*=336, entertainment *n*=341, technology *n*=348). For each domain, we also computed the baseline score using a baseline classifier. This baseline always predicts the most frequent label (i.e. satire or non-satire) found in the training set. It is a useful metric against which we can compare the performance of our classifiers. The baseline scores for each topic are as follows: all articles: 0.48, politics: 0.43, entertainment: 0.42 and technology: 0.48.

4.1. Experiment 1

	Unigrams	***Bigrams***
All articles	0.93 (SVC)	0.88 (SVC)
Politics	**0.94 (SVC)**	**0.93 (NB)**
Entertainment	0.90 (SVC)	0.81 (NB)
Technology	0.93 (NB)	0.88 (NB)

Table 2: F-scores experiment 1 BOW model using unigrams and bigrams.

4.2. Experiment 2

	Textual Features
All articles	0.75 (SVC)
Politics	**0.84 (SVC)**
Entertainment	0.70 (SVC)
Technology	0.79 (SVC)

Table 3: F-scores experiment 2 using 8 textual features.

4.3. Experiment 3

	Unigrams +	***Bigrams +***
All articles	0.93 (SVC)	0.89 (SVC)
Politics	**0.95 (SVC)**	**0.93 (NB)**
Entertainment	0.89 (SVC)	0.80 (SVC)
Technology	0.93 (NB)	0.90 (NB)

Table 4: F-scores experiment 3 using BOW unigrams and bigram models combined with 8 textual features.

5. Discussion

Experiment 1 has shown that classifying articles by means of a BOW approach is a viable method. We obtained high F-scores that easily outperform their baseline counterparts. This corresponds with findings of Mihalcea and Pulman (2007) who have also demonstrated that humorous and non-humorous texts are separable using a BOW approach. The unigram model appears to perform best. Both the unigram and bigram models score lowest on the entertainment domain and perform best on politics related articles. A caveat here is that despite preprocessing and anonymizing all articles, source specific words can still seep through to our model. Specific websites may use specific terminology, so the task then becomes classifying sources instead of text type, which is not what we want. This illustrates one of the pitfalls of the BOW method and its drawback on newswire data. Stripping HTML and website-specific data may sometimes not be enough. BOW models are very sensitive to content, meaning that one has to be extremely wary of content-specific information in article text that may have an influence on model performance. Additionally, we must ask ourselves how far we can go in cleaning data without being too strict and running the risk of influencing the original data by over-editing the running text of articles.

Despite this limitation of our n-gram experiments, our findings regarding the use of a model based on 8 textual features are promising and more robust as they only measure features associated with satire and humor in general. When training classifiers using merely a set of 8 textual features we obtain very decent F-scores on articles from all topics (0.75), politics (0.84), entertainment (0.70) and technology (0.79) that easily outperformed their baselines. Again, classifying entertainment articles seems to be slightly more difficult with this textual feature set. This might indicate that the content of articles from the entertainment domain is different than articles from the technology and politics domain. The 8 features we used might not be as effective in capturing their distinct qualities. On the other hand, it could also be that the content of satirical and non-satirical entertainment articles does not differ as much as is the case for the other topics, making it harder for the classifiers to differ between both classes. When looking at the most informative features using the Gini Importance measure of our Decision Tree classifiers, we see that personal pronouns, negativity and positivity dominate the top 3 meaning that they are effective features. Features related to punctuation often end up in the bottom top 3 suggesting that they are less informative.

Enhancing the n-gram models with the 8 textual features did not lead to any significant improvement over the original models based on n-grams only. This can be explained because they are both related to the words within the articles meaning that they implicitly hold same kind of information. This is in line with Burfoot & Baldwin (2009) who found no significant improvements after enhancing their unigram models with lexical features

6. Conclusion

We have described SatiricLR, a language re- source containing a balanced collection of satirical and non-satirical news texts. It contains news articles in the domains of politics, entertainment and technology. In total in contains 3411 texts, of which 1706 are satirical.

We obtain very reasonable results on the classification task of classifying a text as either satircal or non-satirical. We believe this is a very valuable resource for future work on computational modelling of satirical language.

Our model consisting of 8 textual features has yielded promising results and we might argue that it has several advantages over BOW models. Especially after having acknowledged the possible limitations of n-gram models. First, the feature vectors are sparser as they consist of only 8 features per instance meaning that it takes less time to train a classifier than when using a BOW approach. Additionally, our textual feature model is less reliant on the exact article content, making it more robust and less prone to source-specific content within the running text of articles. This is because it only measures predefined textual features found in literature on irony and humor detection. Consequently, one does not have to worry as much to have perfectly clean data, which is beneficial when working with web scraped data. Moreover, as these features rely on domain specific knowledge, they are more informative and give a better insight into what types of textual elements characterise satirical and non-satirical texts.

The features we used are simple but have nonetheless shown to be effective. In order to enhance their performance, future research could be done to explore the use of more sophisticated features. One idea might be to focus on capturing the semantic dissimilarity between words, as satirical news articles are often characterised by a certain implausibility.

7. Bibliographical References

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.

Carvalho, P., Sarmento, L., Silva, M. J., & De Oliveira, E. (2009). *Clues for detecting irony in user-generated contents: oh...!! it's so easy;-).* Paper presented at the Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion.

Claridge, C. (2011). *Hyperbole in English: A corpus based study of exaggeration*. New York: Cambridge University Press.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. Paper presented at the Proceeding CoNLL '10 Proceedings of the Fourteenth Conference on Computational Natural Language Learning.

Fellbaum, C. (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Mihalcea, R., & Pulman, S. (2007). Characterizing humour: An exploration of features in humorous texts. Paper presented at the 8th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2007.

Reyes, A., & Rosso, P. (2011). Mining Subjective Knowledge from Customer Reviews: A Specific Case of Irony Detection. Paper presented at the Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011.

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.

Veale, T. & Hao, Y. (2007). Comprehending and Generating Apt Metaphors: A Web-driven, Case-based Approach to Figurative Language. Paper presented at the Proceeding AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence.

Veale, T., & Hao Y. (2010). Detecting Ironic Intent in Creative Comparisons. Paper presented at the Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence.

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). *Recognizing contextual polarity in phrase-level sentiment analysis*. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 347-354). Association for Computational Linguistics.