Open Secrets and Wrong Rights: Automatic Satire Detection in English Text

Aishwarya N. Reganti
IIIT-Sri City, Chittoor
Sri City, AP, India, 517588
aishwarya.r14@iiits.in

Tushar Maheshwari
IIIT-Sri City, Chittoor
Sri City, AP, India, 517588
tushar.m14@iiits.in

Upendra Kumar
IIIT-Sri City, Chittoor
Sri City, AP, India, 517588
amitava.das@iiits.in

Amitava Das
IIIT-Sri City, Chittoor
Sri City, AP, India, 517588
upendra.k14@iiits.in

Erik Cambria
Nanyang Technological
University, Singapore
50 Nanyang Ave, Singapore
cambria@ntu.edu.sg

CSCW '17 Companion, February 25 - March 01, 2017, Portland, OR, USA
ACM 978-1-4503-4688-7/17/02.
http://dx.doi.org/10.1145/3022198.3026344

Abstract
Satire is an element of figurative language which often conveys feelings contrary to what is literally stated. It refers to a trenchant wit, irony, or sarcasm used to expose discredit vice or folly. The presence of a satirical utterance in text can entirely change the sentiment of the statement, hence it is necessary to distinguish be-tween true positive statements and satirical ones.In this paper, we identify key value components and features for automatic satire detection. Our experiments have been carried out on three data sets, namely, tweets, product reviews and newswire articles. We examine the impact of a number of state of the art features as well as new generalised textual features.

Corpus Statistics
Amazon Product Reviews (i) : We have used the cor-pus created by Filatova [3] in 2012. This data set con-sists of 1,254 Amazon product reviews reviews, of which 437 are ironic and 817 are non-ironic. Since we started with the notion that satire is a super class of language devices including irony and sarcasm, we used this cor-pus to test our models.

No.	Group	Features
1	Baseline Features(BF)	character n-grams, word n-grams, word skipgrams
2	Lexical Features(LF)	NRC Emotion lexicon, SentiWordNet
3	Literary device features(LD)	Hyperbole, Alliteration, Inversions, Imagery, Onomatopoeia
4	Sentiment Amplifiers(SA)	Brackets, Ellipses, Quotes, Question marks, Exclamation marks, Interjections, Emoticons, Slang words, Acronyms
5	Speech Act Features(SAF)	As Illustrated in Table 3
6	Sensicon Features(SE)	Sense scores for Sight, Hearing, Taste, Smell and Touch
7	Sentiment Continuity disruption features(SCD)	Count of Flips

Table 1: Feature groups used for Satire Detection

Feature group	Feature length
BF	len(Ngrams)
LF	11
LD	5
SA	9
SAF	11
SE	5
SCD	1

Table 2: Feature length

Corpus	Total	satirical	Non-satirical
(i)	1254	437	817
(ii)	4000	233	3767
(iii)	8,000	3,000	5,000

Table 3: Corpus Statistics

Newswire Articles (ii) : This corpus was released by [1] in 2009, This corpus contains a total of 4000 newswire documents and 233 satire news articles. The newswire documents were randomly sampled from the English Gigaword Corpus.

Twitter Dataset (iii) : As of the first quarter of 2016, the micro-blogging service Twitter averaged at about 236 million monthly active users, with around 6,000 tweets being posted every second. Therefore, twitter is definitely a rich source of data. The data was retrieved using the search query option of twitter4j rest API. We used #satire, #irony #sarcasm as the query terms. Inter-annotator agreement between 3 proficient English speakers was used to re-check if the tweet was actually satirical. Statistics of all the three Corpus are mentioned in Table 3.

Architecture and features used

We model the task of satire detection as a supervised classification problem in which each instance is categorized as being satirical or non-satirical. We examine different classifiers and features that affect the accuracy of our system. We use seven sets of features to build our model. In the next subsections, we describe the features used and the set of classifiers compared. Table1 provides an overview of the group of features in our model and table 2 elucidates the length of the features vector in each group.

Ngrams: N-grams are good task-independent features for any kind of textual classification. Hence we chose n-grams as our baseline features, since task-independent features are necessary to detect satire as the difference between positive and negative classes is subtle and us-ing only task-specific features does not yield very good accuracy. We retrieved word n-grams/ Bag of words(bi-grams and tri-grams) and skipgrams(bi-grams) from our corpus.

Sentiment Lexica: Two sentiment lexicons were made use of. The NRC lexicon [5] has affect annotations for each word. Each word is tagged with either one of the 2 sentiments: negative positive, or one of the 8 emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust. From these words, only words with annotations anger, anticipation, disgust, fear, joy, sadness & sur-prise were chosen, as satirical sentences generally con-tain words with extreme emotions like anger or joy. The frequency of words in each category were added as fea-tures. SentiWordNet [2] is one of the largest sentiment lexicons with about words. SentiWordNet assigns to each synset of WordNet three sentiment scores: posi-tivity, negativity, objectivity. The nett score was obtained by adding the positive and negative scores, and added as a feature

No	Features	Accuracy
1	Only bag-of-words(BW)	55.75%
2	BW + WH-words(wh)	57.02%
3	BW + Question mark(qm)	62.97%
4	BW + SentiWordNet(senti)	67.75%
5	BW + NRC(nrc)	63.22%
6	BW + wh + qm	64.18%
7	BW + wh + qm + senti	69.41%
8	BW + wh + qm + senti + nrc	70.33%

Table 4: Speech-Act Classifier Results

Sentiment Amplifiers: As a general trend, it can be observed that almost all satirical utterances use one or the other form of sentiment amplifiers. Sentiment amplifiers are those elements which highlight an emotion or inten-sify it. Amplifiers such as exclamation marks, quotes, ellipses etc. are used to emphasize the sentiment con-veyed in the statement. The presence/absence of these amplifiers was used as features.

Sentiment Continuity disruption features: Generally, it can be observed that in large text, consisting of more than one sentence, and satirical statements, the polar-ity flips at least once, either when the satirical statement ends or when the satirical sentence starts. The more the number of flips in the text, more the number of satir-ical sentences, and hence a stronger satire. We used the number of flips in the text as the Sentiment Con-tinuity disruption feature. The feature however will not perform very well on short utterances like tweets.

Speech Act: A speech act in linguistics is an utterance that has performative function in language and commu-nication [4]. In short, it is the action that lies in utter-ances such as apology, appreciation, promise, thank-ing, etc. Here, in this paper, we use 11 major speech acts (Action Directive, Apology, Appreciation, Response Acknowledgment, Statement Non-Opinion, Statement Opinion, Thanking, Wh Question Yes Answers, Yes-No Question & Other) to classify text . We made a speech act classifier with an accuracy of 70.33% (details in Fig-ure 4), and used the predictions of the classifier as fea-tures. The classifier was trained using the corpus, de-scribed in [4].

Sensicon: Sensicon is a sensorial lexicon that asso-ciates English words with senses [7]. It contains words with sense association scores for the five basic senses: Sight, Hearing, Taste, Smell, and Touch. The cumu-lative sensicon scores for each instance of the dataset were used as features, therefore a total of 5 features, re-ferring to each of the 5 senses were added as features.

No	Features	LR	RF	SVM	DT	Ensemble
1	BF	70.33 %	66.57 %	65.25 %	66.71 %	73.91%
2	BF+ LF	71.83 %	67.09 %	65. 66 %	66.79 %	73.82%
3	BF+ LD	69.89 %	66.82 %	66.02 %	66.68 %	73.22%
4	BF + SA	71.33 %	66.93 %	65.23 %	66.92 %	73.15%
5	BF + SAF	73.42 %	68.02 %	65.99%	65.22 %	75.66%
6	BF + SE	71.22 %	65.60 %	65.33 %	66.03 %	73.33 %
7	BF + SCD	72.01 %	66.61 %	65.56 %	67.05 %	74.88 %
8	All Features	75.30 %	68.93 %	66.63 %	67.11 %	77.96 %

Table 5: F-Scores for product review corpus

No	Features	LR	RF	SVM	DT	Ensemble
1	BF	73.23 %	69.16 %	72.89 %	68.85 %	74.99 %
2	BF+ LF	73.32 %	69.27 %	72.76 %	68.82 %	74.82 %
3	BF+ LD	73.24 %	70.48 %	72.99 %	68.63 %	74.11 %
4	BF +SA	75.26 %	71.22%	73.91 %	70.13 %	76.91 %
5	BF + SAF	74.32 %	70.48 %	72.10%	68.89 %	74.86 %
6	BF +SE	73.02 %	70.02 %	74.03 %	69.66 %	74.58 %
7	BF +SCD	73.24 %	70.71 %	72.56 %	68.01 %	74.43 %
8	All features	76.89 %	71.06 %	74.03 %	68.11 %	78.16 %

Table 6: F-Scores for Twitter posts Corpus

No	Features	LR	RF	SVM	DT	Ensemble
1	BF	70.12 %	62.12 %	68.11 %	63.16 %	69.04 %
2	BF+ LF	72.77 %	61.23 %	68.45 %	63.55 %	75.32 %
3	BF+ LD	71.33 %	63.33 %	68.18 %	65.11 %	75.11 %
4	BF +SA	70.16 %	62.19 %	69.67 %	63.44 %	74.66 %
5	BF + SAF	73.24 %	61.90 %	68.99%	63.77 %	76.98 %
6	BF +SE	69.08 %	62.88 %	68.78 %	62.68 %	74.77 %
7	BF +SCD	71.88 %	63.22 %	69.12 %	63.33 %	75.77 %
8	All features	75.88 %	63.89 %	69.34 %	63.22 %	79.02 %

Table 7: F-Scores for Newswire Corpus

Literary device features: Since satire is all about ex-pressing ones disgust/anger in a creative indirect way, we detect the presence of certain literary devices like metaphor, alliteration, onomatopoeia and Hyperbole that are generally used in satirical statements, using them as features.

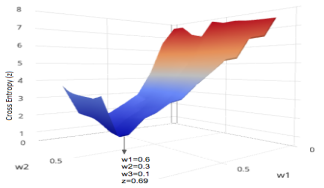

Figure 1: Plot displaying minima of cross entropy values over weight space

	LR	RF	SVM	DT
LR	1.00	0.74	0.57	0.44
RF	–	1.00	0.63	0.38
SVM	–	–	1.00	0.71
DT	–	–	–	1.00

Table 8: Pearson correlation between classifier predictions

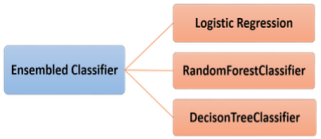

Figure 2: Ensemble Classifier

.

Evaluation

The tables 5, 6 and 7 show the F-scores obtained on three corpora, using 5-fold cross validation. Five different classifiers have been used, Logistic Regression(LR), Random Forest(RF), Support Vector Machine(SVM), De-cision Tree(DT) and an ensemble of classifiers for better performance. An ensemble(Ens.) of different machine learning classifiers (as shown in Figure 2 can be used to achieve broad solution spaces by multiplying com-binations of best component search spaces. In order to select or design best component search spaces, the individual components should be independent in order to assimilate less correlated information from the data. Therefore, Pearson correlation (Table 8) was found be-tween different classifie predictions. The choice of the three classifiers was based on good performance of lin-ear regression and low correlation between Random Forest and Decision tree. We explored the best com-binations of weights for each of the three components in the search space by iteratively running over all com-binations of w1 (weight of LR), w2 (Weight of RF), w3 Weight of DT) and choosing a value where minimum cross entropy was obtained. The cross entropy results are reported in Figure 1. We observe that the minima exists at w1 = 0.6, w2 = 0.3, w3 = 0.1.

$$S : \{(w1, w2, w3) | w1 + w2 + w3 = 1.0 \\ where\, w1, w2, w3 \in \{0.1, 0.2, ..., 0.9\}\}. \quad (1)$$

Conclusion

From the obtained results as displayed in the tables, we observe that the features proposed work reason-ably well for all corpora since our system outperforms the state-of-the-art for product review corpus (by 6%without star ratings) and equals the state-of-the-art for newswire corpus (leads by 2%). Since the twitter post corpus was created by us, we cannot draw any compar-ative analysis on it. We observe that some features are more useful than the others depending on the source of text. A crucial observation worth mentioning is that the Ensemble classifier, performed best on all three Cor-pora. Therefore, our choice of classifiers for the ensem-ble, based on the cross entropy calculations proved to be worthwhile. Further work can include deep learning approaches as in [6] for better results.

References

[1] Clint Burfoot and Timothy Baldwin. 2009. Automatic Satire Detection: Are You Having a Laugh?. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (ACLShort '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 161–164.

[2] Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, Vol. 6. Citeseer, 417–422.

[3] Elena Filatova. 2012. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing. In *Proceedings of the Eighth International Conference on Lan-guage Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*. 392–398.

[4] Geoffrey Leech and Martin Weisser. 2003. Generic speech act annotation for task-oriented dialogues. In *Procs. of the 2003 Corpus Linguistics Conference, pp. 441Y446. Centre for Computer Corpus Research on Language Technical Papers, Lancaster University*.

[5] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.

[6] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks. In *COLING*.

[7] Serra Sinem Tekiroğlu, Gözde Özbal, and Carlo Strapparava. 2014. Sensicon: An automatically constructed sensorial lexicon. (2014).