Ranking Right-Wing Extremist Social Media Profiles by Similarity to Democratic and Extremist Groups

Matthias Hartung
CITEC, Bielefeld University

Roman Klinger
IMS, University of Stuttgart

Franziska Schmidtke and **Lars Vogel**
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 are used by an increasing number of political actors. A small subset of these is interested in pursuing extremist motives such as mobilization, recruiting or radicalization activities. In order to counteract these trends, online providers and state institutions reinforce their monitoring efforts, mostly relying on manual workflows. We propose a machine learning approach to support manual attempts towards identifying right-wing extremist content in German Twitter profiles. Based on a fine-grained conceptualization of right-wing extremism, we frame the task as ranking each individual profile on a continuum spanning different degrees of right-wing extremism, based on a nearest neighbour approach. A quantitative evaluation reveals that our ranking model yields robust performance (up to 0.81 F_1 score) when being used for predicting discrete class labels. At the same time, the model provides plausible continuous ranking scores for a small sample of borderline cases at the division of right-wing extremism and New Right political movements.

1 Introduction

Recent years have seen a dramatic rise in importance of social media as communication channels for political discourse (Parmelee and Bichars, 2013). Political actors use social platforms to engage directly with potential 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 (Berger, 2016). Twitter is comparably under-moderated in comparison to other platforms and therefore constitutes a predestinated channel for such activities (Blanquart and Cook, 2013).

State institutions, platform providers or companies spend growing efforts into monitoring extremist activities in social media. Extremism monitoring aims at detecting *who* is active (possibly separating opinion leaders from adopters, and discovering dynamics of network evolution), *what* they say (identifying prominent topics and possibly hate speech or fake news), and *which purpose* they pursue (revealing strategic objectives such as mobilization or recruiting). Currently, these goals are mostly pursued in time-consuming manual work. For instance, the Amadeu Antonio foundation, a non-governmental organization countering right-wing extremism in Germany, conducts an annual report that relies on a "qualitative method" (Amadeu-Antonio-Stiftung, 2016). Furthermore, the Anti Defamation League issued a report on anti-semitic harassment on Twitter, based on manually reviewed 2.6 million Tweets (ADL, 2016).

In this paper, we propose an approach to support the first of the above-mentioned aspects, *i. e.*, the identification of extremist users in Twitter. In particular, we aim at detecting potential right-wing extremist content in German Twitter profiles, based

Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 24–33
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

on lexical information and patterns of emotion underlying language use (cf. Ghazi et al., 2010; Suttles and Ide, 2013; Wang et al., 2012). Contrary to previous work (Hartung et al., 2017), we phrase the problem as ranking between manually selected groups of Twitter profiles which constitute seeds of right-wing extremists and non-extremist users. We show that our ranking model achieves robust performance in discrete binary categorizations, while also being capable of predicting plausible continuous ranking scores for a sample of borderline cases which specifically address the notoriously hard delimitation of right-wing extremism from New Right political movements in Germany and Europe. This lazy machine learning approach outperforms the eager method proposed in previous work on the same data set (Hartung et al., 2017).

2 Background and Related Work

Background. 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. It encompasses aggressive behavior and the underlying attitudes of *xenophobia*, *racism*, *anti-Semitism*, *social Darwinism*, as well as *national chauvinism*, *glorification of the historical national socialism* and *support for dictatorship* (Stöss, 2010).

When transforming this concept into patterns used in Twitter communication, certain domain-specific contextual opportunities and restrictions have to be considered. First, Tweets are motivated by latent attitudes, but they are manifest communicative behavior. The transformation of attitudes into behavior is, however, conditional. While attitudes are usually revealed in the secrecy of anonymous interviews, Twitter requires to display attitudes in public. This may lead to strategies of camouflage and the use of codes. Second, these attitudes are revealed by commenting on particular topics requiring that their changing saliency over time must be considered. Third, expressing some of these attitudes publicly in a particular manner can become relevant to criminal law. Thus, especially the glorification of national socialism is not suited to serve as a distinctive criterion, since its public expression in a non-subtle manner is avoided by Twitter users. Finally, research has repeatedly demonstrated that *some* of the attitudes mentioned above (*e. g.*, xenophobia) are widespread among the German population (Best et al., 2016; Zick et al., 2016), whereas right-wing extremism is defined by adopting *all* or at least a *majority* of these attitudes.

Related Work. There is only limited work with a focus on right-wing extremism detection. However, other forms of extremism have been the subject of research. As an early example, Ting et al. (2013) aim at identification of hate groups on Facebook. They build automatic classifiers based on social network structure properties and keywords. While this work focuses on detection of groups, Scanlon and Gerber (2014) deal with specific events of interaction, namely the recruitment of individuals on specific extremist's websites. Their domain are Western Jihadists. In contrast, Ashcroft et al. (2015) identify specific messages from Twitter. Similarly, Wei et al. (2016) identify Jihadist-related conversations.

Recently, the identification of Twitter users displaying different traits or attitudes of extremism has gained growing attention. For instance, Ferrara et al. (2016) identify ISIS members among Twitter users, while Kaati et al. (2015) focus on multipliers of Jihadism on Twitter. In very recent work, Wei and Singh (2017) present an approach to detecting Jihadism on Twitter both at the level of user profiles and individual Tweets, using a graph-based approach. The only approach towards automated detection of right-wing extremist users on Twitter we are aware of is our previous work (Hartung et al., 2017).

As a common assumption, all of the latter models rely on discrete output spaces; more specifically, they frame the profile identification task as a binary classification problem. In this paper, we argue that this assumption is overly simplistic as (i) it obscures the complexity of the spectrum of political attitudes, and (ii) it is unable to capture different degrees of radicalization. Therefore, we propose a ranking approach which is capable of projecting user profiles to a continuous range spanning different degrees of similarity to known (groups of) right-wing extremist or non-extremist users.

Extremism detection can also be seen as special case of profiling users of social network platforms in a more general way, *e. g.*, classification of personality traits (Golbeck et al., 2011; Quercia et al., 2011). Such approaches can be seen as extensions to sentiment analysis in general (Liu, 2015). More recently, there is a growing interest in particular aspects such as hate speech (Schmidt and Wiegand, 2017; Waseem and Hovy, 2016), racism (Waseem, 2016), violence or threat detection (C. Basave et al., 2013; Wester et al., 2016).

3 Profile Ranking

Right-wing extremism is defined by adopting all or at least a majority of the attitudes mentioned in Section 2. It is, accordingly, appropriate to investigate entire Twitter profiles rather than individual Tweets. We frame the task of detecting right-wing extremism in Twitter as ranking of user profiles according to their relative proximity to (groups of) other users in high-dimensional vector space.

3.1 Conceptualizing the Dimensions of Right-Wing Extremism

Our approach is based on the general assumption that linguistic variables serve as informative predictors of user's underlying attitudes. We mainly focus on the vocabulary and certain semantic patterns the use of which may be considered as communicative behavior that is motivated by the ideology of right-wing extremism. In the following, we justify this choice by a more thorough description of the conceptual dimensions of right-wing extremism (as introduced in Section 2) and highlight presumable links to linguistic behavior.

National-chauvinism. Migration is currently the most salient topic of German right-wing extremism, touching upon the attitudes of *national-chauvinism* combined with *xenophobia*. In the view of right-wing extremists, migration is perceived as a threat to the homogeneity of the superior German nation (in-group) by migrants from inferior nations (out-group). National-chauvinism expresses the presumed superiority and demanded homogeneity of the in-group, while xenophobia encompasses the imagined inferiority of the out-group and its potential threat to the in-group. Relevant words and hashtags may be "Rapefugees" or "Invasoren" ("invaders"), for instance.

Racism. Although related to national-chauvinism and xenophobia, *racism* is distinct, since it defines the in- and out-group in terms of race rather than nationality. Racism becomes especially obvious with references to the physical appearance of out- and in-group members, as expressed by, *e. g.*, "Neger" ("nigger"), #whitepower or #whiteresistance.

Social Darwinism builds upon racism, but claims that fight either between or within races is an unavoidable means to leverage the survival of the strongest race. Violence is legitimated as a basic law of society and any deviation from violence, *e. g.*, by peaceful agreement, is considered to undermine the chances for survival and is thus illegitimate. The imagined homogeneity and purity of the own race needs to be defended; hence, political opponents and other people who are perceived as not fitting are considered as enemies who can be fought without any reservation. Indicative are thus words and semantic structures which aggressively offend the opponents as enemies refusing any agreement with them, *e. g.*, "Abschaumpresse" ("scum press"), "Volksverräter" ("betrayer of the nation"). Expressions conveying negative emotions such as anger or disgust when referring to opponents may be indicative as well.

Democracy vs. dictatorship. In turn, democracy is considered as weakening the in-group by substituting violent struggle by peaceful competition, negotiation and acceptance of universal rights. Instead, *dictatorship* is favored, since given the homogeneity of the nation or the race, political parties and their competition is considered needless. In the current debate on migration, the rejection of democracy has been fused with conspiracy theories. Indicative for the rejection of democracy and accompanying conspiracy theories are vocabulary like "Lügenpresse" ("lying press"), "Gehirnwäsche" ("brainwash"), or #stopislam.

National socialism. The glorification of the historical *national socialism* by explicitly referring to its symbols or the denial of the Holocaust is relevant to German criminal law. However, using legal references to national socialism or symbolic codes can circumvent this restriction. Indicative are words and number codes like "Heil", 18 or 88 (one and eight representing the letters A and H, respectively, thus abbreviating "Heil Hitler" or "Adolf Hitler").

Additionally, indications of behavior clearly associated to right-wing extremist organizations or parties can be used to classify the profiles. Indicative are therefore expressions of approval, affinity of even membership in such organizations, for instance by following them, or posting hashtags in an affirmative manner such as #NPD, #DritteWeg, #Die Rechte (all referring to German right-wing extremist parties).

3.2 Features

In this section, we describe how the previously discussed dimensions of right-wing extremism are incorporated as features into our ranking model.

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 after stopword filtering. This frequency profile is able to capture lexical expressions described in the previous section. Twitter-specific vocabulary such as "RT" (indicating re-tweets) or short links (URLs referring to websites external to Twitter) are filtered; however, hashtags and references to other Twitter users (*e. g.*, `@NPD`) are kept in the lexical profile.

Emotion Features. Similarly to previous research on emotion detection on Twitter (Ghazi et al., 2010; Suttles and Ide, 2013; Wang et al., 2012), we estimate a single-label classification model for various emotion categories, *viz.*, anger, disgust, fear, joy, love, sadness, shame, surprise, trust (motivated by fundamental emotions (Ekman, 1970; Plutchik, 2001)) on a subsample of approx. 1.2 Million English and German Tweets from March 2016 until November 2016. All English Tweets are machine translated to German via Google translate[1] to receive a more comprehensive training set. We use a weak supervision approach by utilizing the emotion hashtags (which are disregarded during training). As features in our downstream ranking model, we use confidence scores derived from the single-label classification model (capturing the most prominent emotions and the proportion of emotionally charged Tweets per user profile).

Pro/Con Features. We use lexico-syntactic patterns encoding shallow argumentation patterns to capture 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>

As a fundament to apply these rules, noun detection is performed with regular expressions for capitalization, which works well in German, instead of incorporating a full-fledged (and slower) part-of-speech tagger. An arbitrary number of intermediate tokens is accepted between the prepositional cue and the closest subsequent noun denoting the objective of support or disaffirmation.

The following examples[2] illustrate these patterns (pro and con objectives in boldface):

(1) a. *#Muslimefürfrieden bringen Antwort auf die Broschüre der AfD in die Öffentlichkeit: **Aufklärung** statt **Hetze**...*
b. *#Muslimefürfrieden publicly reply to AfD brochure: **awareness** rather than **agitation***

(2) a. *Demo gegen **Abschiebung**: In Erfurt demonstrierten am 25. Januar etwa 200 Menschen gegen die **Abschiebungen** der R...*
b. *Demonstration against **deportation**: On January 25, 200 people demonstrated in Erfurt against the **deportations** of...*

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

unser_ ... <NOUN>
our_ ... <NOUN>

In this pattern, all morphological variants of the lexical cue are considered (*e. g.*, *unsere*, *unseren*), as indicated by the _ symbol. The following example illustrates this pattern:

(3) a. *RT @... Das war klar, es sind Muslime, sie wollen nur Teilhabe an unserem **Wohlstand** haben, ansonsten verachten sie uns...*
b. *RT @... Obviously, they are muslims, they only want to participate in our **wealth**, apart from that they scorn us...*

Both pro/con features and social identity features are primarily intended to capture aspects of

[1] `http://translate.google.com`

[2] In all examples throughout this paper, original German Tweets are presented in (a.), with our translation to English given in (b.), respectively.

national-chauvinism and social darwinism (*cf.* Section 3.1).

Transformation of Feature Values. After extracting the previously described features, the resulting feature vector describing each profile is transformed by following the tf·idf scheme (Manning et al., 2008). This is a standard approach in information retrieval to increase the relative impact of features that are (i) prominent in the respective profile and (ii) bear high discriminative power in the sense that they occur in a relatively small proportion of all profiles in the data.

3.3 Ranking Model

Our approach in this work can be seen as a generalization of nearest neighbour classification in a vector space framework (Manning et al., 2008): Twitter profiles are represented as points in a high-dimensional vector space using the features described in Section 3.2. Assuming a set of seed profiles that are labeled with one of the categories *right-wing extremist* (R) or *non-extremist* (N), the task is to rank an unseen profile $\vec{x}$ on a continuous scale spanning the range from right-wing extremist to non-extremist (N) content. Profiles are ranked according to their similarity to groups of nearest neighbours in the seed profiles.

We define centroids of non-extremist and right-wing nearest neighbours of $\vec{x}$, namely $C_N(\vec{x})$ and $C_R(\vec{x})$, respectively, as

$$C_N(\vec{x}) = \frac{1}{|N_k(\vec{x})|} \sum_{\vec{x}' \in N_k(\vec{x})} \vec{x}' \tag{1}$$

$$C_R(\vec{x}) = \frac{1}{|R_\ell(\vec{x})|} \sum_{\vec{x}' \in R_\ell(\vec{x})} \vec{x}' , \tag{2}$$

where $N_k(\vec{x})$ and $R_\ell(\vec{x})$ denote the sets of the k and ℓ nearest neighbours of $\vec{x}$ in the respective class in the training data. Then, the ranking score of the model is determined as the relative similarity of $\vec{x}$ to each centroid:

$$\mathrm{score}(\vec{x}) = \mathrm{sim}(\vec{x}, C_N(\vec{x})) - \mathrm{sim}(\vec{x}, C_R(\vec{x})) \tag{3}$$

With sim being instantiated as cosine similarity, this score ranges from -1 ($\vec{x}$ maximally similar to right-wing groups) to $+1$ ($\vec{x}$ maximally similar to non-extremist groups); borderline cases between both categories are expected to center around 0 (indicating equidistance of $\vec{x}$ to both groups). Setting k=1 and ℓ=1 renders the model an instance of nearest neighbour ranking.

4 Evaluation

4.1 Data Set

In our experiments, we use the data set previously discussed in Hartung et al. (2017). Annotations are provided by domain experts at the level of individual user profiles. These annotations comprise a set of 37 *seed profiles* of political actors from the 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. (2016). Non-extremist seed profiles contain political actors of the governing parties and single-issue associations (*e. g.*, nature conservation, social equality) (Quent et al., 2016).

In five other user profiles, the annotators were unable to reach a consensus on whether to classify the user as R or N. The latter profiles were kept in the data set as unlabeled *differential profiles*.

The test set comprises 100 randomly sampled profiles from followers of the seed users which have been annotated as being members of the R or N category.

4.2 Experiments and Results

4.2.1 Discrete Decoding

Given that ground truth annotations in the testing data are only available in terms of discrete labels (rather than continuous scores; cf. Section 4.1), the ranking model is evaluated in a discrete setting, using the following indicator function as a decision rule that is applied to the model score as given in Equation (3):

$$\mathrm{class}(\vec{x}) = \begin{cases} R, & \mathrm{score}(\vec{x}) < 0 \\ N, & \mathrm{score}(\vec{x}) > 0 \\ \mathrm{None}, & \mathrm{otherwise} \end{cases} \tag{4}$$

Note that discrete decoding can be applied in a *balanced* and *unbalanced* manner by setting the k and ℓ parameters in Equations (1) and (2) to the same or different numbers (thus considering nearest neighbour centroids of equal or different sizes).

Baseline Classifier. As a baseline classification model for comparison, we train a support vector machine (Cortes and Vapnik, 1995) with a linear kernel on the seed profiles (comprising 45,747 Tweets in total, among them 15,911 of category

	Entire sub-sample			Profiles >100 Tweets		
	P	R	F_1	P	R	F_1
discrete decoding unbalanced (k=4, ℓ=5)	0.56	0.79	0.65	0.79	0.85	0.81
discrete decoding balanced (k=10, ℓ=10)	0.55	0.65	0.59	0.80	0.62	0.70
discrete decoding balanced (k=1, ℓ=1)	0.44	0.63	0.52	0.69	0.69	0.69
Classification (Hartung et al., 2017)	0.25	0.95	0.40	0.32	0.92	0.47
Baseline	0.19	1.00	0.32	0.21	1.00	0.35

Table 1: Performance of the ranking model on the test set when being applied in a discrete decoding scenario, compared to a binary SVM classifier and a one-class baseline. Parameters k and ℓ in discrete decoding indicate the number of nearest neighbours in the centroids (cf. Equations 1 and 2).

R and 29,836 of category N) with all features described in Section 3.2. This implementation corresponds to our previous approach (Hartung et al., 2017).

Results. The results of this experiment can be seen in Table 1. We compare three variants of our ranking model in the previously described discrete decoding setting, the classification model and a baseline assigning all profiles to category R.

All models perform well above the baseline. While the classification model has a strong tendency towards recall, the ranking model generally offers a more harmonic precision-recall trade-off. Comparing the balanced and unbalanced model variants, we observe that our ranking approach generally benefits from larger centroids (thus preferring group similarities over individual ones), while the best performance can be obtained by choosing the k and ℓ parameters independently of one another (k=4, ℓ=5).

As can be seen from the right-most column of Table 1, reducing the test set to a subsample of profiles with at least 100 Tweets each (62 profiles remaining) leads to an additional performance increase up to an F_1 score of 0.81 in unbalanced discrete decoding.

All differences of the ranking models as reported in Table 1 are statistically significant over the baseline and the classifier according to an approximate randomization test (Yeh, 2000) at significance levels of $p < 0.05$ or smaller.

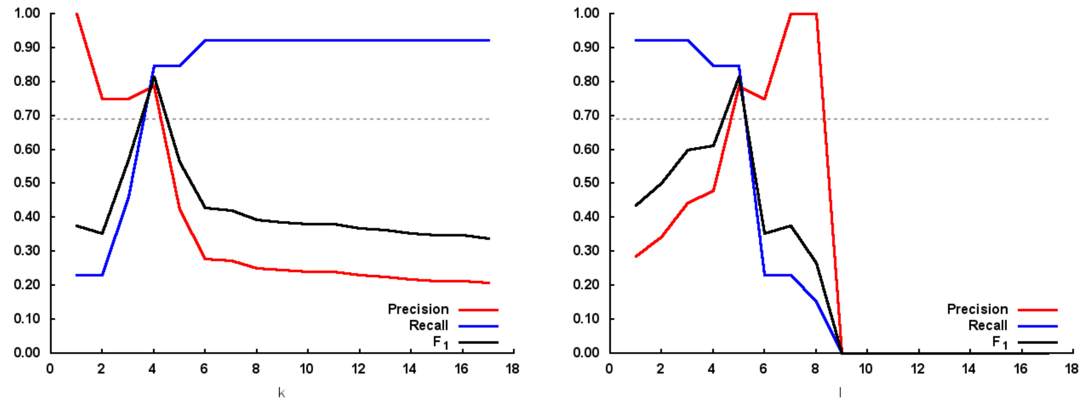
Discussion. In Figure 1 we explore the parameter space for different values of k and ℓ in unbalanced discrete decoding. While analyzing the variation in one parameter, the other one is fixed to its global optimum (k=4 and ℓ=5, respectively). For comparison, the dashed line indicates the performance of the nearest neighbour approach (*i. e.*, setting k=1 and ℓ=1) in terms of F_1 score.

As a general pattern, increasing the number of non-extremist neighbours in unbalanced discrete decoding fosters recall, while increasing the number of right-wing extremist neighbours fosters precision. Having said that, we also observe that the nearest neighbour approach generally yields robust performance which can be outperformed only in very few configurations throughout the parameter space. Apparently, in these configurations the model based on centroids of nearest neighbours is more effective in abstracting from outliers or borderline cases that might otherwise blur the decision boundary.

Figure 1 also illustrates that k and ℓ cannot be set to arbitrary large values without taking a considerable loss in performance. This indicates that, apart from abstracting from outliers, it is also crucial that the centroids are, to some degree, specific for the particular instance to be categorized, rather than a mere class prototype.

4.2.2 Continuous Ranking

In order to evaluate the plausibility of the ranking model scores in the absence of ground truth ranking annotations, we analyze the model predictions on the differential profiles for which no consensus regarding their category membership could be reached among the expert annotators (*cf.* Section 4.1). Being related to some New Right German political movements, which are notoriously hard to be delimited from right-wing extremist political



(a) Increasing k (number of non-extremist neighbours) for a fixed optimal value of ℓ=5

(b) Increasing ℓ (number of right-wing extremist neighbours) for a fixed optimal value of k=4

Figure 1: Exploration of the parameter space of k and ℓ on restricted test set (only profiles $>$100 Tweets). The dashed line indicates the performance of the nearest neighbour approach (*i. e.*, setting k=1 and ℓ=1)

actors, these cases are of particular interest from a social science perspective (*cf.* Zick et al., 2016). Due to their borderline character, we expect the ranking model to produce scores close to 0 for all these profiles.

Results. Figure 2 plots the profiles analyzed here on a continuous scale according to their predicted model score. We rely on the parameter settings which yielded best performance in the previous experiment (*i. e.*, k=4 and ℓ=5)[3]. As expected, all profiles are located closely around 0, which indicates that their predicted relative distance to extremist and non-extremist groups is almost equal. Despite the small sample size underlying this analysis, we consider this result as preliminary evidence of the plausibility of the ranking model on a selection of inherently difficult cases.

Discussion. Each data point in Figure 2 carries two types of information, viz. their position on the R–N spectrum according to the ranking model, and its category label as assigned by the baseline classifier. The latter is indicated in terms of crosses (denoting category N) and circles (category R). Comparing the predictions of both models, we find that they are in agreement in most of the cases. An interesting divergence concerns the case of a prominent member of a New Right German policitical party (explicitly marked by the arrow in Figure 2), who is categorized as R by the classifier, while being projected to the N range of the spectrum by the ranking model. We argue that this finding sheds light on the different methodological underpinnings of the models compared here: Apparently, this profile is sharing many properties with other non-extremist profiles, while the classifier still identifies a critical number of individual features which are taken as evidence in favour of an extremist profile. From our perspective, this finding reflects quite well the observed communicative strategies of the respective political party. Future work should be invested to corroborate this hypothesis.

4.3 Feature Analysis

Table 2 shows the impact of the individual feature groups as described in Section 3.2 in the ranking model when being used in isolation. In this analysis, Pro/Con features and Social Identity features are combined into one group (Pattern features).

We observe that all feature groups are effective to some degree: Emotion features tend to foster recall; pattern features may provide high precision, but suffer from low coverage due to their inherent sparsity. However, there is low complementarity between these feature groups, as the overall performance of the model (*cf.* Table 1) is clearly dominated by the lexical features.[4]

A preliminary analysis of the individual contributions of the emotion and pattern features according to their relative tf·idf weights per class shows that they are conceptually meaningful despite being superseded by other lexical features:

[3] However, the results reported in Figure 2 are largely stable with regard to the relative positions of the profiles to each other, despite some variation in the absolute values of the predicted model scores.

[4] A similar result has been found by Wester et al. (2016) for threat detection in social media.

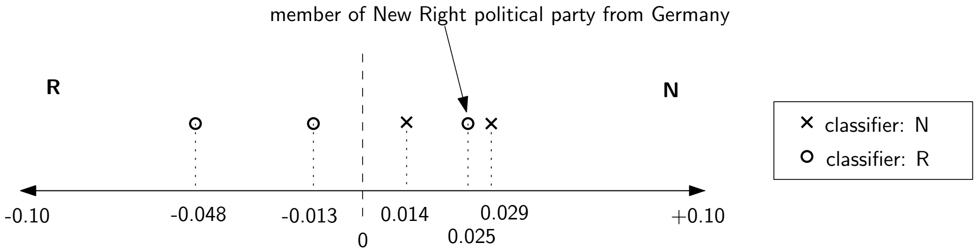


Figure 2: Continuous ranking of differential profiles (*cf.* Section 4.1). Position on the scale indicates the ranking score as given in Equation (3), based on optimal parameters k=4 and ℓ=5. The marked data point is assigned different categories by ranking and classification models (*cf.* discussion in Section 4.2.2).

	Lexical Features			Emotion Features			Pattern Features		
	P	R	F_1	P	R	F_1	P	R	F_1
discrete decoding unbalanced (k=4, ℓ=5)	0.79	0.85	0.81	0.38	0.62	0.47	1.00	0.08	0.14
discrete decoding balanced (k=10, ℓ=10)	0.80	0.62	0.70	0.20	0.38	0.26	0.00	0.00	0.00
discrete decoding balanced (k=1, ℓ=1)	0.69	0.69	0.69	0.48	0.85	0.61	0.63	0.38	0.48

Table 2: Results of analyzing the impact of individual feature groups in the ranking model when being used in isolation (on test set)

First, higher degrees of emotion in language use are clearly associated with category R profiles. Individual emotions most strongly associated with one of the categories are surprise, trust and disgust (for right-wing extremists), and love and sadness (for non-extremist users). Second, the most highly weighted pattern features for category R are GEGEN_Masseneinwanderung ('mass immigration'), UNSER_Politiker ('politicians'), UNSER_Fahne ('banner'), GEGEN_Syrien ('Syria') and GEGEN_Merkel, whereas UNSER_Land ('country'), GEGEN_Rechts ('Right-wing'), GEGEN_Gebietsreform ('territorial reform'), PRO_Aufklärung ('information') and UNSER_Jugendkandidat*innen ('youth contestants') are the most indicative patterns of category N.

5 Conclusions and Outlook

In this paper, we have presented a ranking model to identify Twitter profiles which display traits or attitudes of right-wing extremism. Our work is motivated by the goal of supporting human experts in their monitoring activities which are currently carried out purely manually.

Similarly to standard nearest-neighbour classification approaches, the model is based on estimating the relative proximity of an unseen profile to a limited number of manually annotated groups of seed profiles in high-dimensional vector space. We apply this model in the two settings of discrete decoding and continuous ranking. Our evaluation shows a significant advantage of the ranking model over a binary classification approach (Hartung et al., 2017). At the same time, the ranking model is found to deliver plausible predictions for a sample of borderline cases which specifically address actors from New Right political movements in Germany, whose categorization as right-wing extremists is currently debated in the social sciences (*cf.* Zick et al., 2016).

The latter finding clearly deserves a more thorough investigation based on a larger sample of cases, which we would like to address in future work. Additionally, we aim at developing this method further into a learning-to-rank approach in order to enable the comparison of profiles based on weighted properties. Finally, 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

ADL. 2016. Anti-Semitic Targeting of Journalists during the 2016 Presidential Campaign. A Report from ADL's Task Force on Harassment and Journalism. http://www.adl.org/assets/pdf/press-center/CR_4862_Journalism-Task-Force_v2.pdf.

Amadeu-Antonio-Stiftung. 2016. Rechtsextreme und menschenverachtende Phänomene im Social Web. https://www.amadeu-antonio-stiftung.de/w/files/pdfs/monitoringbericht-2015.pdf.

M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha. 2015. Detecting Jihadist Messages on Twitter. In *Proc. of EISIC*. https://doi.org/10.1109/EISIC.2015.27.

J.M. Berger. 2016. Nazis vs. ISIS on Twitter. A Comparative Study of White Nationalist and ISIS Online Social Media Networks. Technical report, Center for Cyber and Homeland Security, George Washington University, Washington, D.C.

H. Best, St. Niehoff, A. Salheiser, and L. Vogel. 2016. Gemischte Gefühle. Thüringen nach der "Flüchtlingskrise". Ergebnisse des Thüringen-Monitors. http://www.thueringen.de/mam/th1/tsk/thuringen-monitor_2016_mit_anhang.pdf.

G. Blanquart and D. Cook. 2013. Twitter Influence and Cumulative Perceptions of Extremist Support. A Case Study of Geert Wilders. In *Proc. of ACTC*.

A. Elizabeth C. Basave, Y. He, K. Liu, and J. Zhao. 2013. A Weakly Supervised Bayesian Model for Violence Detection in Social Media. In *Proceedings of the JCNLP*.

C. Cortes and V. Vapnik. 1995. Support-Vector Networks. *Machine Learning* 20:273–297.

P. Ekman. 1970. Universal Facial Expressions of Emotion. *California Mental Health Research Digest* 8(4):151–158.

E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan. 2016. Predicting Online Extremism, Content Adopters, and Interaction Reciprocity. *arxiv: https://arxiv.org/abs/1605.00659*.

D. Ghazi, D. Inkpen, and St. Szpakowicz. 2010. Hierarchical versus Flat Classification of Emotions in Text. In *NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*.

J. Golbeck, C. Robles, M. Edmondson, and K. Turner. 2011. Predicting Personality from Twitter. In *IEEE Int. Conference on Privacy, Security, Risk and Trust and IEEE Int. Conference on Social Computing*.

M. Hartung, R. Klinger, F. Schmidtke, and L. Vogel. 2017. Identifying Right-Wing Extremism in German Twitter Profiles: a Classification Approach. In F. Frascinar, A. Ittoo, L.M. Nguyen, and E. Métais, editors, *Natural Language Processing and Information Systems*, Springer, volume 10260 of *LNCS*, pages 320–325.

L. Kaati, E. Omer, N. Prucha, and A. Shrestha. 2015. Detecting Multipliers of Jihadism on Twitter. In *IEEE Int. Conference on Data Mining Workshop (ICDMW)*.

B. Liu. 2015. *Sentiment Analysis. Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Ch. D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

J.H. Parmelee and S.L. Bichars. 2013. *Politics and the Twitter Revolution*. Lexington Books, Landham, MD.

R. Plutchik. 2001. The Nature of Emotions. *American Scientist* .

M. Quent, A. Salheiser, and F. Schmidtke. 2016. Gefährdungen der demokratischen Kultur in Thüringen. http://www.denkbunt-thueringen.de/wp-content/uploads/2016/02/Gef%C3%A4hrdungsanalyse.pdf.

D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. 2011. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *IEEE Int. Conference on Privacy, Security, Risk and Trust and IEEE Int. Conference on Social Computing*.

J. R. Scanlon and M. S. Gerber. 2014. Automatic detection of cyber-recruitment by violent extremists. *Security Informatics* 3(1):5.

A. Schmidt and M. Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Valencia, Spain, pages 1–10.

R. Stöss. 2010. *Rechtsextremismus im Wandel*. Friedrich-Ebert-Stiftung, Berlin.

J. Suttles and N. Ide. 2013. *Distant Supervision for Emotion Classification with Discrete Binary Values*, Springer, Berlin, Heidelberg.

I-H. Ting, H.-M. Chi, J.-S. Wu, and S.-L. Wang. 2013. *An Approach for Hate Groups Detection in Facebook*, Springer Netherlands.

W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth. 2012. Harnessing Twitter "Big Data" for Automatic Emotion Identification. In *IEEE Int. Conference on Privacy, Security, Risk and Trust and IEEE Int. Conference on Social Computing*.

Z. Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas.

Z. Waseem and D. Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*.

Y. Wei and L. Singh. 2017. Using Network Flows to Identify Users Sharing Extremist Content on Social Media. In *Proceedings of PAKDD 2017*, Springer, volume 10234 of *LNAI*, pages 330–342.

Y. Wei, L. Singh, and S. Martin. 2016. Identification of Extremism on Twitter. In *Int. Conference on Advances in Social Networks Analysis and Mining*.

A. Wester, L. Øvrelid, E. Velldal, and H. L. Hammer. 2016. Threat Detection in Online Discussions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

A. Yeh. 2000. More Accurate Tests for the Statistical Significance of Result Differences. In *Proceedings of COLING*.

A. Zick, B. Küpper, D. Krause, R. Melzer, and W. Berghan, editors. 2016. *Gespaltene Mitte – Feindselige Zustände. Rechtsextreme Einstellungen in Deutschland 2016*. Dietz.