Algorithm Aversion

Kopkow, Angermaier, Rohrer, Sonnleitner

February 2022

Contents

Motivation **2**

Data retrieval **2**

Data processing **3**

Analysis **3**

Progression over the years 4

Distribution of classes by topic 5

Conclusion: **5**

Critique **6**

Biblography **7**

Github Repository **7**

Motivation

The utilization of algorithms in a growing number of areas of our everyday lives and their impact on them demonstrates that the topic cannot be elaborated solely from a data science point of view but needs to be considered from a broad interdisciplinary perspective. Psychological, sociological, economical and judicial implications and different methodical approaches need to be integrated. In our current study we want to complement the current research on algorithm aversion with an additional methodical approach in terms of a twitter sentiment analysis. Algorithm aversion either gets defined as a "general resistance to algorithmic judgment" or general negative reaction towards an erring algorithm (Berger et al., 2021). Recent studies suggests that algorithm aversion is among other things less prevalent under the following circumstances:

The results of Dietvorst et al. (2016) confirmed the substantial influence of perceived control over an algorithm's output on the utilization of that algorithm. Participants who were allowed to correct the algorithm's outcame were far more likely to use the algorithm in their forecast. Furthermore it increases the satisfaction with the whole process. Interestingly their results showed that the amount of how much the participants were allowed to correct the algorithm has no effect on the utilization of the algorithm or the satisfaction of the participants.

Berger et al. (2021) found that there is no generell preference of unfamiliar human support over unfamiliar algorithmic support, but experiencing both of them erring leads to preference of human over algorithmic advisors. The reason for this is the implicit assumption of most humans that humans can learn from their mistakes whereas algorithms can not. The results of Berger et al. indicate that explicitly showing an algorithm's ability to learn counteracts algorithm aversion. Castelo et al. (2019) explored under which circumstances and in which domains of application humans are more reluctant to use algorithmic aids. Despite the fact that algorithms often outperform humans in many domains, users trust algorithms even less in subjective tasks and in domains where subjective human expert judgment is perceived as superior to algorithmic judgment. As already partly mentioned, human decision makers have a lot of expectations and prejudices on what an algorithm can and should do, which prevent them from using them. Burton et al. (2020) postulate that development of algorithmic literacy in decision makers will lead to reduced aversion against them. To explore if there is a general aversion against algorithms on twitter and if growing algorithmic literacy in some areas decreases this aversion, we came up with our first research question:

Is there a significant aversion against algorithms on twitter and has it changed over time (last 10 years)?

To explore the impacts of perceived control, the algorithm's ability to improve and the areas of application on algorithm aversion, we came up with our second research question:

Are there differences in algorithm aversion in dependence of influenceability, immutability and setting of application?

Data retrieval

Data retrieval and analysis was performed entirely in R using the respective framework RStudio. To obtain a sufficient amount of data over a period of 11 years, the "academictwitteR" package was used. Its function get_all_tweets() allows not only a search for Tweets containing a specific word, but also to set up a timespan, from where Tweets should be obtained. In order to reduce seasonal effects, 1000 tweets were retrieved from each month on a random day from 2010 to 2021 included. The resulting dataset thus includes 143,721 tweets (on some days, not exactly 1000 tweets were found).

As text analysis tool, we decided to use the unsupervised "VADER" method, as it turned out in the lecture and one according exercise, that it has advantage in precision and reliability compared to other tools like "Syuzhet". With its capabilities, we analyzed a sentiment value for every tweet, ranging from -1 to 1.

A further important resource is the "tidytext" library, especially the unnest_tokens() function in order to categorize tweets into various topics. After searching for words on the Merriam-Webster thesaurus, that do indicate the following topics: Business, Technology, Social Media, Immutability, Changeability, Application,

Aversion. Those word patterns where used to identify a tweet. To do this, the text was disassembled, stopwords were erased in order to maintain better efficiency, and the remaining text was compared to the before initialised word patterns.

Data processing

The value results from the VADER analysis was categorized into four sentiments. A value higher than 0.3 indicates a positive sentiment, between 0.3 and -0.3 a neutral sentiment, between -0.3 and -0.7 a negative sentiment and lower than 0.7 was considered an aversive sentiment. The treshold for the aversive category was set through trying various values, reading the corresponding tweets to find out, what threshold offers an accurate indicator for aversiveness. The aformentioned word pattern comparison uses the number of indicator words in a tweet and saves that value in the respective topic column. With that information, we can than link the tweet to one specific topic. Thanks to this procedure, we are able to group the tweets into the existing topics, which is crucial for research question 2. Research question 1, the development over the years, could be answered by getting the year of a tweet's timestamp, save it in a corresponding column and then group by those years.

The topic categorization needed to be further validated, as our group wanted to make sure, that tweets were ordered accordingly and the significance of such a content check is assured. In order to do so, 30 random tweets of every topic were read and than decided, whether they truly fit the topic. Especially the topics Business, Technology and Social Media were accurately assigned in all of the 30 tweets. The topics Immutability, Influence and Aversion had some ambiguities in a few tweets, but overall still mostly were realted to the topic. Only the topic aversion showed inconsitencies and expressed a different meaning as we intended it to be.

With all that information we have every tweet's topic, sentiment, and year. This enables according filtering and grouping. As mentioned before, research question 1 is analyzed via grouping the years for the overall development. We then calculated the ratio of the four sentiments for every year and created a plot that shows the development via a line plot. To answer research question 2, grouping the topics was necessary. By calculating the percentage of every sentiment for every topic we could create a bar chart to indicate every topic's sentiment ratio. A further interesting look was on the topical sentiment ratio over time. This can be seen as a combination of both research questions. We created a line plot over time but for every topic separately. This enables deeper insight if algorithm aversion developed differently for specific topics. As we had to limit this report, the results can be found in our Git repository.

During our research we increasingly became interested in further topics, that indicated own distributions in detecting algorithm aversion. Therefore we enhanced research question 2 by analyzing the fields of Business, Technology and Social Media.

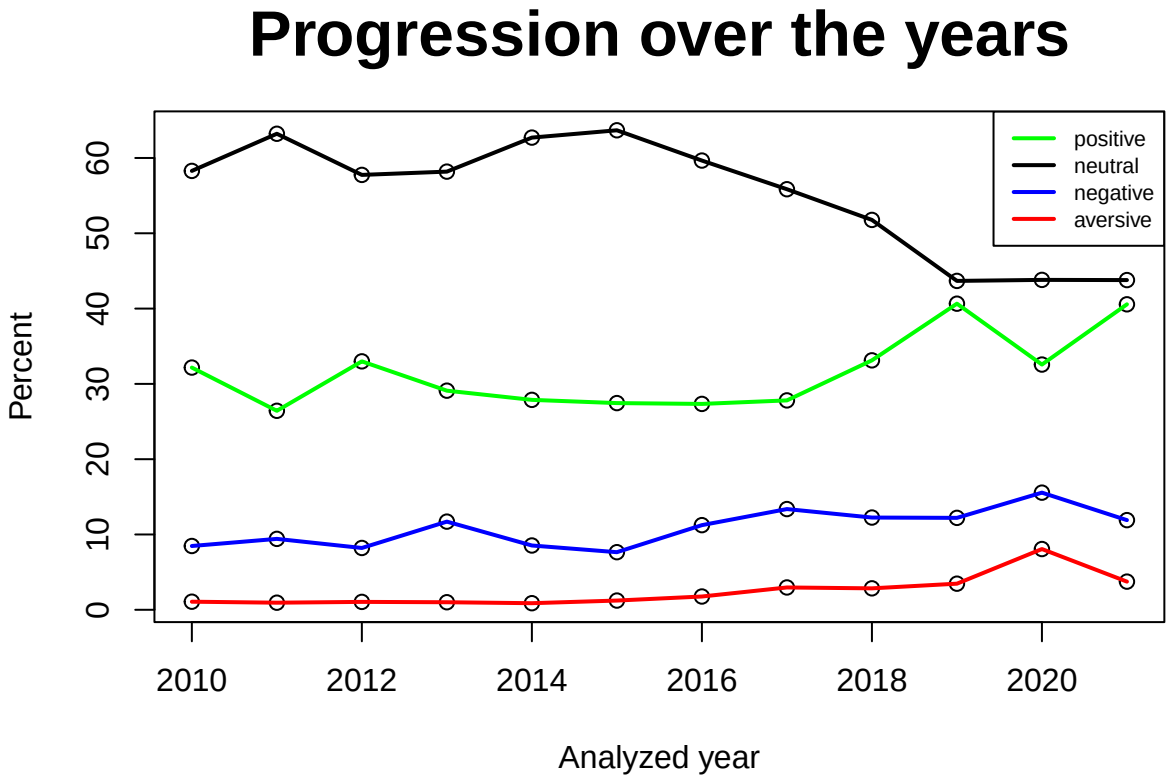
Analysis

The following chapter covers the results of the sentiment analysis of tweets over the years 2010 to 2021 , for which we used the Vader tool. Values between -0.3 and 0.3 were classified as neutral. Those between 0.3 and 1 were classified as positive, as negative between -0.3 and -0.7 and finally as aversive values between -0.7 and -1.

A Confidence interval based on 10000 bootstrap replications was calculated. It revealed that with a probability of 95% that the mean value of the sentiment analysis falls between 0.0766 and 0.1349.

Progression over the years

The diagram shows the percentage progression of the results of the sentiment analysis of tweets over the years 2010 to 2021.A total of 143271 tweets were analyzed. Since all tweets that could not be evaluated in the sentiment analysis were classified as neutral, the highest number of tweets was yielded in the neutral category while the fewest tweets were classified as aversive.



Over time, the number of neutral tweets decreases. In 2010, there were 6934 neutral tweets which amounts to 58.3% of the yearly total. In 2015, the highest number of neutral tweets were posted, with an absolute of 7641 tweets posted (63.8% of the yearly total). In 2019, the lowest number of neutral tweets (5241, 43.7%) was posted. 2021 yielded similar numbers (5254, 43.8%). Over the last 3 years, the number of neutral tweets remained at the same level.

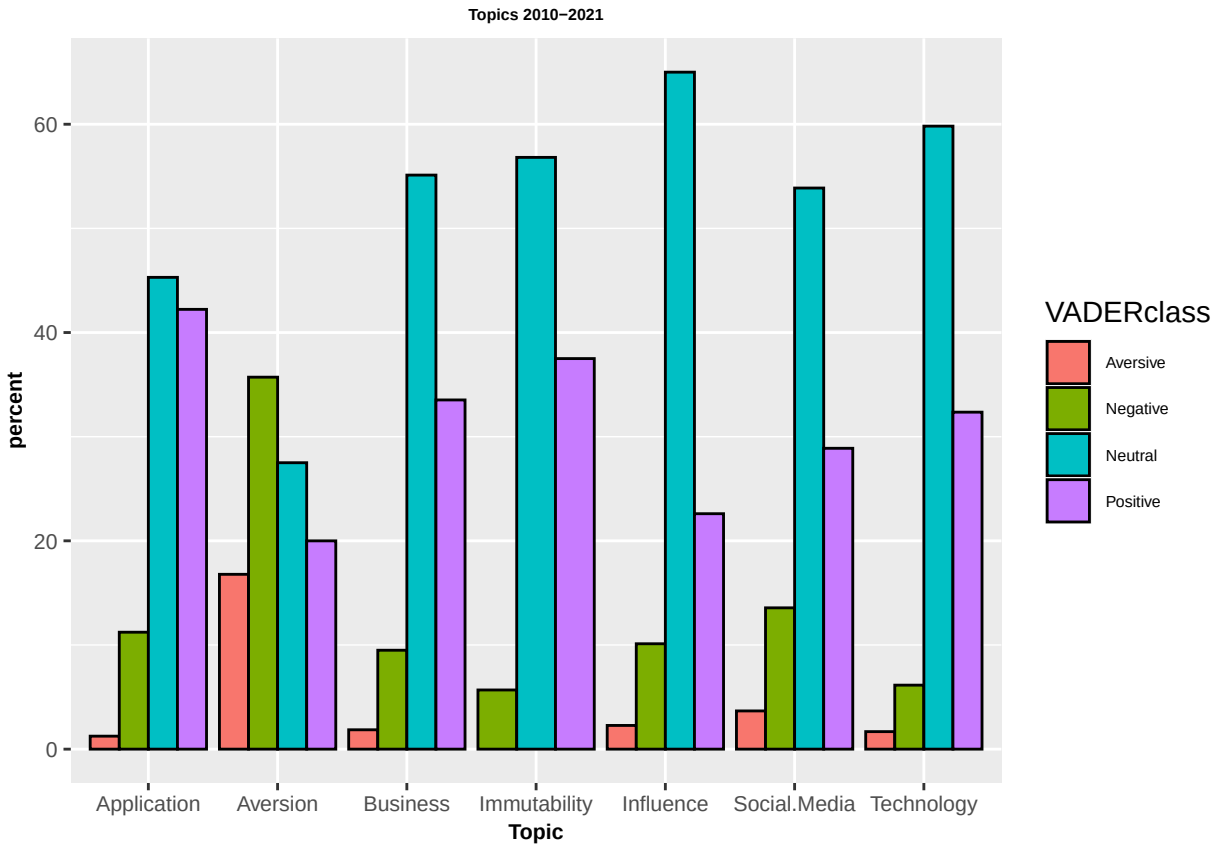
An increasing trend can be seen in the number of positive tweets. In 2010, there were 3827 positive tweets, representing 32.2% of the yearly total. In 2011, the least amount of tweets 3171 (26.7%) were posted. In 2019, the most positive tweets were posted with absolute 4878 (40.7%) posted. Last year, 2021, a total of 4868 positive tweets were posted, which is 40.6% of the annual total. The number of positive tweets remained high over he last 3 years, with a dip in 2020 (3908 total, 32.6%).

Negative tweets show an increasing trend. In 2010, there were 1008 negative tweets constituting 8.5% of the annual total. In 2020 the highest number of negative tweets were posted with absolute 1867 tweets (15.6%) posted. In 2015, the least amount of negative tweets was posted (918, 7.7%). In the last year analyzed, 2021, the number of negative tweets remained high with 1429 total, 11.9% of the annual.

The number of aversive tweets behaves similarly to that of negative tweets. 2011 113 0.95% there were the fewest were 2014 105 and 0.88% with the most there were 2020 968 8.1% and 2021 there were 3.7% and 449 tweets.

Distribution of classes by topic

This graph shows the different topics Business, Social Media, Technology, Immutability, Influence, Application and Aversion as well as the percentage distribution of the VADER classes Aversive, Negative, Neutral, and Positive.



The topic Aversion has the most aversive tweets with 16.8% and 35% negative tweets, so there are more negative tweets in this topic than positive (20%). The smallest amount occurs in Application; here, there are no aversive tweets at all. In the topic Social media, 28.9% of the tweets are positive, 13.6% negative and 3.7% aversive. In total, this was also the topic in which the most tweets occurred with a total of 42020.

Conclusion:

In this chapter, the answers to the following questions are explained:

1. Is there a significant aversion against algorithm on twitter and has it changed over time (last 10 years)?
2. Are there differences in algorithm aversion on twitter in dependence of influenceability, immutability and setting of application.

We did not find any significant aversion to algorithms on twitter. An increase in negative and aversive tweets can be observed.

For the categories application, immutability and influencability the following could be found: That for influencability there are no aversive tweets. The highest number of positive tweets was found for apllication. The categories business, social media, technology and aversion were also examined. It was found that in the

group aversion the most aversive and negative tweets were found. While this was to be expected, it also confirms that the sentiment analysis correctly assigns the tweets.

Since widespread usage of algorithms and research of public opinion about them in longitudinal analysis only emerged in recent years, not too much information about the chronological sequence in this area is available. Nevertheless our results show similarities to other studies with a longitudinal analysis of the public perception of technical and computational innovations: e.g. Artificial Intelligence (Fast and Horowitz, 2016) and Internet of Things (Zubiaga et. al. 2018). Both studies, as well as our study show an increased polarization of positive and negative sentiments towards these technologies. Neutral positions are declining and mostly positive attitudes and expectations are increasing. But characteristic for a polarization of opinions, the negative sentiments are also increasing - although to a much lesser degree. This implies that in generell the positive effects of algorithm utilization and technical innovations are increasingly acknowledged but also certain risks and obstacles are increasingly discerned by the broader public opinion.

Critique

Even though we could assess our research questions and could find differences on the one hand over the years and on the other hand between the assigned topics, one should consider that our findings are only descriptive nature and do not represent significant differences. This should be the next step going further into the data and analyze if the categories and years do really differ on a significant level. This could underline our findings and would make them more reliable. Also, it would be a possibility to check if the used words form the Merriam-Webster thesaurus are exhaustive and represent the assigned topic in the foreseen way. To do this one could either do more research and compare the used words with ones from other studies or could consult stakeholders e.g. from the topics Business, Social Media and Technology.

Apart from this one should consider that, as Dodds et al. (2015) could show, there is a positivity bias in the human language and because of that more tweets would be considered to be positive, even though the meant content would be a more negative. To correct the VADER classes towards the positivity bias one could check literature for a fixed value how much this positivity bias has an influence on tweets. This would make it possible to get closer to the true value of the tweet. If there is nothing like a corrective factor in the literature, it might be interesting to asses the influence of the positivity bias of the human language on tweets.

What also might be interesting to do is to compare the mean values of the last ten years to political or social incidents and if it correlates with the assed values. One of those incidents might be the Cambridge Analytica scandal in 2018, which seems possible because Hinds, Williams & Joinson (2020) reported after conducting interviews with 30 Facebook users that they have a simplified view and do tend to take a position and don't stay neutral on this topic. Findings like these could explain why less tweets got coded as neutral in the last years and that there was this bigger drop in 2019. All in all, our research gives a good first overview how algorithms are discussed and viewed on Twitter over time as well towards different topics. But still further research could dig deeper into it and might tie further connections to other research fields.

Biblography

Berger, B., Adam, M., Rühr, A. et al. (2021). Watch Me Improve - Algorithm Aversion and Demonstrating the Ability to Learn, Business & Information Systems Engineering 63(1):55–68

Burton, J. W., Stein, M-K., Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. Journal of Behaviour and Decision Making, 33, 220-239. https://doi.org/10.1002/bdm.2155

Castelo, N., Bos, M. W. & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. Journal of Marketing Research, 56(5), 809–825. https://doi.org/10.1177/0022243719851788

Dietvorst, B. J,; Simmons, J. P. & Massey, C. (2016). Overcoming Algorithm Aversion: People will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. Management Science, 64 (3), 1155-1170. http://dx.doi.org/10.1287/mnsc.2016.2643

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., . . . Danforth, C. M. (2015). Human language reveals a universal positivity bias. Proceedings of the National Academy of Sciences, 112 (8), 2389–2394. doi: 10.1073/pnas.1411678112

Fast, E., Horvitz (2016). Long-Term Trends in the Public Perception of Artificial Intelligence. AAAI 2017: 963-969. https://arxiv.org/abs/1609.04904

Hinds, J., Williams, E. J. & Joinson, A. N. (2020). "It wouldn't happen to me": Privacy concerns and perspectives following the cambridge analytica scandal. International Journal of Human-Computer Studies, 143. doi: https://doi.org/10.1016/j.ijhcs.2020.102498

Zubiaga A., Procter R., Maple C. (2018) A longitudinal analysis of the public perception of the opportunities and challenges of the Internet of Things. PLoS ONE 13(12): e0209472. https://doi.org/10.1371/journal.pone.0209472

Github Repository

https://github.com/sonnleit/AlgorithmAversion