User Modeling

Finding Trolls and Troll Comments in News Community Forums

Lyubomira Dimitrova

NLP für Social Media
Institut für Computerlinguistik
Universität Heidelberg

20. Juni 2018

Gliederung

Wer ist ein Troll?

Meistens:

Nutzer, die durch *aufrührerische, beleidigende* oder *off-topic Kommentare* versuchen, andere Nutzer zu *ärgern* und zu *provozieren*, oft zur Unterhaltung. (e.g. Chen et al., 2012; Mojica and Ng, 2016)

Hier:

Nutzer, die absichtlich Fehlinformationen verbreiten, um andere Nutzer zu *betrügen* und zu *manipulieren*. (e.g. Adler et al., 2011; Mihaylov and Nakov, 2016)

Wer ist ein Troll?

- **theoretisch:** jemand, der versucht, die öffentliche Meinung zu manipulieren (und wird ggf. auch dafür bezahlt)

- **praktisch:** jemand, der von mindestens *n* anderen Nutzern als Troll bezeichnet wurde (erkannt durch Wörter wie z.B. *'troll'*, *'murzi(lka)'*)

Beispiel: Anklagen

To comment from "Historama": **Murzi**, *you know that you cannot manipulate public opinion, right?*

To comment from "Rozalina": You, **trolls**, *are so funny :) I saw the same signature under other comments :)*

Daten I

- Publikationen + Kommentare + (Nutzer-)Metadaten
- **Quelle:** eine der beliebtesten elektronischen Nachrichtenmedien in Bulgarien - `www.dnevnik.bg`
- **Kategorien:** Bulgarien, Europa, Welt $\rightarrow$ Politik
- **Zeitraum:** 1. Jan. 2013 - 1. Apr. 2015

Daten II

Object	Count
Publications	34,514
Comments	1,930,818
-of which replies	897,806
User profiles	14,598
Topics	232
Tags	13,575

Filter: nur Nutzer mit mind. 100 Komentare

$\rightarrow$ 317 Trolle + 964 nicht-Trolle

Daten III

35

zle_platen_trol
Рейтинг: 1229
12:38, 09 юни 18
Неутрално

До коментар [#30] от "S.A.": // Zum Kommentar [#30] von S.A.:

Всъщност с крайната омраза към България, българите и всичко българско, точно пък ти си типичен пример за " излишен " . Хейтър от, който няма абсолютно никаква полза.

– 4 + 7 !

36

S.A.
Рейтинг:
12:42, 09 юни 18
Неутрално

До коментар [#35] от "zle_platen_trol": // Zum Kommentar [#35] von zle_platen_trol.:

То по твоята логика и докторът, който диагностицира рака е хейтър на пациента. Ма нищо де, живей си с илюзията, че точно аз съм проблемът на Булгаристаня.

– 13 + 1 !

Features I

Aktivität:

- #Kommentare, #kommentierte Publikationen, #Tage seit Registrierung, #Tage mit mind. 1 Kommentar
- alle anderen Features durch diese skaliert

Gruppe	Motivation	Beispiel
Vote-basiert	Trolle bekommen mehr Downvotes	#Kommentare wo $\frac{Upvotes}{Downvotes} < 0.25$
Ähnlichkeit Publ.-Komm.	Trolle versuchen das Thema zu verdrehen	Kosinusähnlichkeit
Komm.-Reihenfolge	Trolle versuchen, unter die ersten zu sein	#Publikationen wo Kommentar in den ersten $k = 3$

Features II

Gruppe	Motivation	Beispiel
Ranking-basiert	Trolle äußern unpopuläre Meinungen	#Publikationen wo Kommentar der meist-downvoted ist
Antworten-basiert	Trolle kommentieren öfter, um Leute zu überzeugen	#Kommentare, die Antworten sind
Zeit-basiert	Trolle könnten bezahlt sein	#Kommentare postet zwischen 9 und 18 Uhr

+ nicht-skalierte Features

Setting

- **Datensatz:** 317 Trolle, 964 nicht-Trolle
- **Features:** normalisiert im Interval [-1, 1]
- **Learner:** SVM mit RBF Kernel (LIBSVM)
- 5-fold Crossvalidation
- **Baseline:** Majority class - 75.25%

Ergebnisse - Ablation

Features	Accuracy	Diff
AS + Non-scaled	94.37(+3.74)	19.13
AS – total comments	91.17(+0.54)	15.93
AS – comment order	91.10(+0.46)	15.85
AS – similarity	91.02(+0.39)	15.77
AS – time day of week	90.78(+0.15)	15.53
AS – trigg rep range	90.78(+0.15)	15.53
AS – time all	90.71(+0.07)	15.46
All scaled (AS)	**90.63**	**15.38**
AS – top loved/hated	90.55(-0.07)	15.30
AS – time hours	90.47(-0.15)	15.22
AS – vote u/down rep	90.47(-0.15)	15.22
AS – similarity top	90.32(-0.31)	15.07
AS – triggered cmnts	90.32(-0.31)	15.07
AS – is rep to has rep	90.08(-0.54)	14.83
AS – vote up/down all	89.69(-0.93)	14.44
AS – is reply	89.61(-1.01)	14.36
AS – up/down votes	88.29(-2.34)	13.04

Av. Accuracy (5-fold cross validation) und Unterschied zur Baseline

Ergebnisse - einzelne Features

Features	Accuracy	Diff
All Non-scaled	93.21	17.95
Only vote up/down	87.67	12.41
Only vote up/down totals	87.20	11.94
Only reply up/down voted	86.10	10.85
Only time hours	84.93	9.68
Only time all	84.31	9.06
Only is reply with rep	82.83	7.57
Only triggered rep range	82.83	7.57
Only day of week	82.28	7.03
Only total comments	82.28	7.03
Only reply status	80.72	5.46
Only triggered replies	80.33	5.07
Only comment order	80.09	4.84
Only top loved/hated	79.39	4.14
Only pub similarity top	75.25	0.00
Only pub similarity	75.25	0.00

Av. Accuracy (5-fold cross validation) und Unterschied zur Baseline

Experimente

Wer ist ein Troll?

min mentions	3	4	5	6
trolls	545	419	317	260
non-troll	964	964	964	964
Accuracy	85.49	87.85	90.87	92.32
Diff	+21.60	+18.15	+15.61	+13.56

Tabelle: Klassifikation von "mentioned" Trolle und nicht-Trolle

Konklusion

- **Aufgabe:** Anhand Nutzerverhalten Trolle erkennen
- **Features** beruhen fast nur auf Metadaten, keine Analyse des Inhalts
- **künftige Arbeit:** weitere, inhaltliche Features - POS, Named Entities...
- **endgültiges Ziel:** Erkennung von *bezahlten* Trollen

1. Trolle in Online Communities

2. Mihaylov et al. (2015)
   Finding Opinion Manipulation Trolls in News Community Forums

3. Mihaylov and Nakov (2016)
   Hunting for Troll Comments in News Community Forums

Trolle

- Nutzer, die versuchen, die öffentliche Meinung zu manipulieren
  - **genannte ('mentioned') Trolle**, die von mindestens *n* versch. User als solche bezeichnet wurden
  - **bezahlte (paid) Trolle**, die wegen durchgesickerten Reputationsmanagment-Vertragen bekannt geworden sind
- nicht-Trolle: Nutzer, die *nie* als Trolle bezeichnet wurden

Daten I

- **paid Troll-Kommentare:** alle Kommentare von Nutzern, deren Vertrage durchgesickert worden sind
- **'mentioned' Troll-Kommentare:** Kommentare, die als Reply eine Anklage bekommen haben

Beispiel: Anklage

"To comment from "Prorok Ilia": I can see that you are a red troll by the words that you are using."
$\rightarrow$ Der ursprüngliche Kommentar von 'Prorok Ilia' wird als Troll-Kommentar im DS aufgenommen

Daten II

- **nicht-Troll Kommentare:** zu jedem Troll-Kommentar, ein Kommentar aus demselben Thread (von einem Nutzer mit mind. 100 Kommentaren, der *nie* als Troll bezeichnet wurde)

Label	Comments
Paid troll comments	650
Mentioned troll comments	578
Non-troll comments	650+578

Tabelle: Datensätze

Kommentar-Features I

- Bag of Words (ohne Stopwörter), Bag of Stems
- Wort N-Gramme (2-3), Zeichen N-gramme (3-4), Wort-Präfix und -Suffix (erste bzw. letzte 3-4 Zeichen)
- #Emoticons, #Punktuation, #Tokens, #ALLCAPS
- Metadaten: Zeit (Werktag/Wochenende; während der Arbeitszeit/Nachts)
- Word2Vec Cluster: Modell trainiert auf alle Publikationen und Kommentare; K-Means: 5,372 Wortvektoren-Cluster

Kommentar-Features II

- Sentiment: übersetzte Lexika; Sentimentanalyse-Pipeline, angepasst fürs Bulgarische
- #Bad Words: übersetztes Lexikon; Ergänzung: zu jedem Wort die 3 ähnlichsten w2v Vektoren
- #Mentions: neuerstellte Lexika mit Namen und Spitznamen bulgarischer Politiker
- POS-Tags: unterschiedliche Granularität - *Npmsi, Np, N*; z.B. $\frac{\#Np}{\#tokens}$ als Feature
- Named Entities - übersetztes Lexikon

Experimente

- **Datensätze:** 650/650 paid Troll/nicht-Troll Kommentare, 578/578 'mentioned' Troll/nicht-Troll Kommentare
- **Features:** normalisiert im Interval [0, 1]
- **Learner:** Logistic Regression mit L2-Regularisierung (LIBLINEAR)

Ergebnisse - Ablation

Features	F	Acc
All – char n-grams	79.24	78.54
All – word suff	78.58	78.20
All – word preff	78.51	78.02
All – bow stems	78.32	77.85
All – bow with stop	78.25	77.77
All – bad words	78.10	77.68
All – emoticons	78.08	77.76
All – mentions	78.06	77.68
All	78.06	77.68
All – (bow, no stop)	78.04	77.68
All – NE	77.98	77.59
All – sentiment	77.95	77.51
All – POS	77.80	77.33
All – w2v clusters	77.79	77.25
All – word 3-grams	77.69	77.33
All – word 2-grams	77.62	77.25
All – punct	77.29	76.90
All – metadata	70.77	70.94
Baseline	50.00	50.00

mentioned troll vs non-troll

Features	F	Acc
All – char n-grams	81.08	81.77
All – word suff	81.00	81.77
All – word preff	80.83	81.62
All – bow with stop	80.67	81.54
All – sentiment	80.63	81.46
All – word 2-grams	80.62	81.46
All – w2v clusters	80.54	81.38
All – word 3-grams	80.46	81.38
All – punct	80.40	81.23
All – mentions	80.40	81.31
All	80.40	81.31
All – bow stems	80.37	81.31
All – emoticons	80.33	81.15
All – bad words	80.09	81.00
All – NE	80.00	80.92
All – POS	79.77	80.69
All – (bow, no stop)	79.46	80.38
All – metadata	75.37	76.62
Baseline	50.00	50.00

paid troll vs non-troll

Einzelne Featuregruppen

Features	F	Acc
All	78.06	77.68
Only metadata	84.14	81.14
Sent,bad,pos,NE,meta,punct	77.79	76.73
Only bow, no stop	73.41	73.79
Only bow with stop	73.41	73.44
Only bow stems	72.43	72.49
Only word preff	71.11	71.62
Only w2v clusters	69.85	70.50
Only word suff	69.17	68.95
Only word 2-grams	68.96	69.29
Only char n-grams	68.44	68.94

Features	F	Acc
Only word 3-grams	64.74	67.21
Only POS	64.60	65.31
Sent,bad,pos,NE	63.68	64.10
Only sent,bad	63.66	64.44
Only emoticons	63.30	64.96
Sent,bad,ment,NE	63.11	64.01
Only punct	63.09	64.79
Only sentiment	62.50	63.66
Only NE	62.45	64.27
Only mentions	62.41	64.10
Only bad words	62.27	64.01
Baseline	50.00	50.00

mentioned troll vs non-troll Kommentare

Einzelne Featuregruppen

Features	F	Acc
All	80.40	81.31
Sent,bad,pos,NE,meta,punct	78.04	78.15
Only bow, no stop	75.95	76.46
Only word 2-grams	75.55	74.92
Only bow with stop	75.27	75.62
Only bow stems	75.25	76.08
Only w2v clusters	74.20	74.00
Only word preff	74.01	74.77
Sent,bad,pos,NE	73.89	73.85
Only metadata	73.79	72.54
Only char n-grams	73.02	74.23

Features	F	Acc
Only POS	72.94	72.69
Only word suff	72.03	72.69
Only word 3-grams	69.20	68.00
Only punct	66.80	65.00
Only NE	66.54	64.77
Sent,bad,ment,NE	66.04	64.92
Only sentiment	64.28	62.62
Only mentions	63.28	61.46
Only sent,bad	63.14	61.54
Only emoticons	62.95	61.00
Only bad words	62.22	60.85
Baseline	50.00	50.00

paid troll vs non-troll Kommentare

'A Witch Hunt'

Problem: Anklagen als Grundwahrheit zu nehmen ist wenig sinnvoll; manche 'Trolle' evtl. ungerecht angeklagt

Idee: Vergleiche Nutzer mit mehreren als Troll bezeichneten Kommentaren mit nicht-Troll Nutzern.

	5	**10**	**15**	**20**
Acc	80.70	81.08	83.41	85.59
Diff	+8.46	+18.51	+30.81	+32.26

Abbildung: Erkennung von Nutzern mit 5, 10... Troll-Kommentaren

Weitere Arbeiten

Zannettou et al. (2018)

- Analyse des Verhaltens russischer Trolle ('state-sponsored actors') auf Twitter
- Geolocation, Inhalt des Tweets
- Bewertung des Einflusses solcher Trolle auf das soziale Netzwerk

Vielen Dank für die Aufmerksamkeit!

Fragen?

Lexika I

Sentiment	MPQA: (Wilson et al., 2005) http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/ NRC: (Mohammad and Turney, 2013) http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm (Hu and Liu, 2004) https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
Bad Words	https://web.archive.org/web/20130704010355/http://urbanoalvarez.es:80/blog/2008/04/04/bad-words-list/
Bulgarian	https://github.com/tbmihailov/gate-lang-bulgarian-gazetteers/

'SVM with RBF kernel'

$C = 32$, $\gamma = 0.0078125$

$$w = \underset{w,\xi}{\operatorname{argmin}} \ \frac{1}{2}\|w\|^2 + C\sum_i^n \xi_i$$

$$\textit{sodass} \quad y_i(w \cdot \phi(x_i)) \geq 1 - \xi_i, \ \xi_i \geq 0$$

C - Regularisierungsparameter; je höher, desto mehr werden Fehler bestraft

$$\textit{Radial basis function kernel}: \quad K(x, x') = \exp(-\gamma \left\|x - x'\right\|^2)$$

Referenzen I

Adler, B. T., L. De Alfaro, S. M. Mola-Velasco, P. Rosso, and A. G. West
2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *International Conference on Intelligent Text Processing and Computational Linguistics*, Pp. 277–288. Springer.

Chen, Y., Y. Zhou, S. Zhu, and H. Xu
2012. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, Pp. 71–80. IEEE.

Hu, M. and B. Liu
2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Pp. 168–177. ACM.

Mihaylov, T., G. Georgiev, and P. Nakov
2015. Finding opinion manipulation trolls in news community forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, Pp. 310–314.

Referenzen II

Mihaylov, T. and P. Nakov
2016. Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, Pp. 399–405.

Mohammad, S. M. and P. D. Turney
2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Mojica, L. G. and V. Ng
2016. Modeling trolling in social media conversations. *arXiv preprint arXiv:1612.05310*.

Wilson, T., J. Wiebe, and P. Hoffmann
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.

Referenzen III

Zannettou, S., T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn
2018. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. *arXiv preprint arXiv:1801.09288*.