Aggression in online comments

Text recognition for the SNF Digital Lives research project «Online Aggression From a Sociological Perspective»

sotomo
gesellschaft, politik & raum

Commissioned by

Dr. Lea Stahel
Dr. Sebastian Weingartner
Institute of Sociology
University of Zurich
Andreasstrasse 15
8050 Zurich

Contractor

sotomo Research Institute
Dolderstrasse 24
8032 Zurich

Authors

Bruno Wüest
Eliane Müller

Zurich, September 2020

Inhaltsverzeichnis

1 Executive summary **4**
2 Starting position **4**
3 Text corpus **4**
4 Annotation of the training data **6**
5 Machine Learning **8**
5.1 Features 9
5.2 Classification models 10
6 Goodness of fit of the classification **11**
7 Results and Sampling **12**
8 Conclusion **13**
9 References **14**
10 Appendix **15**

1 Executive summary

Within the framework of the project «Online Aggression From A Sociological Perspective», the research centre sotomo carried out a supervised classification of aggressive content in about 1.3 million comments from Openpetition.de.

For the training data set, 3006 comments were annotated by two experts as to whether the statement contained an insult, harassment, threat or other kinds of inappropriateness. On this training dataset a machine learning with 195'172 models of varying features and model parameters could be performed.

With the final model applied to all comments, a share of 3.5 and 6 percent, respectively, aggressive comments is calculated. Based on this estimate, a sample of persons stratified according to the aggressiveness of their online commentary behaviour can be drawn. In the further course of the project, surveying this sample will enable systematic comparisons between persons with aggressive and persons with non-aggressive online behaviour.

2 Starting position

The Digital Lives research project «Online Aggression From A Sociological Perspective» explores the causes of aggression in online commentary and proposes appropriate countermeasures.

An empirical core element of the project is the combination of online comments with quantitative surveys. Therefore, online comments by the petition platform Openpetition.de are first divided into comments with aggressive content and those with non-aggressive content. After this identification, a sample of online commentators stratified according to this division is drawn, who are then contacted for a survey.

The central tasks of this study on one hand include the creation of training data for the machine classification of aggressive content and on the other hand machine learning for the development of an automated text recognition of online comments according to their aggressiveness.

3 Text corpus

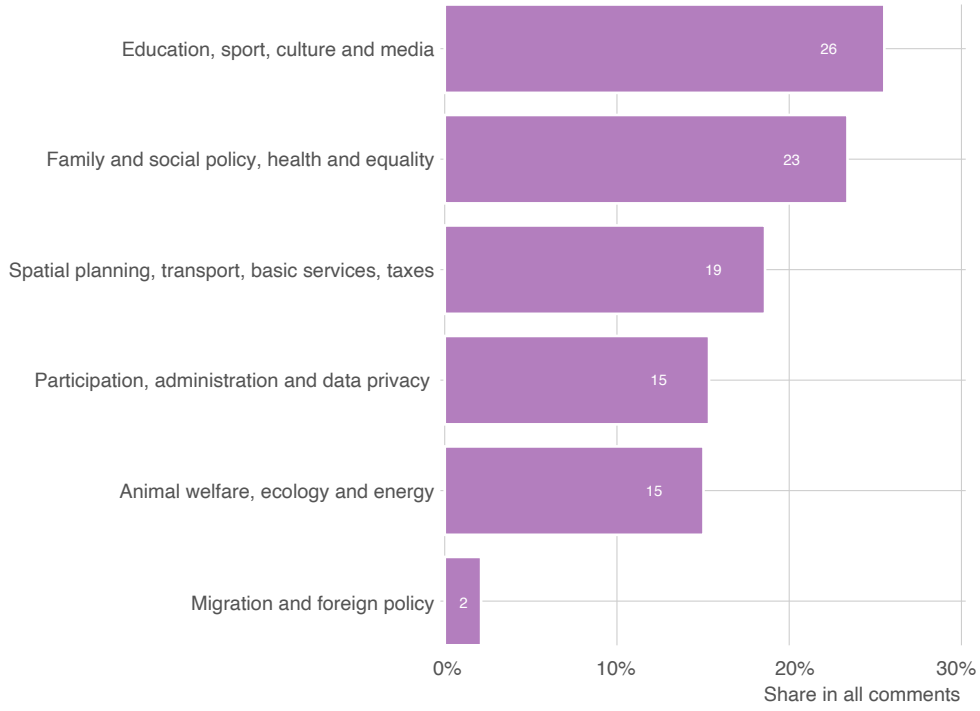
Thanks to a cooperation with Openpetition.de, a non-profit online platform for citizens' initiatives from Germany, the project can draw on around 1.3 million comments published on 4124 different petitions. These comments were submitted by 1,107,718 users of the platform who signed one or more petitions between 1 March 2019 and 29 February 2020. More than 997,000 of the commentators have written only one commentary in this period of time, while almost 85,000 of them

still wrote two. The highest number of comments from a single commentator is 2017, but this is a lonely outlier.

24'156 of the 1'276'497 comments are written in English, 3545 in French and 5206 in one of 22 other languages. Because the recognition of aggressive content is language-specific, all non-German texts except those with single words (e.g. «Nice!» or «Fuck») have been left aside. There are also 111,138 duplicate[1] or empty entries which were also omitted. In total, 1'132'452 comments are available for analysis after the data has been cleaned up.

Figure 1 shows the distribution of the comments on the political topics of the petitions. For clarity, the original 26 topic categories have been grouped into 6 larger categories. About a quarter of the comments were on concerning education, sport and culture as well as social, equality and health policy each. Infrastructure issues (transport, basic services or spatial planning) account for 19 percent, while issues related to political institutions (participation and data protection) make up for 15 percent. Ecological issues, including energy policy, account for 15 percent. Migration and foreign policy, finally, are the subject of only 2 percent of the petitions.

Abbildung 1: Share of petition topics in the comments



[1] comments entered by the same person, on the same petition, on the same day and with the same content.

The comments entered on the platform Openpetition.de are available in their unfiltered form. That means they are similar in form, style and content to other user input on the internet (social media posts, contributions in the commentary columns of online newspapers etc.). They are rather short and have a high stylistic range. Exclamations or bumpy formulations with grammatical or spelling mistakes such as «Weil es Blödsinn ist!» [«Because it's nonsense!»] and «Ich liebe Tattoos und sollte behalten bleiben» [«I love tattoos and should be kept»] appear just as much as eloquently written comments such as «Ein Verbot würde meiner Meinung nach die Kunstfreiheit, die im Grundgesetz Artikel 5 Absatz 3 verankert ist, extrem einschränken. Wie soll man die Kunstfreiheit ausüben, wenn die hierfür benötigten Farben fehlen? » [«In my opinion, a ban would extremely restrict artistic freedom, which is enshrined in Article 5 (3) of the German Constitution. How should one exercise artistic freedom if the necessary colours are missing?»].

4 Annotation of the training data

The decision to perform deductive recognition with a supervised classification led to the need for a corpus of annotated texts created specifically for the recognition task. For this purpose, Danielle Isler, Master's graduate in African Studies / Ethnology, and Eva Bühlmann, Bachelor's student at the Institute for Computational Linguistics, annotated a corpus of randomly and layered drawn comments.

The annotation work was prepared and carried out on the basis of the recommendations of Neuendorf (2017) for carrying out quantitative content analyses. For the preparation, the project team first developed a comprehensible and applicable definition of the concept of «aggressive language»:

«Aggressive is defined as content that contains an insult, harassment, threat or other inappropriateness.»

This definition was then formulated in a comprehensive description of the annotation task («codebook», see appendix) into decision rules and recommendations. Specifically, for each comment it should be answered whether the overall statement ...

... is interpreted as insulting, harassing or threatening? (B3)

... is interpreted in any other way as inappropriate? (U)

The decision between B3 and U should be based in particular on the target objects and the strength of the content. Thus, comments are aggressive in the sense of B3 if they are directed against a relatively clearly recognizable target

object[2]. To be categorized as B3, a commentary must also have rather strong or harsh wording. This includes linguistic strength features such as swear words or superlatives, but also strength features in terms of content such as the severity of the accusation. Furthermore, the indicator U is based on B3, which means that a commentary that is annotated as B3 is also included as aggressive in U.

After an initial joint discussion of the annotation instructions, three pre-tests were conducted, i.e. training rounds, each with individual processing of 20 (1st round) and 30 (2nd and 3rd round) test comments, intensive discussions of the annotations, and adaptation of the codebook. After the third training session, a weighted Cohen's Kappa of over 0.6 was measured for intercoder reliability for both target categories. According to Landis and Koch (1977) and Fleiss and Cohen (1973), this is a substantially good agreement for complex annotation tasks.

The task of evaluating the mostly short comments as aggressive without knowledge of the context is indeed difficult. Especially with regard to the target objects and the strength of the formulations, there is a comparatively large scope for subjective interpretations.

The level of 0.6 reached in the pre-tests was almost always reached in the three actual annotation rounds, as table 1 shows. For the indicator U, except in the first round, even a Kappa of over 0.7 was achieved. In the first round, the annotators were presented with comments in which half of the keywords from dictionaries were recognised as aggression.[3] For the other half, the comments were drawn at random (round 1 in table 1). The second round of annotation consisted of a purely random sample (round 2 in table 1). In the third round, half of the comments were selected completely randomly. The other half was made up of comments that were difficult to classify for a baseline classifier (a linear Support Vector Machine). Specifically, the baseline classifier showed a probability of over 90 percent for a strikingly large number of comments that they are not aggressive. For the sampling of this second half, only comments with a probability of less than 90 percent were taken into account (round 3 in table 1).

Round of annotation	Cohen's Kappa for B3	Cohen's Kappa for U	Number of comments
1	0.62	0.74	53
2	0.65	0.62	60
3	0.58	0.72	85
Overall	0.66	0.75	198

Tabelle 1: Intercoder reliability: Cohen's Kappa (weighted) for two Annotators

[2]Target objects can be individuals, social groups and organizations, but also behavioral norms, traditions, social conventions and value systems/worldviews.

[3]The dictionaries come from the predecessor project (Stahel, Rost and Gnehm 2019), from the website schimpfworte.de as well as from Hermann, Krähenbühl, Müller and Wüest (2020).

Therefore, we can assume valid, high-quality training data for the supervised classification. For the classification, two indicators were formed from the annotations. The first indicator to be estimated is simply B3, and the second is the intersection of B3 and U. The second indicator thus covers a broader, but also somewhat vague definition of aggressive language, while the first represents a concise but also rather restrictive definition.

In total, 3006 annotated comments are available for machine learning:

- 210 per annotator in round 1
- 330 per annotator in round 2
- 488 per annotator in round 3
- 475 per annotator from the follow-up of a first estimation

Table 2 shows the distribution of B3 and U in the training data. In general, it can be assumed that the real proportion of the two classes is somewhat lower, because the proportion of aggressive comments was deliberately increased in order to capture as many linguistic facets of aggressive language as possible. In the first annotation rounds keyword matches of aggressive terms probably heightened the share of aggressive comments. The proportion was also increased in the third round of annotation with somewhat difficult cases from a baseline classification. This distortion of the training data is counteracted with a two-step procedure, where the first step is a filter to exclude as many non-aggressive comments as possible.

Class	N for B3	Shares in B3	N for U	Shares in U
aggressive	471	16 %	772	26 %
not aggressive	2532	84 %	2234	74 %

Tabelle 2: Overview over the training data

5 Machine Learning

To prepare the machine learning, features[4] were extracted from the commentary texts, the metadata of the texts, preparatory analyses and the parameters of the models. For all of the 576 possible combinations of the mentioned features, it was therefore checked whether including or omitting them would optimize the

[4]Features in machine learning correspond to the independent variables of empirical social science, i.e. the properties of the objects of investigation, which are used for the estimation of a further property, in this case the aggressiveness of the texts.

estimation of the aggressiveness in the texts. The text-level features can reduce or enrich the information content of the texts. Using the metadata features, important exogenous information about the comments can be included into the calculation. The pre-analyses can add non a priori existing features like the sentiment to the estimation.

5.1 Features

For the features at the level of commentary texts, four different text pre-processing steps were initially used (see Jurafsky and Martin 2009, pp. 2-28). This means that for each classification it was tested whether the texts are best included in the calculation a) in lower case letters, b) with removed stop words[5], c) with only the root forms of the words or d) with the words and their part of speech.[6] It was also considered whether a TfIdf weighting («term frequency / inverse document frequency»; Manning, Raghavan and Schütze 2008, p. 118) would improve the estimate. Finally, the frequency of the words as such, the frequency of the 2-grams[7] and the frequency of the combination of individual words and 2-grams was tested.

As metadata the already mentioned topic categories of the petition were available, but also the number of comments per user and the status of the users on the petition platform. The latter contains the information whether the users are registered anonymously or their comments are entered in a blacklist.

In addition, the following indicators have been established on the basis of preparatory analyses. Using an opinion mining method according to Rauh (2018), one positive and one negative sentiment each was calculated and included in the estimate. As additional features from the preparatory analyses, it was tested whether the summarized keyword hits from the dictionaries for aggression by Stahel, Rost and Gnehm (2019), from the website schimpfworte.de and from Hermann, Krähenbühl, Müller and Wüest (2020) were able to improve the classification. Finally, information was included on whether there was excessive use of exclamation and question marks and capital letters (e.g. «WAS SOLL DAS!!!!» [approx. «WHAT THE HELL!!!!»]).

All continuous meta-indicators, i.e. sentiment and number of comments per user, were standardized from 0 to 1. The petition topic categories were considered as

[5] Stop words are functional words that often have little content substance, such as pronouns or auxiliary verbs.

[6] As an example, the sentence «Das ist eine Frechheit.» [«This is an outrage.»] in simplified form is prepared in the following four variants: a) «das ist eine Frechheit» [«this is an outrage»], b) «frechheit» [«outrage»], c) «das sein ein Frechheit» [«that being an outrage»] and d) «das-ARTIKEL ist-VERB ein-PRONOMEN Frechheit-NOMEN» [«the-ARTICLE is-VERB a-PRONOUN outrage-NOUNE»].

[7] N-grams are the combinations of N consecutive words. The 2-gram of the exclamation «So eine Frechheit» [«Such an outrage»], for example, is «So-eine eine-Frechheit» [«Such-an an-outrage»] (see Jurafsky and Martin 2009, p. 31).

binary indicators («One-Hot-Encoding», Cerda, Varoquaux and Kégl 2018). The remaining indicators were binary-encoded and used as such.

5.2 Classification models

The first model tested was a simple naïve Bayes with varying additive Laplace smoothing parameters (Zhang 2004a). Naïve Bayes models apply Bayes' theorem with the assumption of conditional independence between every pair of features. Such models are usually very performant and reliable, thus providing a good basis for estimating the potential of a supervised classification on the available data. Subsequently, various «loss»-functions and regularization terms were evaluated for a Support Vector Machine (Zhang 2004b).[8] The specific Support Vector Machine used here implements regularized linear models with stochastic gradient descent learning. These models are generally very powerful in automated text analysis and were an industry standard before the advent of neural networks.

There are individual text classification tasks where hierarchical models outperform the two mentioned above. Therefore, a nearest a Random forest classifier with varying maximum tree length was tested (Tibshirani 2002, Breimann 2001) in order to verify that the classification problem of aggressive language is not better solved by alternative approaches.

Further, a neural network was considered the most promising model. Neural networks have also revolutionized the field of automated text analysis in recent years (Glorot and Bengio 2010). The multi-layer Perceptron classifier used here calculates with 100 hidden layers, which are activated by a rectified linear unit function.

Finally, word embeddings based models have become very popular in recent years (see Joulin et al. 2016), whereby fastText is a very efficient and versatile implementation of such models. In order to optimize the classification of aggressive content, the number of epochs and the learning rate of fastText has been systematically varied.

In total, the combinations of all features and model parameters resulted in a grid search of over 195,172 models for the model that best masters the task of estimating B3 and U[9]. To reduce the risk of overfitting[10] each model was validated on three different, randomly drawn samples of training data. Each draw divides the training data into a training set (80% of the data) and test set (20%). The former is used for machine learning, the latter is retained for testing the trained model.

[8] For the Support Vector Machine and the Random forest classifier (see next section), it was also possible to perform the estimates weighted by class size.

[9] The estimates were performed with «scikit-learn» (see Pedregosa et al. 2011).

[10] «Overfitting» refers to a situation where a model fits the training data too well but cannot be generalized well for application to new data (Hawkins 2004).

6 Goodness of fit of the classification

Because from the beginning a strongly skewed distribution of the classes was assumed[11], the aggressive comments were not only drawn excessively often for the training data, but a two-stage procedure was designed.

- A first classification round should detect as many non-aggressive comments as possible with a very high degree of precision, in order to better balance the ratio between aggressive and non-aggressive comments for the second round. Comments rated as non-aggressive in this round will not be considered for the second round and will be shown as non-aggressive in the final results. Precision means the minimization of false positives, i.e. comments that are classified as aggressive by the classification but are not aggressive. 90 percent was set as the minimum target.
- A second round of recognition is designed to detect aggressive comments with the highest possible precision and exhaustion, thus allowing a final distinction between aggressive and non-aggressive comments. Exhaustion means the minimization of false negatives, i.e. aggressive comments that were not recognized as such by the classification. 85 percent for F1, the harmonic mean between precision and exhaustion, was set as a minimum target.

The first round therefore focuses on precision in the recognition of non-aggressive comments. The exhaustion is of secondary importance, but should also not be too low. In other words, a considerable number of non-aggressive comments and at the same time as few aggressive comments as possible should be excluded. Table 3 shows the evaluation of the estimate of B3 – insulting, harassing, or threatening comments – and U – additionally inappropriate statements – for the best runs of the five models used.

	B3			U		
model	**precision**	**recall**	**F1**	**precision**	**recall**	**F1**
Naïve Bayes	0.89	0.91	0.90	0.85	0.82	0.83
Support Vector Machine	0.98	0.35	0.48	0.96	0.21	0.32
FastText	0.89	0.94	0.91	0.85	0.90	0.87
Multi-layer Perceptron	0.90	0.92	0.91	0.87	0.89	0.88
Random forest	0.87	0.99	0.92	0.83	0.97	0.90

Tabelle 3: Key figures of the first machine learning round for the recognition of non-aggressive comments

The first figures for the Naïve Bayes model, which is intended as a baseline to sound out the difficulty of the task, show that in principle very good precision

[11] The rough expectation based on earlier studies (Stahel, Rost and Gnehm 2019; Hermann, Krähenbühl, Müller and Wüest 2020) was that about 5 to 10 percent aggressive comments could be found.

can be achieved. . For B3 and U, the model is already just below the minimum requirement. Of the other models, the Support Vector Machine is slightly ahead of the other two models. Based on the experience in comparable studies (Struß et al. 2019) it is somewhat surprising that the neural network model (Multi-layer Perceptron) as well as fastText do not perform better.

The first classification round at B3 should therefore clearly be carried out with the Support Vector Machine[12] with the following features: POS-tagging for preprocessing, Tfldf weighting and no N-Gram, modified huber loss function, l2-regularisation and exclamations and capital letters, hits with aggressive keywords as well as the number of comments per user as pre-analysis and metadata features.

With U, the Support Vector Machine again comes into play with the following features: Stemming for preprocessing, no weighting and 2-grams, log loss function, l2-regularisation and sentiment, exclamations and capital letters, hits with aggressive keywords, the subject categories of the petitions as well as the number of comments per user as pre-analysis and metadata features.

The evaluation for the second machine learning round in Table 4 shows that the minimum requirements can only be exceeded for B3. For U, the average of precision and recall over both classes remains slightly below the requirements for all models. The Support Vector Machine is only just ahead of the neural network model with a slightly higher precision and F1 value for B3. The features used are stoppword removal (B3), stemming (U), Tfidf weighting, no N-grams (B3) as well as words and n-grams (U), modified huber loss function, l2-regularisation as well as user status (U only), exclamations and capital letters and hits with aggressive keywords as pre-analysis and metadata features.

	B3			U		
model	**precision**	**recall**	**F1**	**precision**	**recall**	**F1**
Naïve Bayes	0.82	0.85	0.83	0.78	0.79	0.79
Multi-layer Perceptron	0.84	0.86	0.85	0.83	0.83	0.83
Support Vector Machine	0.85	0.86	0.86	0.83	0.83	0.83
Random forest	0.87	0.87	0.82	0.83	0.83	0.81
Nearest centroid	0.84	0.86	0.84	0.80	0.80	0.80

Tabelle 4: Key figures of the second training round: Average values over both aggressive and not aggressive comments weighted by the number of comments in the test set

7 Results and Sampling

The application of the support vector machines, that were specifically trained in each of the two classification rounds, yields a recognition of aggressive comments

[12] A qualitative follow-up check of 20 randomly drawn, non-aggressive comments also confirmed that the 98 percent precision is not an artifact, for example due to overfitting of the classifier.

that is summarized in table 5. For B3, it is estimated that the proportion of aggressive comments is 3.5 percent. For U, it is estimated to be 6 percent. Since this is rather in the lower range of what could have been expected (see Stahel, Rost und Gnehm 2019 and Hermann, Krähenbühl, Müller and Wüest 2020), there is probably no overestimation of the proportion of aggressiveness in online comments.

Classification	N for B3	Shares in B3	N for U	Shares in U
aggressive	39'102	3,5 %	68'350	6 %
not aggressive	1'093'350	96,5 %	1'064'102	84 %

Tabelle 5: Overview over the classified data

Because the evaluation of machine learning for B3 produced the better results, this indicator was preferred to U. For the sampling, however, the commentators and not the comments themselves are the relevant unit of analysis. This is because the text classification needs to be able to connect to the survey, which is the next step within the research project. The aim is to draw a sample from the total of 995,610 persons, who are deliberately stratified according to their commentary behavior. First, only those commentators who wrote at least one commentary within one year before the expected field start of the survey on September 20, 2020 were considered. This is still 456,822 persons. Secondly, three groups were formed from the results of the text recognition for the survey. From the first group of 437,632 commentators who did not make an aggressive comment, 8500 were randomly selected. Likewise, 8500 of the 18'838 people who made an aggressive comment were randomly selected. The remaining 574 commentators with more than one aggressive comment were included as a third group.

8 Conclusion

The research center sotomo was able to carry out a supervised machine learning task on behalf of the project «Online Aggression From A Sociological Perspective». With a sample of about 3000 annotations and extensive training on text-related features, metadata on comments and parameters of five classification algorithms, it was possible to perform the recognition with high quality. The results made it possible to draw a sample stratified according to the aggressiveness of the online commentary behaviour, thus taking the research project one step further.

9 References

Breiman, Leo (2001). «Random Forests». Machine Learning, 45, 5–32. doi: 10.1023/A:1010933404324

Cerda, Patricio, Gaël Varoquaux and Balázs Kégl (2018). «Similarity encoding for learning with dirty categorical variables». Machine Learning 107, 1477–1494. doi: 10.1007/s10994-018-5724-2.

Fleiss, Joseph L. and Jacob Cohen (1973). «The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability». Educational and Psychological Measurement. 33 (3): 613–619. doi:10.1177/001316447303300309.

Glorot, Xavier and Yoshua Bengio (2010). «Understanding the difficulty of training deep feedforward neural networks». Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010). Proceedings of Machine Learning Research 9, 249–256.

Hawkins, Douglas M. (2004). «The problem of overfitting». Journal of Chemical Information and Modeling. 44 (1): 1–12. doi:10.1021/ci0342472.

Herman, Michael, David Krähenbühl, Eliane Müller and Bruno Wüest (2020). «Digitale Debatten in Zeiten der Corona-Krise. Eine Analyse von Kommentaren auf Twitter und Online-News-Plattformen». Forschungsstelle sotomo.

Jurafsky, Daniel and James H. Martin (2009). «Speech and Language Processing. An Introduction to Natural Language Processing,Computational Linguistics, and Speech Recognition». Upper Saddle River, NJ: Prentice Hall. doi: 10.5555/1214993.

Landis, J. Richard and Gary G. Koch (1977). «The measurement of observer agreement for categorical data». Biometrics, 33 (1), 159–174. doi:10.2307/2529310.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze (2008). «Introduction to Information Retrieval». New York, NY: Cambridge University Press.

Neuendorf, Kimberley (2017). «The content analysis guidebook». Thousand Oaks, CA: SAGE Publications. doi: 10.4135/9781071802878.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot and Édouard

Duchesnay (2011). «Scikit-learn: Machine Learning in Python». Journal of Machine Learning Research, 12(85), 2825–2830.

Rauh, Christian (2018). «Validating a sentiment dictionary for German political language – a workbench note». Journal of Information Technology and Politics. 15 (4), 319–343. doi: 10.1080/19331681.2018.1485608.

Stahel, Lea, Katja Rost and Anne-Sophie Gnehm. (2020). «Do moral judgements fuel online firestorms? Moral illegitimacy judgements and aggressive commenting in a social media phenomenon». Working Paper. University of Zurich.

Struß, J. Maria, Melanie Siegel, Josep Ruppenhofer, Michael Wiegand and Manfred Klenner (2012). «Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language». Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019). German Society for Computational Linguistics and Language Technology.

Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan and Gilbert Chu (2002). «Diagnosis of multiple cancer types by shrunken centroids of gene expression». Proceedings of the National Academy of Sciences of the United States of America, 99 (10), 6567-6572.

Zhang, Harry (2004a). «The optimality of Naive Bayes». Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS). AAAI Press.

Zhang, Tong (2004b). «Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms». Proceedings of the Twenty-First International Conference on Machine Learning (ICML). Omnipress.

10 Appendix

sotomo
gesellschaft, politik & raum

ONLINE AGGRESSION FROM A SOCIOLOGICAL PERSPECTIVE

AN INTEGRATIVE VIEW ON DETERMINANTS AND POSSIBLE COUNTERMEASURES

Annotationsanweisungen zur Identifikation von aggressiven Inhalten in Online-Kommentaren

Dr. Lea Stahel – Universität Zürich
Dr. Sebastian Weingartner – Universität Zürich
Dr. Bruno Wüest – sotomo
Eliane Müller – sotomo

Zürich, 4. Juni 2020

Forschungsstelle sotomo
+41 44 525 84 01 - www.sotomo.ch - info@sotomo.ch

A. Überblick

Diese Annotationsanweisungen wurden für die Erstellung von Trainingsdaten für eine maschinelle Erkennung von aggressiven Inhalten in Online-Kommentaren im Rahmen des Projekts «Online Aggression From A Sociological Perspective» erstellt. Aufgrund der negativen sozialen Folgen der weit verbreiteten Aggression in Online-Kommentaren ergründet das Projekt die Ursachen von Aggression und schlägt geeignete Gegenmassnahmen vor.

Ein empirisches Kernelement des Projekts ist die Verbindung von Online-Kommentaren mit quantitativen Befragungen. Die von der Petitionsplattform Openpetition.de gespendeten Online-Kommentare werden zunächst mit Hilfe einer automatisierten Textanalyse in Kommentare mit aggressivem Inhalt und solche mit nicht aggressivem Inhalt aufgeteilt. Dann wird eine nach dieser Aufteilung geschichtete Stichprobe von Online-Kommentierenden gezogen, welche anschliessend für die Befragungen kontaktiert werden.

Ihre Aufgabe steht im Zusammenhang mit der automatisierten Textanalyse der Online-Kommentare und beinhaltet die Erstellung von Trainingsdaten für die maschinelle Erkennung von aggressiven Inhalten. In anderen Worten werden die von Ihnen erstellten Annotationen von Kommentaren dafür verwendet, regelbasierte, statistische und linguistische Verfahren zu trainieren, damit die aggressive von nicht aggressiven Inhalten unterschieden lernen.

B. Untersuchungsgegenstand

Die auf der Plattform Openpetition.de eigegebenen Kommentare sind ähnlich zu anderen Nutzereingaben im Internet (Social Media Posts, Beiträge in den Kommentarspalten von Online-Zeitungen etc.) eher kurz und sie variieren formal und stilistisch auch relativ stark. Sie dürfen sich also nicht durch Rechtschreibfehler oder unpassenden Formulierungen aus der Ruhe bringen lassen.

Ebenso möchten wir Sie darauf hinweisen, dass der Untersuchungsgegenstand des Projekts inhaltlich unangenehme Formen annehmen kann. Wenn Sie diesbezüglich Unterstützung brauchen, können Sie sich jederzeit an das Projektteam wenden. Und werden Sie sich auch bewusst, dass Sie mit Ihrer Teilnahme einen wichtigen Beitrag dazu zu leisten, aggressives Online-Verhalten besser zu verstehen.

C. Aufgabenstellung und Vorgehen

Sie erhalten eine Excel-Tabelle mit einer Auswahl von Kommentaren für eine manuelle Datenerhebung. Ihre Aufgabe besteht dabei darin, diese Kommentare auf aggressive Inhalte hin zu untersuchen. Als aggressiv werden Inhalte definiert, die eine Beleidigung, Belästigung, Bedrohung oder sonstige Unangemessenheit enthalten. Die Beleidigung, Belästigung, Bedrohung oder sonstige Unangemessenheit kann sich ganz unterschiedlich äussern, z.B. in vulgärer und drohender Sprache, aber auch mit anderen rhetorischen Mitteln wie Stereotypisierungen oder Ironie ausgedrückt werden.

Konkret sollten Sie zu jedem Kommentar die folgenden zwei Fragen mit ja (= 1) oder nein (= 0) beantworten. Kann die Gesamtaussage des vorliegenden Kommentars...

...als beleidigend, belästigend oder bedrohend interpretiert werden? (B3)
...auf eine andere Art als unangemessen interpretiert werden? (U)

Die Einteilung soll sich insbesondere nach den Zielobjekten und der Stärke des Inhalts ausrichten.

Zielobjekte

Eine Beleidigung, Belästigung oder Bedrohung richtet sich gegen ein relativ klar erkennbares Zielobjekt. Allerdings können diese Zielobjekte sehr unterschiedlich sein und Folgendes umfassen:

- Einzelpersonen (z.B. Merkel, Promi, Bürgermeister)
- Soziale Gruppen (z.B. Migrant*innen, Politiker*Innen, «Linke», Jogger)
- Organisationen (z.B. Unternehmen, Behörden, Schulen)
- Abstrakteres, z.B.:
    - Verhaltensnormen, Traditionen und gesellschaftliche Konventionen (z.B. das 1.-Augustfeuerwerk, Gesetze, Schlange stehen)
    - Wertehaltungen/Weltanschauungen (z.B. Neoliberalismus, Tier-/Umweltschutz, Religion)

Wir machen generell keine Unterscheidungen zwischen verschiedenen Personen, Sozialen Gruppen und Organisationen. So können zum Beispiel «Migranten» wie auch «Rassisten» das Zielobjekt von Beleidigungen, Belästigungen oder Bedrohungen sein. Meistens geht es in den Kommentaren nicht um so klar konnotierte Personen, Sozialen Gruppen und Organisationen wie im Fall der klar negativ besetzten «Rassisten». Um deshalb Unterschiede in der Annotation zu vermeiden, werden alle genannten Personen, Sozialen Gruppen und Organisationen als Zielobjekte betrachtet.

Stärke des Inhalts

Um als Beleidigung, Belästigung oder Bedrohung kategorisiert zu werden, muss ein Kommentar erstens *eher* starke bzw. krasse Formulierungen aufweisen (Kraftausdrücke, Schimpfworte, Superlative etc.). Eine Formulierung kann aber auch inhaltlich und nicht nur linguistisch stark oder krass sein, wie die beiden Beispiele mit der gleichen Wortwahl aber komplett unterschiedlicher inhaltlicher Schärfe zeigen: «Wer eine schwere Straftat begeht, muss ins Gefängnis», «wer jüdischen Glaubens ist, muss ins Gefängnis». Zweitens sollte dabei auch die Absicht, einem Zielobjekt zu schaden, erkennbar sein.

Wichtig für die Unterscheidung zwischen Beleidigung, Belästigung oder Bedrohung auf der einen Seite und sonstiger Unangemessenheit auf der anderen Seite ist, dass bei Ersterem tendenziell eine Kombination von relativ klar erkennbarem Zielobjekt und starker/krasser Formulierung vorliegen sollte. Bei Zweiterem kann die Ziel-Komponente wegfallen oder der Angriff gegen ein Zielobjekt weniger stark/krass formuliert sein. Beispielsweise hat die Aussage „So ein Scheiss"

kein klares Ziel und ist einfach unangemessen, während „Die Politik der Bürgermeisterin ist doch einfach nur beschissen“ mit der Bürgermeisterin ein klares Ziel der Aggression beinhaltet.

Spezialfall «ich hasse»

Aussagen mit der Formulierung «ich hasse» werden als U eingestuft, wenn sie sich auf Einzelpersonen oder soziale Gruppen beziehen (z.B. «ich hasse Migranten»). Wenn sie sich auf Dinge, Organisationen oder Abstrakteres beziehen, werden sie tendenziell weder als B3, noch als U eingestuft. Eine Einstufung als U ist nur möglich, wenn eine der untenstehenden Bedingungen für Unangemessenheit vorliegt (z.B. übermässige Ausrufezeichen).

Spezialfall Skandalisierung

Skandalisierende Formulierungen wie «es ist eine Frechheit», «es ist ein Skandal», «Unverschämtheit» werden tendenziell weder als B3, noch als U eingestuft. Eine Einstufung als U ist nur möglich, wenn eine der untenstehenden Bedingungen für Unangemessenheit vorliegt (z.B. übermässige Ausrufezeichen, weniger starke Stereotypisierung).

Allgemeine Anmerkungen

Generell gilt, dass der gesamte Inhalt eines Kommentars für die Bestimmung der Aggressivität zu berücksichtigen ist. Dabei gilt, dass die beiden Fragen (B3 vs. U) hierarchisch und exklusiv zu einander stehen. Konkret bedeutet das: sobald eine Äusserung in einem Kommentar einer Beleidigung, Belästigung, oder Bedrohung entspricht, ist der gesamte Kommentar als beleidigend, belästigend oder bedrohend einzustufen. Es ist dann nicht mehr möglich, den Kommentar zusätzlich auf eine andere Art unangemessen einzustufen. Letzteres ist nur dann der Fall, wenn eine unangemessene Äusserung in einem Kommentar vorliegt, die nicht eindeutig einer Beleidigung, Belästigung, oder Bedrohung entspricht. In anderen Worten muss die zweite Frage also nur beantwortet werden, um herauszufinden, ob ein Kommentar auf eine andere Art unangemessen ist, wenn vorher keine Beleidigung, Belästigung, oder Bedrohung erkennbar waren.

Interpretieren Sie immer nur so viel wie für das Verständnis des Inhalts nötig ist. Das heisst, wenden Sie Ihr Kontextwissen und Heuristiken nur an, um inhaltliche Lücken zu füllen. Versuchen Sie also Ihre Einschätzungen der Aggressivität eines Inhalts möglichst ohne Überlegungen zu den Autor*innen oder dem thematischen Zusammenhang des Kommentars zu treffen.

Versuchen Sie auch, die Schwierigkeit bei der Beantwortung einer Frage als Entscheidungskriterium mit einzubeziehen. Müssen Sie beispielsweise überdurchschnittlich lange überlegen, ob eine Aussage ein bestimmtes Objekt zum Ziel hat, gibt es eventuell wirklich kein klares Ziel. In diesem Fall müsste der Kommentar also «lediglich» als unangemessen oder eventuell auch gar nicht als aggressiv kategorisiert werden.

Beleidigend

Erläuterung:

- relativ explizit formulierte Herabsetzungen, Feindseligkeiten und Beschimpfungen
- Aussagen, die auf abwertende Art die psychische Gesundheit, Würde oder Rechtmässigkeit absprechen.

Beispiele:

- *wir haben ALLE ZUSAMMEN MEHR MACHT als irgendwelche geldgeile Mongos mit kleinwüchsigen Verstand..*
- *Diese scheiß verfickten Wix shit Brüder sind solche dummen idiotischen Volidioten Macho Ärsche die sich für verdammt cool halten aber in echt kleine Trottel ohne Sinn!*
- *Migranten sind Schmarotzer. Die sollen dorthin zurück wo sie herkommen.*
- *Die hat doch nicht mehr alle Tassen im Schrank!*
- *Der Kapitalismus ist ein Scheisssystem!*

Belästigend

Erläuterung:

- Sexualisierende, vulgäre oder anzügliche Bemerkungen, die entwürdigend bzw. beschämend wirken.

Beispiele:

- *Die Dame würde ich auch gerne mal rannehmen.*
- *Männer haben ihr Hirn in der Hose.*
- *Der Typ ist doch Alkoholiker.*

Bedrohend

Erläuterung:

- subtile oder direkte Gewaltandrohung, Gewaltfantasien, Aufstacheln zu Gewalt, Rechtfertigen und Verharmlosen von Gewalt.
- Androhung bzw. Nahelegung rechtlicher Sanktionen.

Beispiele:

- *Irgendwann wird es wohl Bombendrohungen oder ähnliches für diese Abzocker-Anstalten geben...*
- *Diese Beamten sollten öffentlich ausgepeitscht werden.*
- *Dem sollte man mal einen Hausbesuch abstatten.*
- *Die gehört in den Knast und nicht in eine Regierung.*
- *Irgendwann werde ich dieses System zerstören.*

Auf eine andere Art unangemessen

Erläuterung:

- was unter diese Kategorie fällt, entspricht einer subjektiven Einschätzung der Unangemessenheit, die nicht durch die drei obigen Kategorien abgedeckt ist. Darunter können z.B. fallen:
    - Weniger starke Pauschalisierungen oder Stereotypisierungen
    - Flüche (ohne klares Ziel)
    - Stark negativ-emotionale Äusserungen (u.a. signalisiert durch übermässige Verwendung von Grossbuchstaben und Ausrufezeichen)

Gehen Sie am besten für jeden Kommentar wie folgt vor:

1. Lesen Sie den Kommentar sorgfältig und werden Sie sich im Klaren, ob Sie den ganzen Inhalt erfasst haben. Lesen Sie den Kommentar wenn nötig nochmals durch.
2. Beantworten sie die zwei Fragen, indem Sie einen Eintrag (0 für nein oder 1 für ja) in das entsprechende Feld einer Kategorie machen. Wenn Sie sich unsicher sind, machen Sie bitte auch einen Eintrag in das entsprechende Feld -- im Zweifelsfall einen Kommentar also als aggressiv einstufen --, tragen aber Ihre Unsicherheit in der separaten Spalte dafür ein. Wenn Sie die erste Spalte mit „ja“ beantworten, können sie die zweite Spalte leer lassen und zum nächsten Kommentar gehen.
3. Überlegen Sie regelmässig selbstkritisch, ob sich Ihre Einschätzungen und somit Ihr Annotationsverhalten im Laufe Ihrer Arbeit geändert hat. Ist dies der Fall ist, korrigieren Sie bitte Ihre vorhergehenden Annotationen entsprechend.
4. Falls Sie in den Kommentaren Formen von Aggression erkennen, die Sie in keine der obigen Definition von Aggression einordnen können, formulieren Sie eine annähernde Definition dieser Aggressionsform in der separaten Kommentarspalte.

Lassen Sie sich bei den ersten Annotationen etwas mehr Zeit, um eine gewisse Routine bei der Anwendung der Definitionen und Einteilungen zu erhalten. Wenn Sie sich sicher fühlen, können Sie das Tempo gerne schrittweise erhöhen.

Bei Fragen zur Annotation wenden Sie sich bitte an Bruno Wüest (bruno.wueest@sotomo.ch).

Vielen herzlichen Dank für Ihre wertvolle Mitarbeit!