Adversarial Training for Satire Detection: Controlling for Confounding Variables

Robert McHardy[1], **Heike Adel**[1,2*] and **Roman Klinger**[1]
[1] Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
[2] Bosch Center for Artificial Intelligence, Renningen, Germany
{firstname.lastname}@ims.uni-stuttgart.de

Abstract

The automatic detection of satire vs. regular news is relevant for downstream applications (for instance, knowledge base population) and to improve the understanding of linguistic characteristics of satire. Recent approaches build upon corpora which have been labeled automatically based on article sources. We hypothesize that this encourages the models to learn characteristics for different publication sources (*e.g.*, "The Onion" vs. "The Guardian") rather than characteristics of satire, leading to poor generalization performance to unseen publication sources. We therefore propose a novel model for satire detection with an adversarial component to control for the confounding variable of publication source. On a large novel data set collected from German news (which we make available to the research community), we observe comparable satire classification performance and, as desired, a considerable drop in publication classification performance with adversarial training. Our analysis shows that the adversarial component is crucial for the model to learn to pay attention to linguistic properties of satire.

1 Introduction

Satire is a form of art used to criticize in an entertaining manner (*cf.* Sulzer, 1771, p. 995ff.). It makes use of different stylistic devices, *e.g.*, humor, irony, sarcasm, exaggerations, parody or caricature (Knoche, 1982; Colletta, 2009). The occurrence of harsh, offensive or banal and funny words is typical (Golbert, 1962; Brummack, 1971).

Satirical news are written with the aim of mimicking regular news in diction. In contrast to misinformation and disinformation (Thorne and Vlachos, 2018), it does not have the intention of fooling the readers into actually believing something wrong in order to manipulate their opinion.

The task of satire detection is to automatically distinguish satirical news from regular news. This is relevant, for instance, for downstream applications, such that satirical articles can be ignored in knowledge base population. Solving this problem computationally is challenging. Even human readers are sometimes not able to precisely recognize satire (Allcott and Gentzkow, 2017). Thus, an automatic system for satire detection is both relevant for downstream applications and could help humans to better understand the characteristics of satire.

Previous work mostly builds on top of corpora of news articles which have been labeled automatically based on the publication source (*e.g.*, "The New York Times" articles would be labeled as regular while "The Onion" articles as satire[1]). We hypothesize that such distant labeling approach leads to the model mostly representing characteristics of the publishers instead of actual satire. This has two main issues: First, interpretation of the model to obtain a better understanding of concepts of satire would be misleading, and second, generalization of the model to unseen publication sources would be harmed. We propose a new model with adversarial training to control for the confounding variable of publication sources, *i.e.*, we debias the model.

Our experiments and analysis show that (1) the satire detection performance stays comparable when the adversarial component is included, and (2) that adversarial training is crucial for the model to pay attention to satire instead of publication characteristics. (3), we publish a large German data set for satire detection which is a) the first data set in German, b) the first data set including publication sources, enabling the experiments at hand, and c) the largest resource for satire detection so far.[2]

* Work was done at University of Stuttgart.

[1] https://www.theonion.com/, https://www.nytimes.com/
[2] Data/code: www.ims.uni-stuttgart.de/data/germansatire.

Proceedings of NAACL-HLT 2019, pages 660–665
Minneapolis, Minnesota, June 2 - June 7, 2019.

2 Previous Work

Previous work tackled the task of automatic English satire detection with handcrafted features, for instance, the validity of the context of entity mentions (Burfoot and Baldwin, 2009), or the coherence of a story (Goldwasser and Zhang, 2016). Rubin et al. (2016) use distributions of parts-of-speech, sentiment, and exaggerations. In contrast to these approaches, our model uses only word embeddings as input representations. Our work is therefore similar to Yang et al. (2017) and De Sarkar et al. (2018) who also use artificial neural networks to predict if a given text is satirical or regular news. They develop a hierarchical model of convolutional and recurrent layers with attention over paragraphs or sentences. We follow this line of work but our model is not hierarchical and introduces less parameters. We apply attention to words instead of sentences or paragraphs, accounting for the fact that satire might be expressed on a sub-sentence level.

Adversarial training is popular to improve the robustness of models. Originally introduced by Goodfellow et al. (2014) as generative adversarial networks with a generative and a discriminative component, Ganin et al. (2016) show that a related concept can also be used for domain adaptation: A domain-adversarial neural network consists of a classifier for the actual class labels and a domain discriminator. The two components share the same feature extractor and are trained in a minimax optimization algorithm with gradient reversal: The sign of the gradient of the domain discriminator is flipped when backpropagating to the feature extractor. Building upon the idea of eliminating domain-specific input representations, Wadsworth et al. (2018) debias input representations for recidivism prediction, or income prediction (Edwards and Storkey, 2016; Beutel et al., 2017; Madras et al., 2018; Zhang et al., 2018).

Debiasing mainly focuses on word embeddings, *e.g.*, to remove gender bias from embeddings (Bolukbasi et al., 2016). Despite previous positive results with adversarial training, a recent study by Elazar and Goldberg (2018) calls for being cautious and not blindly trusting adversarial training for debiasing. We therefore analyze whether it is possible at all to use adversarial training in another setting, namely to control for the confounding variable of publication sources in satire detection (see Section 3.1).

3 Methods for Satire Classification

3.1 Limitations of Previous Methods

The data set used by Yang et al. (2017) and De Sarkar et al. (2018) consists of text from 14 satirical and 6 regular news websites. Although the satire sources in train, validation, and test sets did not overlap, the sources of regular news were not split up according to the different data sets (Yang et al., 2017). We hypothesize that this enables the classifier to learn which articles belong to which publication of regular news and classify everything else as satire, given that one of the most frequent words is the name of the website itself (see Section 4.1). Unfortunately, we cannot analyze this potential limitation since their data set does not contain any information on the publication source[3]. Therefore, we create a new corpus in German (see Section 4.1) including this information and investigate our hypothesis on it.

3.2 Model

Motivated by our hypothesis in Section 3.1, we propose to consider two different classification problems (satire detection and publication identification) with a shared feature extractor. Figure 1 provides an overview of our model. We propose to train the publication identifier as an adversary.

3.2.1 Feature Extractor

Following De Sarkar et al. (2018), we only use word embeddings and no further handcrafted features to represent the input. We pretrain word embeddings of 300 dimensions on the whole corpus using word2vec (Mikolov et al., 2013). The feature generator f takes the embeddings of the words of each article as input for a bidirectional LSTM (Hochreiter and Schmidhuber, 1997), followed by a self-attention layer as proposed by Lin et al. (2017). We refer to the union of all the parameters of the feature extractor as θ_f in the following.

3.2.2 Satire Detector

The gray part of Figure 1 shows the model part for our main task – satire detection. The satire detector feeds the representation from the feature extractor into a softmax layer and performs a binary classification task (satire: yes or no). Note that, in contrast to De Sarkar et al. (2018), we classify satire solely

[3] https://data.mendeley.com/datasets/hx3rzw5dwt/draft?a=377d5571-af17-4e61-bf77-1b77b88316de, v.1, 2017, accessed on 2018-11-23

on the document level, as this is sufficient to analyze the impact of the adversarial component and the influence of the publication source.

3.2.3 Publication Identifier

The second classification branch of our model aims at identifying the publication source of the input. Similar to the satire detector, the publication identifier consists of a single softmax layer which gets the extracted features as an input. It then performs a multi-class classification task since our dataset consists of 15 publication sources (see Table 1).

3.2.4 Adversarial Training

Let θ_f be the parameters of the feature extractors and θ_s and θ_p be the parameters of the satire detector and the publication identifier, respectively. The objective function for satire detection is

$$J_s = -\mathbb{E}_{(x,y_s)\sim p_{\text{data}}} \log P_{\theta_f \cup \theta_s}(y_s, x)\,, \quad (1)$$

while the objective for publication identification is

$$J_p = -\mathbb{E}_{(x,y_p)\sim p_{\text{data}}} \log P_{\theta_f \cup \theta_p}(y_p, x)\,. \quad (2)$$

Note that the parameters of the feature extractor θ_f are part of both model parts. Since our goal is to control for the confounding variable of publication sources, we train the publication identifier as an adversary: The parameters of the classification part θ_p are updated to optimize the publication identification while the parameters of the shared feature generator θ_f are updated to fool the publication identifier. This leads to the following update equations for the parameters

$$\theta_s := \theta_s - \eta \frac{\partial J_s}{\partial \theta_s} \quad (3)$$

$$\theta_p := \theta_p - \eta \frac{\partial J_p}{\partial \theta_p} \quad (4)$$

$$\theta_f := \theta_f - \eta \left(\frac{\partial J_s}{\partial \theta_f} - \lambda \frac{\partial J_p}{\partial \theta_f}\right) \quad (5)$$

with η being the learning rate and λ being a weight for the reversed gradient that is tuned on the development set. Figure 1 depicts the gradient flow.

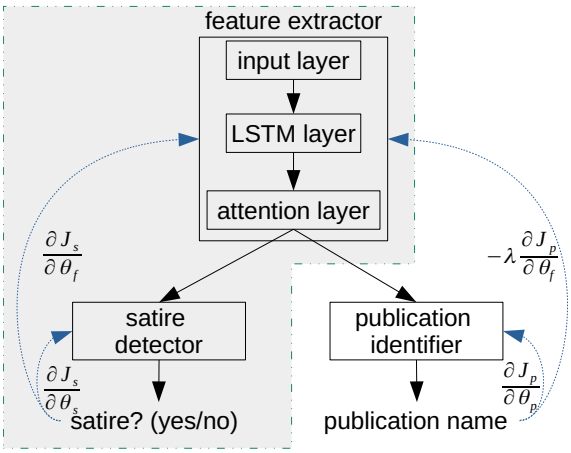

Figure 1: Architecture of the model. The gray area on the left shows the satire detector; the white area on the right is the adversary (publication identifier); the gradient flow with and without adversarial training is shown with blue arrows pointing upwards.

4 Experiments

4.1 Experimental Setting

Dataset. We consider German regular news collected from 4 websites and German satirical news from 11 websites. Table 1 shows statistics and sources of the corpus, consisting of almost 330k articles. The corpus contains articles published between January 1st, 2000 and May 1st, 2018. Each publication has individual typical phrases and different most common words. Among the most common words is typically the name of each publication, *e.g.*, "Der Spiegel" has "SPIEGEL" as fifth and "Der Postillon" "Postillon" as third most common word. We did not delete those words to keep the dataset as realistic as possible. We randomly split the data set into training, development (dev) and test (80/10/10 %) with the same label distributions in all sets. Given the comparable large size of the corpus, we opt for using a well-defined test set for reproducibility of our experiments in contrast to a crossvalidation setting.

Research questions. We discuss two questions. RQ1: How does a decrease in publication classification performance through adversarial training affect the satire classification performance? RQ2: Is adversarial training effective for avoiding that the model pays most attention to the characteristics of publication source rather than actual satire?

Baseline. As a baseline model, we train the satire detector part (gray area in Figure 1) on the satire task. Then, we freeze the weights of the feature extractor and train the publication classifier on top of it. In addition, we use a majority baseline model which predicts the most common class.

Hyperparameters. We cut the input sentences to a maximum length of 500 words. This enables us to fully represent almost all satire articles and

	Publication	#Articles	Average Length Article	Sent.	Title
Regular	Der Spiegel	31,180	556.74	19.04	7.47
	Der Standard	53,632	328.82	18.62	6.3
	Südd. Zeit.	177,605	635.77	17.58	7.74
	Die Zeit	57,802	1,116.53	17.0	5.2
Satire	Der Enthüller	324	404.3	13.87	9.67
	Eulenspiegel	192	1,072.17	17.45	4.38
	Nordd. Nach.	211	188.46	17.84	8.46
	Der Postillon	5,065	225.36	19.59	9.16
	Satirepatzer	193	262.99	12.26	7.53
	Die Tagespresse	1,271	301.28	16.39	10.83
	Titanic	149	292.88	16.04	7.79
	Welt (Satire)	1,249	291.45	21.76	9.02
	Der Zeitspiegel	171	315.76	18.69	9.71
	Eine Zeitung	416	265.16	16.04	13.35
	Zynismus24	402	181.59	17.67	11.96
	Regular	320,219	663.45	17.79	6.86
	Satire	9,643	269.28	18.73	9.52

Table 1: Corpus statistics (average length in words)

capture most of the content of the regular articles while keeping the training time low. As mentioned before, we represent the input words with 300 dimensional embeddings. The feature extractor consists of a biLSTM layer with 300 hidden units in each direction and a self-attention layer with an internal hidden representation of 600. For training, we use Adam (Kingma and Ba, 2014) with an initial learning rate of 0.0001 and a decay rate of 10^{-6}. We use mini-batch gradient descent training with a batch size of 32 and alternating batches of the two branches of our model. We avoid overfitting by early stopping based on the satire F1 score on the development set.

Evaluation. For evaluating satire detection, we use precision, recall and F1 score of the satire class. For publication identification, we calculate a weighted macro precision, recall and F1 score, *i.e.*, a weighted sum of class-specific scores with weights determined by the class distribution.

4.2 Selection of Hyperparameter λ

Table 2 (upper part) shows results for different values of λ, the hyperparameter of adversarial training, on dev. For $\lambda \in \{0.2, 0.3, 0.5\}$, the results are comparably, with $\lambda = 0.2$ performing best for satire detection. Setting $\lambda = 0.7$ leads to a performance drop for satire but also to $F_1 = 0$ for publication classification. Hence, we chose $\lambda = 0.2$ (the best performing model on satire classification) and $\lambda = 0.7$ (the worst performing model on publication identification) to investigate RQ1.

	Model	Satire P	R	F1	Publication P	R	F1
dev	majority class	0.0	0.0	0.0	29.5	54.3	38.3
	no adv	98.9	**52.6**	**68.7**	44.6	56.2	49.7
	adv, $\lambda = 0.2$	**99.3**	50.8	67.2	31.2	55.4	40.0
	adv, $\lambda = 0.3$	97.3	48.9	65.0	31.1	54.8	39.6
	adv, $\lambda = 0.5$	99.1	50.8	67.2	31.7	55.2	40.3
	adv, $\lambda = 0.7$	86.7	44.1	58.4	**26.9**	**0.0**	**0.0**
test	majority class	0.0	0.0	0.0	29.1	53.9	37.8
	no adv.	99.0	**50.1**	**66.5**	44.2	55.7	49.3
	adv, $\lambda = 0.2$	**99.4**	49.4	66.0	**30.8**	54.8	39.5
	adv, $\lambda = 0.7$	85.0	42.5	56.6	31.3	**0.0**	**0.0**

Table 2: Results on dev and independent test data.

5 Results (RQ1)

The bottom part of Table 2 shows the results on test data. The majority baseline fails since the corpus contains more regular than satirical news articles. In comparison to the baseline model without adversarial training (no adv), the model with $\lambda = 0.2$ achieves a comparable satire classification performance. As expected, the publication identification performance drops, especially the precision declines from 44.2 % to 30.8 %. Thus, a model which is punished for identifying publication sources can still learn to identify satire.

Similar to the results on dev, the recall of the model with $\lambda = 0.7$ drops to (nearly) 0 %. In this case, the satire classification performance also drops. This suggests that there are overlapping features (cues) for both satire and publication classification. This indicates that the two tasks cannot be entirely untangled.

6 Analysis (RQ2)

To address RQ2, we analyze the results and attention weights of the baseline model and our model with adversarial training.

6.1 Shift in Publication Identification

The baseline model (no adv) mostly predicts the correct publication for a given article (in 55.7 % of the cases). The model with $\lambda = 0.2$ mainly (in 98.2 % of the cases) predicts the most common publication in our corpus ("Süddeutsche Zeitung"). The model with $\lambda = 0.7$ shifts the majority of predictions (98.7 %) to a rare class (namely "Eine Zeitung"), leading to its bad performance.

Example 1

German original:

no adv	Erfurt (dpo) - Es ist eine Organisation , die ausserhalb von Recht und Ordnung agiert , zahlreiche NPD-Funktionäre finanziert und in nicht unerheblichem Maße in die Mordserie der sogenannten Zwickauer Zelle verstrickt ist .
adv	Erfurt (dpo) - Es ist eine Organisation , die ausserhalb von Recht und Ordnung agiert , zahlreiche NPD-Funktionäre finanziert und in nicht unerheblichem Maße in die Mordserie der sogenannten Zwickauer Zelle verstrickt ist .

English translation:

no adv	Erfurt (dpo) - It is an organization which operates outside of law and order , funds numerous NPD operatives and is to a not inconsiderable extent involved in the series of murders of the so called Zwickauer Zelle .
adv	Erfurt (dpo) - It is an organization which operates outside of law and order , funds numerous NPD operatives and is to a not inconsiderable extent involved in the series of murders of the so called Zwickauer Zelle .

Example 2

German original:

no adv	Immerhin wird derzeit der Vorschlag diskutiert , den Familiennachzug nur inklusive Schwiegermüttern zu erlauben , wovon sich die Union einen abschreckenden Effekt erhofft .
adv	Immerhin wird derzeit der Vorschlag diskutiert , den Familiennachzug nur inklusive Schwiegermüttern zu erlauben , wovon sich die Union einen abschreckenden Effekt erhofft .

English translation:

no adv	After all , the proposal to allow family reunion only inclusive mothers-in-law is being discussed , whereof the Union hopes for an off-putting effect .
adv	After all , the proposal to allow family reunion only inclusive mothers-in-law is being discussed , whereof the Union hopes for an off-putting effect .

Figure 2: Attention weight examples for satirical articles, with and without adversary.

6.2 Interpretation of Attention Weights

Figure 2 exemplifies the attention weights for a selection of satirical instances. In the first example the baseline model (no adv) focuses on a single word ("dpo" as a parody of the German newswire "dpa") which is unique to the publication the article was picked from ("Der Postillon"). In comparison the model using adversarial training ($\lambda = 0.2$) ignores this word completely and pays attention to "die Mordserie" ("series of murders") instead. In the second example, there are no words unique to a publication and the baseline spreads the attention evenly across all words. In contrast, the model with adversarial training is able to find cues for satire, being humor in this example ("family reunion [for refugees] is only allowed including mothers-in-law").

7 Conclusion and Future Work

We presented evidence that simple neural networks for satire detection learn to recognize characteristics of publication sources rather than satire and proposed a model that uses adversarial training to control for this effect. Our results show a considerable reduction of publication identification performance while the satire detection remains on comparable levels. The adversarial component enables the model to pay attention to linguistic characteristics of satire.

Future work could investigate the effect of other potential confounding variables in satire detection, such as the distribution of time and region of the articles. Further, we propose to perform more quantitative but also more qualitative analysis to better understand the behaviour of the two classifier configurations in comparison.

Acknowledgments

This work has been partially funded by the German Research Council (DFG), project KL 2869/1-1.

References

Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. In *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) at 23rd SIGKDD conference on Knowledge Discovery and Data Mining (KDD 2017)*, Halifax, Canada.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.

Jürgen Brummack. 1971. Zu Begriff und Theorie der Satire. *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte*, 45(1):275–377. In German.

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.

Lisa Colletta. 2009. Political Satire and Postmodern Irony in the Age of Stephen Colbert and Jon Stewart. *The Journal of Popular Culture*, 42(5):856–874.

Sohan De Sarkar, Fan Yang, and Arjun Mukherjee. 2018. Attending Sentences to detect Satirical Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3371–3380. Association for Computational Linguistics.

Harrison Edwards and Amos Storkey. 2016. Censoring Representations with an Adversary. In *International Conference on Learning Representations*.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(1):2030–2096.

Highet Golbert. 1962. *The Anatomy of Satire*. Princeton paperbacks. Princeton University Press.

Dan Goldwasser and Xiao Zhang. 2016. Understanding Satirical Articles Using Common-Sense. *Transactions of the Association for Computational Linguistics*, 4:537–549.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Cambridge, MA, USA. MIT Press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Ulrich Knoche. 1982. *Die römische Satire*. Orbis Biblicus Et Orientalis - Series Archaeologica. Vandenhoeck & Ruprecht. In German.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *International Conference on Learning Representations*.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3384–3393, Stockholmsmässan, Stockholm Sweden. PMLR.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop at International Conference on Learning Representations*.

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.

Johann Georg Sulzer. 1771. *Allgemeine Theorie der Schönen Künste*, 1. edition. Weidmann; Reich, Leipzig. In German.

James Thorne and Andreas Vlachos. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. In *Proceedings of the Conference on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, Stockholm, Sweden.

Fan Yang, Arjun Mukherjee, and Eduard Dragut. 2017. Satirical News Detection and Analysis using Attention Mechanism and Linguistic Features. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1979–1989. Association for Computational Linguistics.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 335–340, New York, NY, USA. ACM.