University of Stuttgart
Institute for
Natural Language Processing

Adversarial Training for Satire Detection: Controlling for Confounding Variables

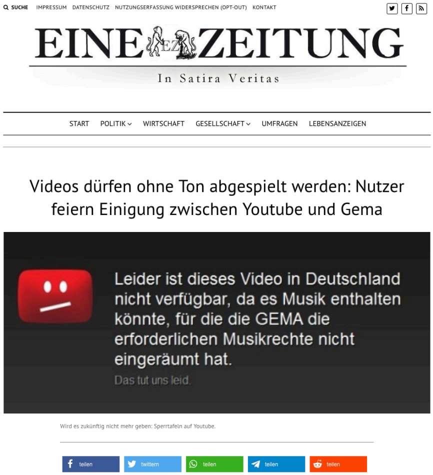
June 3rd, 2019

Robert McHardy, Heike Adel
and Roman Klinger

Motivation 1: Satire or not?

"After years of fighting there finally is a settlement between the Gema and Youtube . It became known today , that in future every music video is allowed to be played back in Germany again, as long as the audio is removed"

(translated from German)



Videos dürfen ohne Ton abgespielt werden: Nutzer feiern Einigung zwischen Youtube und Gema

Leider ist dieses Video in Deutschland nicht verfügbar, da es Musik enthalten könnte, für die die GEMA die erforderlichen Musikrechte nicht eingeräumt hat.

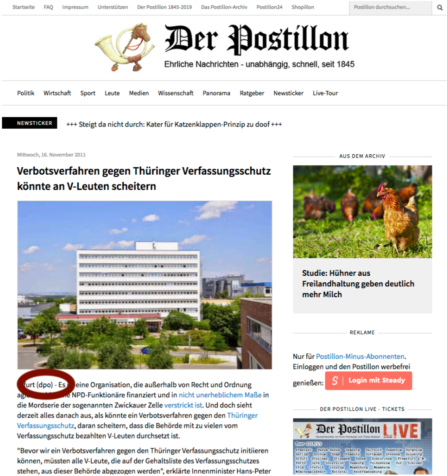
Das tut uns leid.

Wird es zukünftig nicht mehr geben: Sperrtafeln auf Youtube.

Motivation 2: Satire or not?

“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.”

(translated from German)



Startseite FAQ Impressum Unterstützen Der Postillon 1845-2019 Das Postillon-Archiv Postillon24 Shopillon Postillon durchsuchen...

Der Postillon

Ehrliche Nachrichten - unabhängig, schnell, seit 1845

Politik Wirtschaft Sport Leute Medien Wissenschaft Panorama Ratgeber Newsticker Live-Tour

NEWSTICKER +++ Steigt da nicht durch: Kater für Katzenklappen-Prinzip zu doof +++

Mittwoch, 16. November 2011

Verbotsverfahren gegen Thüringer Verfassungsschutz könnte an V-Leuten scheitern

furt (dpo) - Es eine Organisation, die außerhalb von Recht und Ordnung agi NPD-Funktionäre finanziert und in nicht unerheblichem Maße in die Mordserie der sogenannten Zwickauer Zelle verstrickt ist. Und doch sieht derzeit alles danach aus, als könnte ein Verbotsverfahren gegen den Thüringer Verfassungsschutz, daran scheitern, dass die Behörde mit zu vielen vom Verfassungsschutz bezahlten V-Leuten durchsetzt ist.

"Bevor wir ein Verbotsverfahren gegen den Thüringer Verfassungsschutz initiieren können, müssten alle V-Leute, die auf der Gehaltsliste des Verfassungsschutzes stehen, aus dieser Behörde abgezogen werden", erklärte Innenminister Hans-Peter

AUS DEM ARCHIV

Studie: Hühner aus Freilandhaltung geben deutlich mehr Milch

REKLAME

Nur für Postillon-Minus-Abonnenten. Einloggen und den Postillon werbefrei genießen: Login mit Steady

DER POSTILLON LIVE - TICKETS

Der Postillon LIVE
Tour 2018/19
Arnstadt Baden Baden Bamberg Bayreuth Bensheim Bergheim
Berlin Bochum Bonn Chemnitz Dortmund Dresden Duisburg
Erfurt Erkelenz Erlangen Essen Euskirchen Frankfurt a.M.
Fürth Gera Gütersloh Hamburg Hildesheim Hof Koblenz
Köln Krefeld Leipzig Magdeburg Mannheim
Mönchengladbach Nettetal Neubrandenburg

DPA is a German news agency – DPO does not exist (in this context).

Outline

1. Satire, Previous Work and Research Goals
2. Model and Data
3. Experiments & Results
4. Conclusion & Availability

Satire

- Form of art to critize in an entertaining manner
- Stylistic devices include humor, irony, sarcasm
- Goal: Mimic regular news in diction
- It's not misinformation or desinformation (fake news):
  Articles typically contain satire markers
  (similar to irony or sarcasm)

Automatic Satire Detection

Automatically distinguish satirical news from regular news
⇒ Challenging task (even for humans)

Previous Work

Yang et al. 2017, De Sarkar et al. 2018

- Created data sets which are automatically labeled from publication source
- Potential limitation: Models might learn characteristics of publication sources instead of actual characteristics of satire
- (evaluation is not faulty, they use different publication sources for validation than for training)

⇒ Bad generalization to unseen publication sources?

⇒ Interpretation of models (regarding concepts of satire) misleading?

Our Contributions

- We propose adversarial training: Improve robustness of model against confounding variable of publication sources
- We show that adversarial training is crucial for the model to pay attention to satire instead of publication characteristics
- We publish a large German data set for satire detection.
  - First dataset in German
  - First dataset including publication sources
  - Largest resource for satire detection so far

Outline

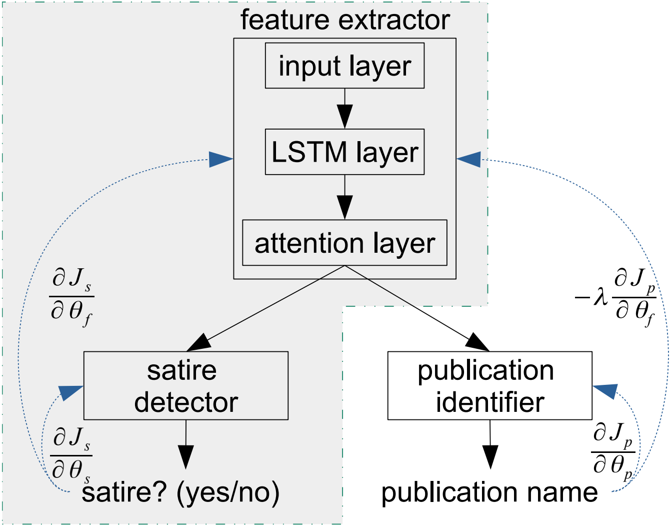
1 Satire, Previous Work and Research Goals

2 Model and Data

3 Experiments & Results

4 Conclusion & Availability

Model


feature extractor
input layer
LSTM layer
attention layer
$\frac{\partial J_s}{\partial \theta_f}$
$-\lambda \frac{\partial J_p}{\partial \theta_f}$
satire detector
publication identifier
$\frac{\partial J_s}{\partial \theta_s}$
$\frac{\partial J_p}{\partial \theta_p}$
satire? (yes/no)
publication name

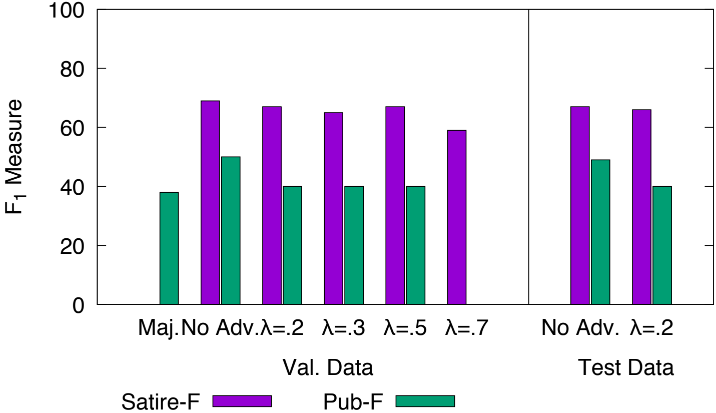
Data Collection and Selection

- Regular news:
  Der Spiegel, Der Standard, Die Zeit, Süddeutsche Zeitung
- Satire:
  Der Enthüller, Eulenspiegel, Nordd. Nach., Der Postillon, Satirepatzer, Die Tagespresse, Titanic, Welt (Satire), Der Zeitspiegel, Eine Zeitung, Zynismus24
- Articles from January 1st, 2000 and May 1st, 2018

		Average Length		
Publication	#Articles	Article	Sent.	Title
Regular	320,219	663.45	17.79	6.86
Satire	9,643	269.28	18.73	9.52

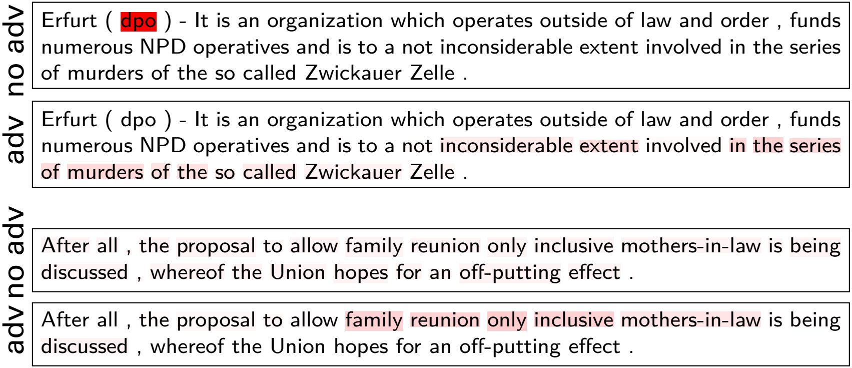
Research Question 1: Performance

How does a decrease in publication classification performance through adversarial training affect the satire classification performance?



Research Question 2: Attention Weights

Is adversarial training effective for avoiding that the model pays most attention to the characteristics of publication source rather than actual satire?



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 .

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 .

Conclusion and Availability

- Observation: Satire detection models learn characteristics of publication sources

Our Contributions

- Adversarial training to control for this confounding variable
  ⇒ Considerable reduction of publication identification performance while satire detection remains on comparable levels
  ⇒ Attention weights show effectiveness of our approach
- First German dataset for satire detection
  ⇒ Dataset and code available at:
  http://www.ims.uni-stuttgart.de/data/germansatire

University of Stuttgart
Institute for
Natural Language Processing

Adversarial Training for Satire Detection: Controlling for Confounding Variables

June 3rd, 2019

Robert McHardy, Heike Adel
and Roman Klinger