Identifying Right-Wing Extremism in German Twitter Profiles: a Classification Approach

Matthias Hartung[1,2], Roman Klinger[2,3], Franziska Schmidtke[4], and Lars Vogel[4]

[1] Semantic Computing Group, CITEC, Bielefeld University
[2] Semalytix GmbH, Bielefeld
[3] Institut für Maschinelle Sprachverarbeitung, University of Stuttgart
[4] Kompetenzzentrum Rechtsextremismus, Friedrich-Schiller-Universität Jena
mhartung@cit-ec.uni-bielefeld.de, klinger@ims.uni-stuttgart.de,
{franziska.schmidtke, lars.vogel}@uni-jena.de

Abstract. Social media platforms are used by an increasing number of extremist political actors for mobilization, recruiting or radicalization purposes. We propose a machine learning approach to support manual monitoring aiming at identifying right-wing extremist content in German Twitter profiles. We frame the task as profile classification, based on textual cues, traits of emotionality in language use, and linguistic patterns. A quantitative evaluation reveals a limited precision of 25 % with a close-to-perfect recall of 95 %. This leads to a considerable reduction of the workload of human analysts in detecting right-wing extremist users.

Keywords: extremism monitoring, classification, social media

1 Introduction

Recent years have seen a dramatic rise in importance of social media as communication channels for political discourse [9]. Various political actors use different kinds of social platforms to engage directly with voters and supporter networks in order to shape public discussions, induce viral social trends, or spread political ideas and programmes for which they seek support.

With regard to extremist political actors and parties, a major current focus is on recruiting and radicalizing potential activists in social media. For instance, the American white nationalist movements have been able to attract a 600 % increase of followers on Twitter since 2012 [3]. Twitter is comparably undermoderated in comparison to other platforms (*e. g.*, YouTube or Facebook[5]) and can therefore be seen as the predestinated channel for such activities [4].

Growing efforts are spent into monitoring extremist activities in social media by state institutions, platform providers or companies. Extremism monitoring aims at detecting *who* is active (possibly separating opinion leaders from followers, and discovering dynamics of network evolution), *what* they say (identifying

[5] https://www.facebook.com, https://www.youtube.com

prominent topics and possibly hate speech or fake news), and *which purpose* they pursue (revealing strategies such as mobilization or recruiting). Currently, these goals are mostly pursued in time-consuming manual work [1].

We propose a method to support the identification of extremist users in Twitter and aim at detecting right-wing extremist content in German Twitter profiles, based on lexical information and patterns of emotionality underlying language use. Our application scenario is an instance of semi-automatic knowledge base completion in that automated classification methods are applied in order to facilitate manual efforts to grow the data pool of known German right-wing extremists. We are aware that binary categorization of political attitudes is an oversimplification; yet, we consider this a valid approach for the given use case, as it is not intended to work in isolation from experts' decisions.

2 Related Work

There is only limited work with a focus on right-wing extremist detection. However, other forms of extremism have been the subject of research. As an early example, Ting et al. aimed at identification of hate groups on Facebook [15]. They build automatic classifiers based on social network structure properties and keywords. While this work focused on detection of groups, Scanlon et al. dealt with specific events of interaction, namely the recruitment of individuals [12] on specific extremist's websites. Their domain are Western Jihadists. In contrast, Ashcroft et al. identify specific messages from Twitter [2]. Similarly, Wei et al. identify Jihadist-related conversations [16].

Similar to our work is the approach of Ferrara et al., who identify ISIS members among Twitter users [6]. Kaati et al. identify multipliers of Jihadism on Twitter, as their aim is also the identification of specific profiles, for a different domain, though [7].

3 Profile Classification

Right-wing extremism is an ideology of asymmetric quality of social groups, defined by race, ethnicity or nationality, and a related authoritarian concept of society. Table 1 shows an overview of the conceptial dimensions of right-wing extremism, following Stöss [13].

Right-wing extremism is defined by adopting all or at least a majority of the attitudes in Table 1. It is, accordingly, appropriate to investigate entire Twitter profiles rather than individual Tweets. Therefore, we frame the task of detecting right-wing extremism in Twitter as supervised classification of user profiles into the target categories R (right-wing extremist) and N (non-extremist). We use support vector machines with a linear kernel [5].

Under the assumption that linguistic variables serve as informative predictors of users' underlying attitudes, we mainly focus on the vocabulary and certain semantic patterns as features of the model, as described in the following.

Table 1. Conceptual dimensions of right-wing extremism (following Stöss [13])

Dimension	Definition
National-chauvinism	presumed superiority and demanded homogeneity of the in-group
Xenophobia	imagined inferiority of out-groups and potential threat to in-group
Racism	definition of in- and out-groups strictly in terms of race
Social Darwinism	imagined homogeneity and purity of own race; fight between races as unavoidable means to leverage survical of the strongest race
Support of dictatorship	perception of democracy as weakening the in-group by substituting violent struggle by peaceful competition, negotiation and universal rights
National socialism	glorification of historical national socialism by referring to its symbols (using symbolic codes) or denial of the Holocaust

Lexical Features. We create a *bag-of-words frequency profile* of all tokens (unigrams and bigrams) used by an author in the entirety of all messages in their profile. Stopwords, Twitter-specific vocabulary such as "RT" (indicating re-tweets) and short links (URLs referring to websites external to Twitter) are filtered, while keeping hashtags and references to other Twitter users.

Emotion Features. We estimate a single label classification model for emotional categories, *viz.*, anger, disgust, fear, joy, love, sadness, shame, surprise, trust (motivated by Plutchik [10]) on a sample of 1.2M English and German Tweets from March until November 2016.[6] Similarly to Suttles et al. [14], we follow a weak supervision approach by utilizing emotion hashtags. As features in our downstream prediction model, we use confidence scores for each emotion as provided by this classifier.

Pro/Con Features. We use lexico-syntactic patterns capturing the main political goals or motives to be conveyed by an author in their messages:

gegen ... `<NOUN>` / against ... `<NOUN>`
`<NOUN>` ... statt ... `<NOUN>` / `<NOUN>` ... instead of ... `<NOUN>`

Social Identity Features. Based on the assumption that collective identities are constructed by means of discursive appropriation, we apply another pattern to detect real-world entities that are recurrently used in appropriation contexts:

`unser[e|en|em|er|es]? ... <NOUN>` / `our ... <NOUN>`

Transformation of Feature Values. The resulting feature vectors are transformed by the tfidf metric [8] in order to increase the relative impact of features that are (i) prominent in the respective profile and (ii) bear high discriminative power, *i. e.*, they occur in a relatively small proportion of all profiles in the data.

4 Evaluation

4.1 Data Collection and Annotation

Annotations are provided by domain experts at the level of individual user profiles. These annotations comprise 37 *seed profiles* of political actors from the

[6] All English Tweets are translated to German via Google translate (`http://translate.google.com`) to receive a more comprehensive training set.

Table 2. Classification results in detecting profiles of category R, using 10-fold cross-validation on seed profiles (left part), a held-out test set and a restricted test set comprising only profiles with at least 100 Tweets (right part)

Features	Precision	Recall	F_1 Score
all	0.87	1.00	0.93
BOW	0.91	1.00	0.95
Bigrams	0.80	1.00	0.89
Emotions	0.78	0.90	0.84
Pro/Con	0.63	1.00	0.77
Identity	0.70	0.95	0.81
Baseline	0.54	1.00	0.70

	Precision	Recall	F_1 Score
Test Set	0.25	0.95	0.40
Baseline	0.19	1.00	0.32
Test $Set_{>100}$	0.32	0.92	0.47
Baseline	0.21	1.00	0.35

German federal state Thuringia. They are split into 20 profiles labeled as right-wing and 17 non-extremist ones. Right-wing seed profiles contain organizations as well as leading individuals within the formal and informal extremist scene as documented by Quent et al. [11]. Non-extremist seed profiles contain political actors of the governing parties and single-issue associations [11]. Using the Twitter REST API[7], entire timelines (comprising all Tweets by the respective user until December 15, 2016) are acquired for all seed profiles and their followers.

4.2 Experimental Results

We train a classification model on the seed profiles (comprising 45,747 Tweets in total, among them 15,911 of category R and 29,836 of category N) with all features described in Section 3.

Results of a *10-fold cross-validation on the seed profiles* can be seen from Table 2 (left part). We compare the performance of the classifier relying on all features, each feature group in isolation, and a majority baseline categorizing all profiles as "R". All feature groups turn out as reliable predictors of political orientation. Bag-of-words (BOW) lexical cues have the strongest individual contribution and cannot be outperformed by any other feature combination.

The *test set* comprises 100 randomly sampled profiles from followers of the seed users which have been annotated by one of the authors of this paper. Table 2 (right part) shows the results of this evaluation: In comparison to cross-validation on the seed profiles, the performance drops considerably, while still outperforming the baseline of categorizing all profiles as "R" by 8 points in F_1 score. Limiting the test data to profiles containing at least 100 Tweets each (*cf. Test* $Set_{>100}$ in the table) yields an improved performance of F_1=0.47.

4.3 Qualitative Discussion

We perform a feature analysis by sorting features according to their weight in the model and manually categorize them into different semantic classes (*word*

[7] `https://dev.twitter.com/rest/public`

Table 3. Top features from a global list of 100 features categorized in feature classes for category N and R, according to feature weights in the model.

	Category N	Category R
words	heute, wir, beim, 's, the, uhr, geht, für_die, to, of, glückwunsch, Stellenangebot, feuerwehren, morgen, danke	d18, deutschland, #wehrdich, brd, deutsche, DTL, asylanten, Bürgerinitiative_wir, demonstration, asyl, herbstoffensive, asylbewerber, durchgeführt_!, volk, !_!
network	nabu, ldkth, awo, mikemohring, bodoramelow, mdr_th, cdu_fraktion, naju, lmvth, der_AWO, lakoth16, DGB_Jugend, jef_de, dgb, bund_net	NPD, FN, suegida, identitäreaktion, weimarticker, sternde, spiegelonline, der_NPD, goldenerlöwe, npdde, Thüringer_NPD, agnordost, antifa, NPD-Landesverband, AB_ERFURT
local	erfurt, in_Erfurt	hildburghausen, Jena, nordhausen, Saalfeld, Erfurt_), suhl, Webetgasse, Bauvereinstraße, Hußstraße, Gera_Zschochernstraße
regional	Thüringen, thüringer, in_Thüringen, r2g, hochwasser, gebietsreform	kyffhäuserkreis, wir_lieben, lieben_Thüringen
national/global	TTIP, #mitredeneu, Bund, ews2014	#merkelmussweg, genozid
emotionality	— None —	`EMO_TWEETS_PROP>0.3`, `MOST_FREQ_EMOTION=Disgust`

features, references to other profiles in the *network* structure, *emotionality*, and topics of *local*, *regional* or *global* geographical scope), shown in Table 3.

We find that users from category N presumably aim at promoting other users only if they share similar political orientations. Contrarily, right-wing users also follow delimitation strategies, by referring to media platforms (e.g., *sternde*) or left-wing political groups (e.g., *antifa*) in provocative or offensive ways. In general, the tonality in the language used by the right-wing authors in our data is much more aggressive (e.g., *wehr dich*/*fight back*, *Herbstoffensive*/*autumn offensive*) and characterized by emphasizing identitary contrasts (e.g., *Deutsche* vs. *Asylanten*/*Germans* vs. *refugees*) pointing into the direction of nationalism or chauvinism. Among the most discriminative features in our model are codes such as *d18*, which are indicative of a glorification of national socialism. Beyond that, we find that a substantial proportion of emotionally rich messages or messages conveying disgust are good predictors of right-wing extremism in our data. These characteristics clearly reflect aspects involved in the conceptual definition of right-wing extremism (cf. Table 1), which we consider supportive evidence in favor of the plausibility of our approach from a theoretical perspective.

5 Conclusions and Outlook

We presented a machine learning approach to identify Twitter profiles which correspond to right-wing extremism. Our work aims at supporting human experts in their monitoring activities which are currently carried out purely manually. Our classification model achieves high recall on right-wing extremist profiles

such that only very few candidates are missed. At the same time, inconspicuous profiles are effectively filtered, which reduces the work load by 25 %.

With its focus on discrete classification, the present study is certainly affected by an oversimplification, as the spectrum of political attitudes is more complex than only the two target categories considered. We are currently working on a ranking model based on similarity metrics in order to project unseen profiles on a continuous scale of political attitudes. In future work, we aim at developing this method further into a learning-to-rank approach. In addition, we propose the development of features that are based on deeper methods of natural language analysis in order to be able to address more fine-grained aspects in the conceptualization of right-wing extremism.

References

1. Amadeu-Antonio-Stiftung: Rechtsextreme und menschenverachtende Phänomene im Social Web. https://www.amadeu-antonio-stiftung.de/w/files/pdfs/monitoringbericht-2015.pdf (2016)
2. Ashcroft, M., Fisher, A., Kaati, L., Omer, E., Prucha, N.: Detecting Jihadist messages on Twitter. In: Proc. of EISIC (2015)
3. Berger, J.: Nazis vs. ISIS on Twitter. A Comparative Study of White Nationalist and ISIS Online Social Media Networks. Tech. rep., Center for Cyber and Homeland Security, George Washington University, Washington, D.C. (2016)
4. Blanquart, G., Cook, D.: Twitter Influence and Cumulative Perceptions of Extremist Support. A Case Study of Geert Wilders. In: Proc. of ACTC (2013)
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20, 273–297 (1995)
6. Ferrara, E., Wang, W.Q., Varol, O., Flammini, A., Galstyan, A.: Predicting online extremism, content adopters, and interaction reciprocity. arxiv: https://arxiv.org/abs/1605.00659 (2016)
7. Kaati, L., Omer, E., Prucha, N., Shrestha, A.: Detecting multipliers of jihadism on twitter. In: IEEE Int. Conference on Data Mining Workshop (ICDMW) (2015)
8. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
9. Parmelee, J., Bichars, S.: Politics and the Twitter Revolution. Lexington Books, Landham, MD (2013)
10. Plutchik, R.: The nature of emotions. American Scientist (2001)
11. Quent, M., Salheiser, A., Schmidtke, F.: Gefährdungen der demokratischen Kultur in Thüringen. http://www.denkbunt-thueringen.de/wp-content/uploads/2016/02/Gef%C3%A4hrdungsanalyse.pdf (2016)
12. Scanlon, J.R., Gerber, M.S.: Automatic detection of cyber-recruitment by violent extremists. Security Informatics 3(1), 5 (2014)
13. Stöss, R.: Rechtsextremismus im Wandel. Friedrich-Ebert-Stiftung, Berlin (2010)
14. Suttles, J., Ide, N.: Distant Supervision for Emotion Classification with Discrete Binary Values. Springer, Berlin, Heidelberg (2013)
15. Ting, I.H., Chi, H.M., Wu, J.S., Wang, S.L.: An Approach for Hate Groups Detection in Facebook. Springer Netherlands (2013)
16. Wei, Y., Singh, L., Martin, S.: Identification of extremism on twitter. In: Int. Conference on Advances in Social Networks Analysis and Mining (2016)