





An Identity-Based Framework for Generalizable Hate Speech Detection

Joshua Uyheng^(✉)  and Kathleen M. Carley 

CASOS Center, Institute for Software Research, Carnegie Mellon University,
Pittsburgh, PA 15213, USA
{juyheng,kathleen.carley}@cs.cmu.edu

Abstract. This paper explores the viability of leveraging an identity-based framework for generalizable hate speech detection. Across a corpus of seven benchmark datasets, we find that hate speech consistently features higher levels of abusive and identity terms, robust to social media platforms of origin and multiple languages. Using only lexical counts of abusives, identities, and other psycholinguistic features, heuristic and machine learning models achieve high precision and weighted F1 scores in hate speech prediction, with performance on a three-language dataset comparable to recent state-of-the-art multilingual models. Cross-dataset predictions further reveal that our proposed identity-based models map hate and non-hate categories with each other in a conceptually coherent fashion across diverse classification schemes. Our findings suggest that conceptualizing hate speech through an identity lens offers a generalizable, interpretable, and socio-theoretically robust framework for computational modelling of online conflict and toxicity.

Keywords: Hate speech · Social media · Identities · Machine learning

1 Introduction

Developing computational methods for the detection of hate speech is an important task for the emerging science of social cyber-security [2, 14]. While a vast literature in machine learning and allied fields has sought to develop cutting-edge tools to address this problem [4, 6, 16], some consequences of this proliferation of work include a divergence in hate speech definitions, and with the increasing traction of complex deep learning models, the interpretability of model predictions [15].

This work was supported in part by the Knight Foundation and the Office of Naval Research grants N000141812106 and N000141812108. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS) and the Center for Informed Democracy and Social Cybersecurity (IDEaS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research or the U.S. government.

Table 1. Summary of datasets used. Collapsed classes are specified in parentheses. Frequencies reflect retrievable data, as some data points may be unavailable.

Dataset	Class	Frequency
Chung [3]	Hate	857 (11.19%)
	Counter	6804 (88.81%)
Davidson [4]	Hate	1430 (5.77%)
	Offensive	19190 (77.43%)
	Regular	4163 (16.80%)
De Gibert [5]	Hate	1196 (10.93%)
	Regular (Relation/Non-Hate)	8748 (89.07%)
Founta [6]	Hate	541 (7.28%)
	Abusive	1956 (26.30%)
	Spam	848 (11.40%)
	Regular	4089 (54.99%)
Mathew [9]	Hate	6854 (34.02%)
	Offensive	5480 (27.20%)
	Regular	7814 (38.78%)
Qian [10]	Hate	25344 (40.80%)
	Regular	36779 (59.20%)
Waseem [16]	Hate (Racism/Sexism)	2161 (27.41%)
	Regular	5723 (72.59%)

In this work, we use social scientific theory about *identities* to build grounded computational models of hate speech. Utilizing a multilingual lexicon of known terms which reference identities [7], we show that it is possible to detect hate speech in an accurate, interpretable, and generalizable way across a variety of datasets based on diverse definitional taxonomies, languages, and social media platforms [3–6, 9, 10, 16]. Through this work, we contribute an enhanced social scientific understanding of hate speech that may inform future modelling efforts, as well as a general and scalable method in its own right that may be deployed for hate speech detection in emergent and applied settings [2, 12, 13].

2 Data and Methods

2.1 Dataset Curation

To facilitate systematic analysis, this study relied on a curated collection of hate speech datasets for systematic analysis. Beginning with the online repository¹ of Vidgen and Derczynski [15], we filtered out datasets with fewer than 1000 examples, which included only hate (and no negative examples), or which conflated

¹ <https://hatespeechdata.net>.

hate with bullying (outside the current scope). We settled on seven datasets summarized in Table 1.

Because several datasets involved social media data (e.g., Twitter), Table 1 reports the frequency of each class label as retrievable from each dataset as of February 2021. Due to possible suspensions of hateful utterances online, these reported frequencies may differ from statistics reported in their original papers. Additionally, for classes which had few examples (e.g., less than 100), we collapsed several classes into a single class, made explicit in Table 1. For each dataset, we uniformly identify the “hate” category. We note that non-hate categories across datasets may vary widely, including regular speech [4–6, 9, 10, 16], abusive or offensive but not hateful speech [4, 6, 9], spam [6], and counter-hate [3]. While most datasets originate from Twitter, others also include utterances generated offline [3], and from other platforms like Reddit [10], Gab [9, 10], and Stormfront [5]. One dataset moreover includes multilingual data [3].

2.2 An Identity Lexicon with Psycholinguistic Features

To study the use of identities in the collected hate speech datasets, we drew upon social scientific theorizing around identities, or concepts of socially embedded selves and groups. We leverage recent work which expands and validates a lexicon of *identity terms* for computational modelling [7]. We specifically use the Netmapper software² which counts these identity terms across several dozen languages, and has been previously used in applied settings of psycholinguistic analysis for social cyber-security [12, 13].

We additionally use Netmapper to measure various other psycholinguistic features that have been associated with a wide variety of cognitive and emotional states [11]. These include pronoun usage, various positive and negative emotion words, and patterns of punctuation. Of particular interest, however, we also probe the joint presence of identity terms alongside known *abusive terms*. Taken together, we propose that identities and abusives constitute general, reliable, and theoretically grounded empirical touchstones for hate speech detection.

2.3 Problem Formulation and Experimental Setup

Utilizing the foregoing psycholinguistic measurements across the seven datasets in this study, we examine several research questions of interest. We divide the results that follow into three stages, beginning with a statistical analysis of identities and abusives across the curated datasets (RQ1), a predictive modelling analysis that evaluates the use of psycholinguistic features for hate speech detection within individual datasets (RQ2), and a generalizability analysis that maps out the quality and consistency of cross-dataset predictions (RQ3).

² <https://netanomics.com/netmapper/>.

RQ1: Does hate speech contain more abusive and identity terms?

To answer our first question, we perform two types of regression analysis. We begin with separate analyses of abusives and identities for each dataset. Here, we perform linear regression to predict the number of abusives and identities in a given text based solely on its label. Based on extant definitions of hate speech, we expect that, across datasets, hate speech will indeed have higher levels of *both* abusive and identity terms.

Next, we perform a binary logistic regression analysis over a consolidated dataset, where we predict whether a given text is classified as hate or not hate. Here, the predictors are now the number of abusive and identity terms in a given text. In addition, we control for each dataset’s platform of origin and its multilingual focus by including them as covariates in the regression model. In this case, we also hypothesize positive effects for both abusives and identities, robust to the explored controls.

RQ2: How can abusive and identity terms be used to detect hate speech?

In the second stage of this work, we shift from statistical analysis to predictive modelling. Here, our objective is now to evaluate whether hate speech can be accurately detected using abusives and identities. We begin by evaluating the precision of a heuristic model that simplistically predicts that a text is hate speech if it contains at least one abusive term, and at least one identity term. Given only two features, this heuristic may not achieve particularly high precision. However, we do expect it to significantly outperform a random baseline, as it aligns with our identity-based theoretical understanding of hate speech.

We then assess the use of psycholinguistic counts more broadly - which include abusives and identities - for hate speech detection using a range of machine learning models, including logistic regression, random forests, and support vector machines. We augment feature inputs with their squared values and pairwise products to capture interactions between psycholinguistic measures, since these second-order terms capture signals from when pairs of variables have simultaneously high values.

We compare these results against a deep learning benchmark that uses the full text. We specifically use a classical convolutional neural network (CNN) for sentence classification [8]. We use an Adam optimizer and perform a grid search over word embedding dimension, filter size, and dropout rate. We use the average weighted F1 measure obtained from five-fold cross-validation for evaluation. Here, we expect that the deep learning model will consistently achieve the best performance, but we also expect that the psycholinguistically informed machine learning models will not be severely worse.

RQ3: How generalizable is identity-based hate speech detection?

Finally, we consider the generalizability of the proposed identity-based framework for hate speech detection. Here, we adopt both confirmatory and exploratory tools to flesh out a rich examination of cross-dataset dynamics.

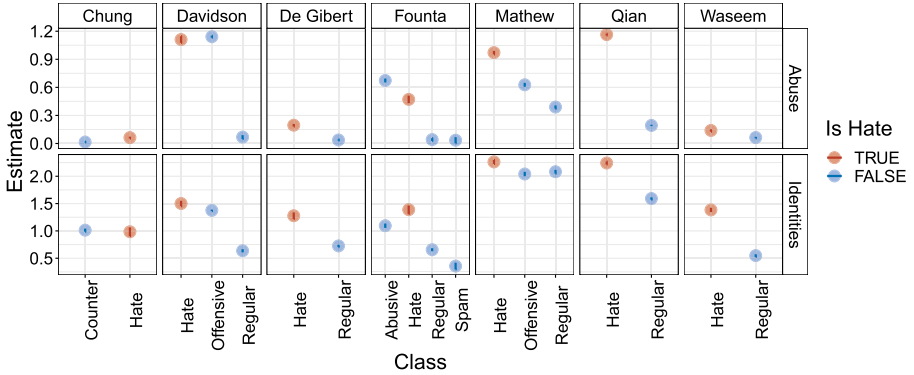


Fig. 1. Results of linear regression on single datasets, predicting abusives and identities based on class labels. In general, hate speech has higher (red) values of *both* abuse and identities than non-hate (blue). (Color figure online)

We first take the best-performing machine learning model from the previous stage of analysis, and use it to generate predictions for all other datasets in our corpus. For instance, a model trained using the Davidson dataset [4] would generate predictions for the Founta dataset [6]. So all instances in the Founta dataset of Hate, Abusive, Spam, and Regular, will be assigned a Davidson-based label of Hate, Offensive and Regular. Given the diversity of definitional taxonomies, we assess the extent to which each model accurately maps its own associated hate class to the hate class of other datasets.

From an exploratory standpoint, we also perform principal component analysis on the cross-dataset predictions. This allows us to qualitatively assess a shared, low-dimensional visualization of text classes across datasets. In both analyses of generalizability, we expect that if our identity-based psycholinguistic approach is successful, hate categories will be mapped to each other across datasets, and non-hate categories likewise.

3 Results

3.1 Hate Speech Consistently Features Identity Abuse

Figure 1 visualizes the levels of abusive and identity-based language in hate speech (red) and other classes (blue) in each dataset. In the case of the De Gibert [5], Mathew [9], Qian [10], and Waseem [16] datasets, comparisons are straightforward, as hate speech shows higher levels of both measures. Interestingly, however, we note that in both Davidson [4] and Founta [6] datasets, the Offensive and Abusive classes respectively have more abusive terms. Yet in both cases, hate speech features more identity terms than these classes, while *also* containing more abusives than the designated Regular classes. Conversely, in the Chung [3] dataset, hate speech shows fewer identity terms than the Counter

Table 2. Results of binary logistic regression analysis over consolidated dataset ($N = 140979$). Regression 1 is the baseline, while Regression 2 includes covariates. Positive and statistical significant coefficients for Abusives and Identities point to their robust associations with hate across platforms and languages.

Factors	Regression 1		Regression 2	
	Coefficient	p	Coefficient	p
Intercept	-1.7436***	<.001	–	–
Abusives	0.8895***	<.001	1.0909***	<.001
Identities	0.1196***	<.001	0.0156***	<.001
Gab	–	–	1.7283***	<.001
Reddit	–	–	-2.7653***	<.001
Stormfront	–	–	-2.2021***	<.001
Twitter	–	–	-3.1943***	<.001
Multilingual	–	–	-2.1225***	<.001

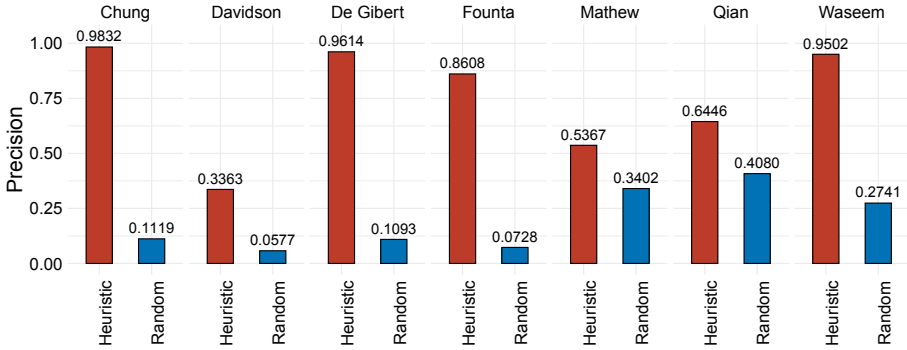


Fig. 2. Evaluation of heuristic predictive model precision against random baselines.

class; but as expected, hate speech has more abusive terms. Thus, our hypothesis for this analysis holds: hate speech systematically features *both* abusive and identity terms, at significantly higher levels than other kinds of text.

We strengthen this per-dataset observation by analyzing the consolidated corpus ($N = 140979$). Table 2 shows that regardless of dataset, indeed, higher levels of abusive (Model 1: $\beta = 0.8895, p < .001$; Model 2: $\beta = 0.1196, p < .001$) and identity (Model 1: $\beta = 1.0909, p < .001$, Model 2: $\beta = 0.0156, p < .001$) terms predict higher likelihood of a text being hate speech, with and without controlling for platforms of origin and multilingualism suggest the robustness of these effects.

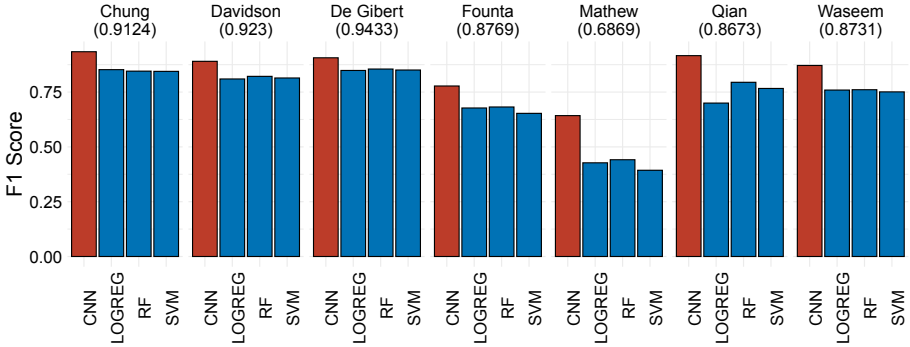


Fig. 3. Cross-validated weighted F1 scores of machine learning models versus CNNs. Datasets are labelled with the best relative performance of machine learning models.

3.2 Evaluating Identity Abuse for Hate Speech Detection

Predictive modelling results bolster our statistical findings by showing the practical utility of abusive and identity terms in detecting hate speech. Figure 2 shows that a heuristic that predicts hate speech solely using these two features achieves 86.08–98.32% precision for four of the seven datasets, with the remaining three datasets showing 33.63–64.46% precision. Note we only evaluate precision because the two features alone do not account for all non-hate classes.

These results indicate that for many datasets of hate speech, our proposed theoretical identity-based framework may have predictive utility. However, more features need to be accounted for in other hate taxonomies, in view of the latter three datasets. Yet despite their relatively low performance, we note with interest that across all datasets, the heuristic vastly outperformed all random baselines. Two-sample proportion tests with continuity correction affirm these differences to be highly statistically significant ($p < .001$).

Pushing our psycholinguistic framework further, we now evaluate various machine learning models against a deep learning benchmark [8]. On average, the F1 scores of psycholinguistic machine learning models are 12.11% higher than the precision scores of the two-feature heuristics using only identities and abusives. Yet as expected, a CNN with high-dimensional representations of the entire text achieves better performance compared to machine learning models only given psycholinguistic counts as inputs.

However, in all cases but one [9], the best machine learning model performance is 86.73–94.33% that of the CNN model. That means using our approach there is practically between a 5.67–13.27% performance reduction from a deep learning model, even with significantly fewer parameters, and with much greater interpretability. We also note that in the multilingual dataset [3], the F1 scores of the machine learning models (0.8445–0.8523) were comparable to a state-of-the-art analysis of deep learning models in a multilingual setup (0.6651–0.8365) [1]. This may be crucial to systematically explore further.

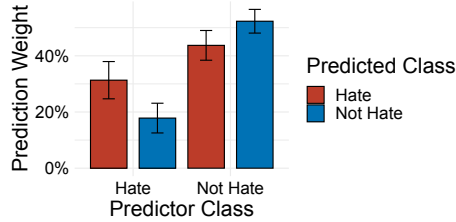


Fig. 4. Cross-dataset predictions using the best psycholinguistic machine learning models. Error bars show 95% confidence intervals.

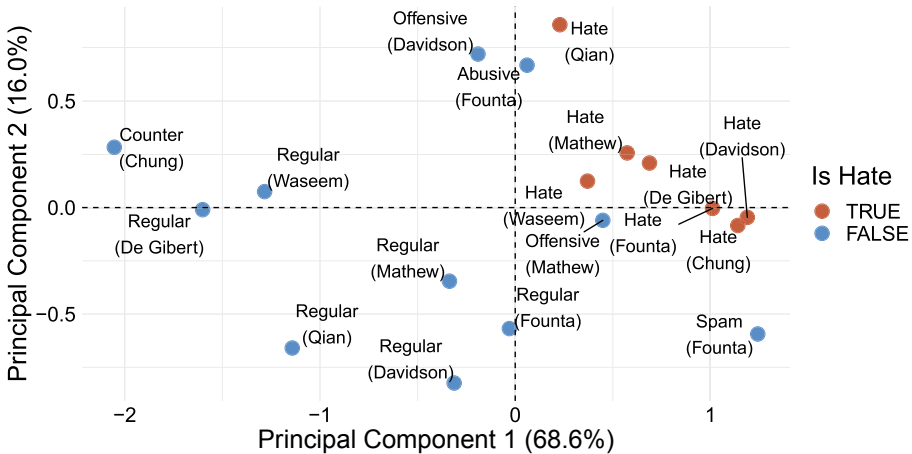


Fig. 5. Principal component analysis of cross-dataset predictions, with hate in red, and non-hate in blue. Principal components are presented with variance explained. (Color figure online)

For the outlier case [9], though machine learning models at best achieve 68.69% of the CNN performance, the CNN itself only obtains a weighted F1 score of 64.27%. As the dataset was used to train explainable models with human-generated context, these examples may be difficult to classify without this additional information, as here we only leverage the raw text.

3.3 Mapping Cross-dataset Generalizability

Finally, we turn to generalizability analysis. Figure 4 shows the accuracy of cross-dataset predictions for all hate and non-hate classes. Across datasets, we find that our proposed identity-based framework reliably maps hate to hate, and non-hate to non-hate. However, non-negligible discrepancies may still arise, likely due to intermediate categories like Offensive, Abusive, and Spam classes.

To explore these nuances further, Fig. 5 shows the results of a principal component analysis on cross-dataset prediction weights. Remarkably, we find that

the Hate classes for all seven datasets tightly cluster on the top-right corner of a low-dimensional visualization of the first two principal components (84.6% of the variance). As might be expected, the other classes plotted closest to hate speech are the harmful yet non-hate classes of concern. Notably, Spam is distinctly located in a separate location from Offensive and Abusive, which are conceptually more similar to each other. Furthermore, most Regular classes are clustered near the left and bottom-left corner of Fig. 5, with the Counter class occupying the left-most coordinate. Conceptually, the Counter class, which actively combats hate [3], may also be understood as the most distinct form of speech relative to both hate and other irregular yet non-hateful texts.

Collectively, then, we find a highly theoretically coherent mapping of the classes considered in our corpus, with cross-dataset model predictions empirically resonating with latent conceptual relationships. This suggests that our proposed identity-based framework may valuably capture shared, underlying features of hate speech and related constructs that cut across datasets.

4 Conclusions and Future Work

This paper showed that an identity-based approach reflects key features of hate speech across several datasets [7]. Hate speech not only features more identity and abusive terms (RQ1), but these terms may also detect hate speech with enhanced interpretability relative to state-of-the-art models (RQ2) [1]. Identity-based representations further produce generalizable models that map different forms of hate and non-hate in a theoretically coherent fashion (RQ3).

This work shows the conceptual benefit of an identity perspective for understanding hate as abuse targeted against social groups, with practical benefits for generalizable, interpretable, and scalable computational modelling [2, 14]. Future applied work may leverage the multi-dataset effort presented here for even richer maps of hate dynamics - alongside linked phenomena like spam and counter-hate [3, 6] - to capture the multi-faceted nature of online toxicity, and potentially inform more nuanced policy and platform responses [14]. Our theory-based method could also be potentially used to classify the targets of hate and measure the coordination of hate in information operations [2, 12, 13].

To this end, extensions to our work may readily be pursued through an expanded data corpus and set of models. Psycholinguistic and identity-based features could also be utilized alongside - rather than independent from - word or sentence embeddings already prevalent in cutting-edge hate speech detection applications [15]. It is also extremely promising to more systematically compare - or combine - our approach against more state-of-the-art models in multilingual hate speech prediction [1]. Finally, from a qualitative standpoint, it is crucial to consider kinds of hate that may not explicitly abuse identities, yet take on silent but salient and harmful forms - these are issues outside the framework we propose, and form vital directions for further, multidisciplinary research [2, 14].

References

1. Aluru, S.S., Mathew, B., Saha, P., Mukherjee, A.: Deep learning models for multilingual hate speech detection. arXiv preprint [arXiv:2004.06465](https://arxiv.org/abs/2004.06465) (2020)
2. Carley, K.M.: Social cybersecurity: an emerging science. *Comput. Math. Org. Theory* **26**(4), 365–381 (2020)
3. Chung, Y.L., Kuzmenko, E., Tekiroglu, S.S., Guerini, M.: CONAN-COunter NAratives through Nichesourcing: a multilingual dataset of responses to fight online hate speech. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2819–2829 (2019)
4. Davidson, T., Warmley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11 (2017)
5. De Gibert, O., Perez, N., García-Pablos, A., Cuadros, M.: Hate speech dataset from a white supremacy forum. In: *Proceedings of the 2nd Workshop on Abusive Language Online*, pp. 11–20 (2018)
6. Founta, A., et al.: Large scale crowdsourcing and characterization of Twitter abusive behavior. In: *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12 (2018)
7. Joseph, K., Wei, W., Benigni, M., Carley, K.M.: A social-event based approach to sentiment analysis of identities and behaviors in text. *J. Math. Sociol.* **40**(3), 137–166 (2016)
8. Kim, Y.: Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, October 2014
9. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: HateXplain: a benchmark dataset for explainable hate speech detection. arXiv preprint [arXiv:2012.10289](https://arxiv.org/abs/2012.10289) (2020)
10. Qian, J., Bethke, A., Liu, Y., Belding, E., Wang, W.Y.: A benchmark dataset for learning to intervene in online hate speech. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 4757–4766 (2019)
11. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* **29**(1), 24–54 (2010)
12. Uyheng, J., Carley, K.M.: Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines. *J. Comput. Soc. Sci.* **3**(2), 445–468 (2020)
13. Uyheng, J., Carley, K.M.: Characterizing network dynamics of online hate communities around the COVID-19 pandemic. *Appl. Network Sci.* **6**(1), 1–21 (2021). <https://doi.org/10.1007/s41109-021-00362-x>
14. Uyheng, J., Magelinski, T., Villa-Cox, R., Sowa, C., Carley, K.M.: Interoperable pipelines for social cyber-security: assessing Twitter information operations during NATO Trident Juncture 2018. *Comput. Math. Organ. Theory* **26**, 1–19 (2019)
15. Vidgen, B., Derczynski, L.: Directions in abusive language training data, a systematic review: garbage in, garbage out. *PLoS ONE* **15**(12), e0243300 (2020)
16. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In: *Proceedings of the NAACL Student Research Workshop*, pp. 88–93 (2016)