# „coding racism in social media data”:

# A Dictionary-based Approach to Racism Detection in Dutch Social Media: https://arxiv.org/abs/1608.08738 We present a dictionary-based approach to racism detection in Dutch social media comments, which were retrieved from two public Belgian social media sites likely to attract racist reactions. These comments were labeled as racist or non-racist by multiple annotators. For our approach, three discourse dictionaries were created: first, we created a dictionary by retrieving possibly racist and more neutral terms from the training data, and then augmenting these with more general words to remove some bias. A second dictionary was created through automatic expansion using a \texttt{word2vec} model trained on a large corpus of general Dutch text. Finally, a third dictionary was created by manually filtering out incorrect expansions. We trained multiple Support Vector Machines, using the distribution of words over the different categories in the dictionaries as features. The best-performing model used the manually cleaned dictionary and obtained an F-score of 0.46 for the racist class on a test set consisting of unseen Dutch comments, retrieved from the same sites used for the training set. The automated expansion of the dictionary only slightly boosted the model's performance, and this increase in performance was not statistically significant. The fact that the coverage of the expanded dictionaries did increase indicates that the words that were automatically added did occur in the corpus, but were not able to meaningfully impact performance. The dictionaries, code, and the procedure for requesting the corpus are available at:

# Us and them: identifying cyber hate on Twitter across multiple protected characteristics:

# <https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0072-6> Hateful and antagonistic content published and propagated via the World Wide Web has the potential to cause harm and suffering on an individual basis, and lead to social tension and disorder beyond cyber space. Despite new legislation aimed at prosecuting those who misuse new forms of communication to post threatening, harassing, or grossly offensive language - or cyber hate - and the fact large social media companies have committed to protecting their users from harm, it goes largely unpunished due to difficulties in policing online public spaces. To support the automatic detection of cyber hate online, specifically on Twitter, we build multiple individual models to classify cyber hate for a range of protected characteristics including race, disability and sexual orientation. We use text parsing to extract typed dependencies, which represent syntactic and grammatical relationships between words, and are shown to capture ‘othering’ language - consistently improving machine classification for different types of cyber hate beyond the use of a Bag of Words and known hateful terms. Furthermore, we build a data-driven blended model of cyber hate to improve classification where more than one protected characteristic may be attacked (e.g. race and sexual orientation), contributing to the nascent study of intersectionality in hate crime.

# Automated Hate Speech Detection and the Problem of Offensive Language:

# https://arxiv.org/abs/1703.04009

# A key challenge for automatic hate-speech detection on social media is the separation of hate speech from other instances of offensive language. Lexical detection methods tend to have low precision because they classify all messages containing particular terms as hate speech and previous work using supervised learning has failed to distinguish between the two categories. We used a crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords. We use crowd-sourcing to label a sample of these tweets into three categories: those containing hate speech, only offensive language, and those with neither. We train a multi-class classifier to distinguish between these different categories. Close analysis of the predictions and the errors shows when we can reliably separate hate speech from other offensive language and when this differentiation is more difficult. We find that racist and homophobic tweets are more likely to be classified as hate speech but that sexist tweets are generally classified as offensive. Tweets without explicit hate keywords are also more difficult to classify.

# machine learning antisemitism social media

# - detecting tweets against blacks: <http://www.aaai.org/ocs/index.php/AAAI/AAAI13/paper/download/6419/6821>

# - Detecting hate speech on the world wide web

# <https://dl.acm.org/citation.cfm?id=2390377>

# We present an approach to detecting *hate speech* in online text, where hate speech is defined as abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation. While hate speech against any group may exhibit some common characteristics, we have observed that hatred against each different group is typically characterized by the use of a small set of high frequency stereotypical words; however, such words may be used in either a positive or a negative sense, making our task similar to that of words sense disambiguation. In this paper we describe our definition of hate speech, the collection and annotation of our hate speech corpus, and a mechanism for detecting some commonly used methods of evading common "dirty word" filters. We describe pilot classification experiments in which we classify anti-semitic speech reaching an accuracy 94%, precision of 68% and recall at 60%, for an F1 measure of. 6375.

# Ranking right-wing extremist social media profiles by similarity to democratic and extremist groups

# https://pub.uni-bielefeld.de/publication/2912757

# Social media are used by an increasing number of political actors. A small subset of these is interested in pursuing extrem- ist 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 F1 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.

## **Alt\_Right white lite: Trolling, hate speech and cyber racism on social media**

# <https://search.informit.com.au/documentSummary;dn=309784538174296;res=IELHSS>

# **Abstract:**The rapid growth of race hate speech on the Internet seems to have overwhelmed the capacity of states, corporations or civil society to limit its spread and impact. Yet by understanding how the political economy of the Internet facilitates racism it is possible to chart strategies that might push back on its negative social effects. Only by involving the state, economy and civil society at both the global level, and locally, can such a process begin to develop an effective 'civilising' dynamic. However neo-liberalism and democratic license may find such an exercise ultimately overwhelmingly challenging, especially if the fundamental logical drivers that underpin the business model of the Internet cannot be transformed. This article charts the most recent rise and confusion of the Internet under the impact of the Alt\_Right and other racist groups, focusing on an Australian example that demonstrates the way in which a group could manipulate the contradictions of the Internet with some success. Using an analytical model developed to understand the political economy and sociology of mass media power in the later stages of modernity, before the Internet, the author offers a series of proposals on how to address racism on the Internet.

# Methods for monitoring and mapping online hate speech

# <http://www.gsdrc.org/docs/open/hdq1121.pdf>

# online extremism machine learning

# Interpreting text and image relations in violent extremist discourse: A mixed methods approach for big data analytics

# http://www.tandfonline.com/doi/abs/10.1080/09546553.2016.1233871

# This article presents a mixed methods approach for analysing text and image relations in violent extremist discourse. The approach involves integrating multimodal discourse analysis with data mining and information visualisation, resulting in theoretically informed empirical techniques for automated analysis of text and image relations in large datasets. The approach is illustrated by a study which aims to analyse how violent extremist groups use language and images to legitimise their views, incite violence, and influence recruits in online propaganda materials, and how the images from these materials are re-used in different media platforms in ways that support and resist violent extremism. The approach developed in this article contributes to what promises to be one of the key areas of research in the coming decades: namely the interdisciplinary study of big (digital) datasets of human discourse, and the implications of this for terrorism analysis and research.

# Predicting Online Extremism, Content Adopters, and Interaction Reciprocity

# <https://link.springer.com/chapter/10.1007/978-3-319-47874-6_3>

# We present a machine learning framework that leverages a mixture of metadata, network, and temporal features to detect extremist users, and predict content adopters and interaction reciprocity in social media. We exploit a unique dataset containing millions of tweets generated by more than 25 thousand users who have been manually identified, reported, and suspended by Twitter due to their involvement with extremist campaigns. We also leverage millions of tweets generated by a random sample of 25 thousand regular users who were exposed to, or consumed, extremist content. We carry out three forecasting tasks, (i) to detect extremist users, (ii) to estimate whether regular users will adopt extremist content, and finally (iii) to predict whether users will reciprocate contacts initiated by extremists. All forecasting tasks are set up in two scenarios: a post hoc (time independent) prediction task on aggregated data, and a simulated real-time prediction task. The performance of our framework is extremely promising, yielding in the different forecasting scenarios up to 93 % AUC for extremist user detection, up to 80 % AUC for content adoption prediction, and finally up to 72 % AUC for interaction reciprocity forecasting. We conclude by providing a thorough feature analysis that helps determine which are the emerging signals that provide predictive power in different scenarios.

### **Detecting online markers of violent extremism in online enviroments (chapter 18 of** [**Interpreting text and image relations in violent**extremist **discourse: A mixed methods approach for big data analytics**](http://www.tandfonline.com/doi/abs/10.1080/09546553.2016.1233871)**)**

### <https://books.google.de/books?hl=de&lr=&id=qu0ODAAAQBAJ&oi=fnd&pg=PA374&dq=online+extremism+machine+learning&ots=E14RbQu2jS&sig=E21tmdMUGOytFGZ7vZC3pEDgYD4#v=onepage&q=online%20extremism%20machine%20learning&f=false>

# On Textual Analysis and Machine Learning for Cyberstalking Detection

# <https://link.springer.com/article/10.1007/s13222-016-0221-x>

# Cyber security has become a major concern for users and businesses alike. Cyberstalking and harassment have been identified as a growing anti-social problem. Besides detecting cyberstalking and harassment, there is the need to gather digital evidence, often by the victim. To this end, we provide an overview of and discuss relevant technological means, in particular coming from text analytics as well as machine learning, that are capable to address the above challenges. We present a framework for the detection of text-based cyberstalking and the role and challenges of some core techniques such as author identification, text classification and personalisation. We then discuss PAN, a network and evaluation initiative that focusses on digital text forensics, in particular author identification.

# Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter

# <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181405>

# The Islamic State of Iraq and ash-Sham (ISIS) continues to use social media as an essential element of its campaign to motivate support. On Twitter, ISIS’ unique ability to leverage unaffiliated sympathizers that simply retweet propaganda has been identified as a primary mechanism in their success in motivating both recruitment and “lone wolf” attacks. The present work explores a large community of Twitter users whose activity supports ISIS propaganda diffusion in varying degrees. Within this ISIS supporting community, we observe a diverse range of actor types, including fighters, propagandists, recruiters, religious scholars, and unaffiliated sympathizers. The interaction between these users offers unique insight into the people and narratives critical to ISIS’ sustainment. In their entirety, we refer to this diverse set of users as an online extremist community or OEC. We present Iterative Vertex Clustering and Classification (IVCC), a scalable analytic approach for OEC detection in annotated heterogeneous networks, and provide an illustrative case study of an online community of over 22,000 Twitter users whose online behavior directly advocates support for ISIS or contibutes to the group’s propaganda dissemination through retweets.

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

# https://link.springer.com/chapter/10.1007/978-3-319-59569-6\_40

# 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.

# Machine Learning Models that Remember Too Much

# https://dl.acm.org/citation.cfm?id=3134077

Machine learning (ML) is becoming a commodity. Numerous ML frameworks and services are available to data holders who are not ML experts but want to train predictive models on their data. It is important that ML models trained on sensitive inputs (e.g., personal images or documents) not leak too much information about the training data.

We consider a malicious ML provider who supplies model-training code to the data holder, does \emph{not} observe the training, but then obtains white- or black-box access to the resulting model. In this setting, we design and implement practical algorithms, some of them very similar to standard ML techniques such as regularization and data augmentation, that "memorize" information about the training dataset in the model\textemdash yet the model is as accurate and predictive as a conventionally trained model. We then explain how the adversary can extract memorized information from the model. We evaluate our techniques on standard ML tasks for image classification (CIFAR10), face recognition (LFW and FaceScrub), and text analysis (20 Newsgroups and IMDB). In all cases, we show how our algorithms create models that have high predictive power yet allow accurate extraction of subsets of their training data

# Automatic Detection of Hateful Comments in Online Discussion

# <https://link.springer.com/chapter/10.1007/978-3-319-52569-3_15>

# Making violent threats towards minorities like immigrants or homosexuals is increasingly common on the Internet. We present a method to automatically detect threats of violence using machine learning. A material of 24,840 sentences from YouTube was manually annotated as violent threats or not, and was used to train and test the machine learning model. Detecting threats of violence works quit well with an error of classifying a violent sentence as not violent of about 10% when the error of classifying a non-violent sentence as violent is adjusted to 5%. The best classification performance is achieved by including features that combine specially chosen important words and the distance between those in the sentence.

# Understanding Online Radicalisation Using Data Science

# https://www.igi-global.com/article/understanding-online-radicalisation-using-data-science/171450

## Abstract

What characterises social media radicals? And why some people become attracted to radicalisation? To explore answers to these questions, a number of tweets posted by a group of suspected radicals tweeting in Arabic were analysed using social network analysis and machine learning. The study revealed that these suspected radicals' networks showed significant interaction with others; but this interactivity is only significant quantitatively as the interaction is not reciprocated. With regards to why these suspected radicals became attracted to radicalisation, Topic Modelling revealed these suspected radicals' tweets underpinned a perceived injustice that they believed the Secret Police and the Government inflicted upon them. Overall, the study has shown that data science tools have the potential to inform our understanding of online radicalisation. It is hoped this exploratory study will be the basis for a future study in which the research questions will be answered using a larger sample.

# - Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts)

<http://ieeexplore.ieee.org/abstract/document/8004907/>

In this paper we present results of a research on automatic extremist text detection. For this purpose an experimental dataset in the Russian language was created. According to the Russian legislation we cannot make it publicly available. We compared various classification methods (multinomial naive Bayes, logistic regression, linear SVM, random forest, and gradient boosting) and evaluated the contribution of differentiating features (lexical, semantic and psycholinguistic) to classification quality. The results of experiments show that psycholinguistic and semantic features are promising for extremist text detection.

* **Machine Learning for Affect Analysis on White Supremacy Forum**

<http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A955841&dswid=-6324>

Since the inception of the World Wide Web, security agencies, researchers, and analysts have focused much of their attention on the sentiment found on hate-inspired webforums. Here, one of their goals has been to detect and measure users' affects that are expressed in the forums as well as identify how users' affects change over time. Manual inspection has been one way to do this; however, as the number of discussion posts and sub-forums increase, there has been a growing need for an automated system that can assist humans in their analysis. The aim of this thesis, then, is to detect and measure a number of affects expressed in written text on Stormfront.org, the most visited hate forum on the Web. To do this, we used a machine learning approach where we trained a model to recognize affects on three sub-forums: Ideology and Philosophy, For Stormfront Ladies Only, and Stormfront Ireland. The training data consisted of manually annotated data and the affects we focused on were racism, aggression, and worries. Results indicate that even though measuring affects is a subjective process, machine learning is a promising way forward to  analyse and measure the presence of different affects on hate forums.

* [**Detection and Analysis of Online Extremist Communities**](http://repository.cmu.edu/cgi/viewcontent.cgi?article=1988&context=dissertations)

<http://repository.cmu.edu/dissertations/949/>

Online social networks have become a powerful venue for political activism. In many cases large, insular online communities form that have been shown to be powerful diffusion mechanisms of both misinformation and propaganda. In some cases these groups users advocate actions or policies that could be construed as extreme along nearly any distribution of opinion, and are thus called Online Extremist Communities (OECs). Although these communities appear increasingly common, little is known about how these groups form or the methods used to influence them. The work in this thesis provides researchers a methodological framework to study these groups by answering three critical research questions: How can we detect large dynamic online activist or extremist communities? What automated tools are used to build, isolate, and influence these communities? What methods can be used to gain novel insight into large online activist or extremist communities? These group members social ties can be inferred based on the various affordances offered by OSNs for group curation. By developing heterogeneous, annotated graph representations of user behavior I can efficiently extract online activist discussion cores using an ensemble of unsupervised machine learning methods. I call this technique Ensemble Agreement Clustering. Through manual inspection, these discussion cores can then often be used as training data to detect the larger community. I present a novel supervised learning algorithm called Multiplex Vertex Classification for network bipartition on heterogeneous, annotated graphs. This methodological pipeline has also proven useful for social botnet detection, and a study of large, complex social botnets used for propaganda dissemination is provided as well. Throughout this thesis I provide Twitter case studies including communities focused on the Islamic State of Iraq and al-Sham (ISIS), the ongoing Syrian Revolution, the Euromaidan Movement in Ukraine, as well as the alt-Right.

social media machine learning/social media hate machine learning/online racism machine learning

# Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying

# <https://academic.oup.com/jigpal/article-abstract/24/1/42/2893010>

# The use of new technologies along with the popularity of social networks has given the power of anonymity to the users. The ability to create an alter-ego with no relation to the actual user, creates a situation in which no one can certify the match between a profile and a real person. This problem generates situations, repeated daily, in which users with fake accounts, or at least not related to their real identity, publish news, reviews or multimedia material trying to discredit or attack other people who may or may not be aware of the attack. These acts can have great impact on the affected victims' environment generating situations in which virtual attacks escalate into fatal consequences in real life. In this article, we present a methodology to detect and associate fake profiles on Twitter social network which are employed for defamatory activities to a real profile within the same network by analysing the content of the comments generated by both profiles. Accompanying this approach we also present a successful real life use case in which this methodology was applied to detect and stop a cyberbullying situation in a real elementary school.