Delft University of Technology

Pre-registration report

Values of predictions in hate speech detection

Authors:
Philippe Lammerts
Prof. dr. ir. G.J.P.M. Houben, thesis advisor
Dr. J. Yang, TU Delft, daily supervisor
Dr. Y-C. Hsu, TU Delft, co-daily supervisor
P. Lippmann, TU Delft, co-daily supervisor
September 16, 2022

Abstract

This document is prepared for the Human Research Ethics Committee review at TU Delft. It describes the plan for an experiment for the thesis project called "Building a smart rejector for detecting hate speech". This document follows the pre-registration plan suggested by [22] and elaborates on the goal, the exact procedure, and the analysis of the experiment.

In this experiment, we will conduct an anonymous survey where we ask human subjects to judge different scenarios of hate speech detection on social media. The goal of this experiment is to find out the value ratios between these scenarios. We will use the results in the thesis project to build a smart rejector for hate speech detection.

Contents

1 Research Question and Hypotheses

The amount of hateful content that is spread online on social media platforms remains a serious problem. Manual moderation is still the most reliable solution but is simply infeasible due to the large amount of data generated every second on social media platforms [2]. There exist automated solutions for detecting hate speech,

and most of these use Machine Learning models. However, these models tend to be unreliable as they often perform poor on deployment data [2, 8]. One study found that the F1 scores reduce significantly (69% F1 score drop in the worst case) when training a hate speech detection model on one dataset and evaluating it using another dataset [8].

Therefore in this project, we focus on Machine Learning (ML) models with a reject option. The goal of the reject option is to reject a prediction when the model is not confident enough [9]. This thesis project is about building a smart rejector for detecting hate speech. A system in which the machine assists the human in detecting hate speech automatically and in which the human makes the decisions when the machine is not confident enough.

The first goal of the project was to find a confidence metric that calculates the optimal rejection threshold. This threshold is calculated based on the confidence values of the ML predictions, the ground-truth labels of the underlying data samples, and a set of value ratios. These value ratios represent the relative impact of the True Positive (TP: correctly predicting content as hateful), True Negative (TN: correctly predicting content as non-hateful), False Positive (FP: predicting content as hateful while it is non-hateful), False Negative (FN: predicting content as non-hateful while it is hateful) and rejected predictions. Rejecting a prediction means that the ML predictor was not confident enough and, therefore, a human moderator needs to make the final judgment. We understand that there are different ways of handling rejected predictions. In this project, we treat rejected predictions as content that remains publicly visible online for, at most, 24 hours until a human moderator decides to remove/allow it. We based the timespan of 24 hours on a news article that stated that Germany fines social media platforms if they do not remove hate speech within 24 hours[1].

The second step of this project is to find out how we can retrieve the value ratios of TP, TN, FP, FN, and rejected predictions in hate speech detection. By value ratios, we mean to figure out, for example, the ratio between an FP and an FN prediction (e.g. FN predictions are X times worse than FP predictions). Expressing these values in money spent/saved is infeasible in hate speech detection due to many uncertainties [21]. So in this experiment, we aim to define the values in a subjective manner by analyzing the subjects' opinions about different hate speech detection scenarios. The main research question and our hypotheses of this experiment are as follows:

[RQ]: How can we determine the value ratios of rejections and True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions in hate speech detection?

- **We hypothesize that the negative value (cost) of an FN is smaller than the negative value of an FP**. We believe that allowing hateful content to be publicly visible has a more negative impact on social media users than filtering out neutral content. Therefore, we think that the value of an FN is smaller than the value of an FP.
- **We hypothesize that the positive value (gain) of an TP is greater than the gain of an TN.** We believe that correctly predicting hateful content is more valuable to social media users than correctly predicting non-hateful content. Therefore, we think that the gain of an TP is greater than the gain of an TN.
- **We hypothesize that the negative value (cost) of a rejection is greater than the average negative value of an FP and an FN prediction.** The key assumption of using ML models with a reject option is that the negative value of a rejection should always be greater than the negative value of an incorrect decision.
- **We hypothesize that Magnitude Estimation (ME) is a suitable technique for retrieving the value ratios.** For details on ME refer to 2.1.2. ME seems like a promising technique for retrieving ratio data from judgements about hate speech detection scenarios. We use a 100-level numerical scale for validation. We expect that both scales are correlated and will give similar judgements. Although we also expect the 100-level scale to be suitable for retrieving opinions about the different hate speech detection scenarios, it does not provide the ratio data we need. We also expect that the inter-rater reliability for the 100-level scale will be higher than for the ME scale since the ME scale provides more response freedom. We also expect this since the authors of [19] concluded that the inter-rater reliability of the 100-level scale is higher than the ME scale when rating the relevance of documents.

[1] https://www.nytimes.com/2017/06/30/business/germany-facebook-google-twitter.html

2 Method

In this experiment, we show different scenarios to the human subjects using a survey and ask them whether they agree, disagree or are neutral about the decisions of a fictional social media platform we call SocialNet. These scenarios represent TP, TN, FP, FN, and rejected predictions. For example, we simulate an FN by showing a hateful post to the subject and explain that SocialNet did not detect any hate speech. The subject then indicates whether they agree, disagree, or are neutral with this decision. By analyzing the results of a group of subjects, we aim to conclude that an FN scenario is considered, for example, three times as worse as an FP scenario. This section first gives some background information and then elaborates on the exact procedure of the experiment.

2.1 Background

Initially, we considered using Likert scales as the response scales of our survey. However, as we will explain in the following subsection, Likert scales are not suitable for retrieving ratio values between the different scenarios of TP, TN, FP, FN, and rejections. Therefore, we will use a technique called Magnitude Estimation in our survey.

2.1.1 Likert

Likert scales are widely used in academic research for retrieving the opinions of a group of subjects. Likert scales are a series of multiple Likert-type questions where subjects can answer questions with several response alternatives [5]. So in our case, we could use a bipolar scale with seven response alternatives ranging from 'strongly disagree' to 'strongly agree', including a 'neutral' midpoint. However, there is a lot of discussion in the literature about how we should analyze these Likert scales [5, 1, 17, 16]. The scale of the questions is ordinal, which means that we do know the ranking of the responses, but we do not have an exact measurement of the distances between the response items [1]. For example, we know that 'strongly agree' is higher in rank than 'agree', but not the exact distance between the two responses and whether it is greater than the distance between the 'neutral' and the 'somewhat agree' responses. Therefore, we technically cannot use parametric statistics, such as calculating the mean, when analyzing the data [1]. Other papers argue that we can treat a Likert scale that consists of multiple Likert items as interval data and, therefore, applying parametric statistics will not affect the conclusions [5, 17, 16]. So, we can calculate mean scores for TP, TN, FP, FN, and rejection scenarios and compare these with each other. For example, we can then verify that the mean value of FN cases is greater than the mean value of FP cases and conclude that the negative value of an FN is smaller than the negative value of an FP. Analyzing Likert scales from our surveys would at most provide us with interval data (data for which we know the order, and we can measure the distances, but there is no true zero point [1]). However, we need to have ratio data in this project since we want to know the exact value ratios of the TP, TN, FP, FN, and rejection scenarios.

2.1.2 Magnitude Estimation

In 2.1.1, we concluded that Likert scales are not suitable since they do not provide ratio data. In this research, we want to experiment with a technique called Magnitude Estimation (ME). The ME technique originates from psychophysicists where human subjects need to give a quantitative estimation of sensory magnitudes [20]. For example, in one experiment, human subjects are asked to assign any number that reflects their perception of the loudness of a range of sounds [20]. If the human subjects perceive the succeeding sound as twice as loud, they should assign a number to it that is twice as large. Researchers applied the ME technique to different types of physical stimuli (line length, brightness, duration, etc.) and proved that the results are reproducible and that the data has ratio properties [15]. Other works have shown that the ME technique is also useful for rating more abstract types of stimuli, such as judging the relevance of documents [13], the linguistic acceptability of sentences [3], the strength of political opinions [11, 12], and the usability of system interfaces [14]. Therefore, we think that ME is a promising method for judging the value ratios of hate speech detection scenarios.

The main advantage of ME is that it provides the ratio scale properties we need. Another advantage is that the scale is unbounded compared to other commonly used response scales, such as Likert scales. For example, suppose we first show a scenario and the subject provides a 'strongly disagree' judgment. Suppose we now present an even worse scenario. The subject is now limited to the response items in the Likert scale and can only give the same 'strongly disagree' judgement. We do not have this problem when using ME because the subject always has the freedom to assign a value of disagreement that is even larger. However, there are two

main drawbacks of using ME. First, we need to normalize the results. Second, it can be hard to validate if we can use ME to measure the subjects' judgements for different scenarios of hate speech detection.

The resulting data needs to be normalized since each subject can use any value they like. For example, one may judge the scenarios using values of 1, 2, and 10, while another may use 100, 200, and 1000. Luckily, there are different solutions for normalizing ME data, such as modulus normalization [15]. The most commonly used method for modulus normalization is geometric averaging since this preserves the ratio information [15, 14]. However, as opposed to the unipolar scales used in [3, 14], we are using bipolar scales (disagree-agree). By including 0 (neutral) and negative values (disagree), we cannot use geometric averaging anymore because it uses log calculations [15]. Using the algorithmic mean is also not an option since it would destroy the ratio scale properties [15]. Therefore, we normalize the magnitude estimates by dividing all estimates of each subject by the maximum given value. This way, all magnitudes estimates are in the range [100, 100] while maintaining the ratio properties.

Most papers that use the ME method apply some form of validation. Cross-modality validation is a technique that is often applied to validate the ME results [3]. Psychophysicists compare the magnitude estimates to the physical stimuli [3]. They analyze the correlation between the magnitude estimates and the physical stimuli by taking the log of each value and plotting them against each other [3]. In the case of estimating line lengths, we can easily vary the line length, for example, by showing a line that is twice as long as the previous line. Subjects can then estimate the line length using a number twice as large. However, this becomes more difficult in the social science and psychology domains. In hate speech detection and other applications in social science and psychology, we do not have an exact measure of the stimulus [3]. Luckily, related work has shown that ME is still a suitable technique for eliciting opinions about different types of non-physical stimuli [3, 14, 13, 11]. We can validate the magnitude estimates by adopting the cross-modality technique but instead compare judgements against judgements [3, 11]. Some papers analyze the correlation between different ME scales for validation, such as handgrip measurements or drawing lines [3, 12]. Others compare ME with another validated scale that can be of any type. For example, in [13] about judging the relevance of documents, the authors compared the ME scale with two validated ordinal scales for the same dataset [13]. So, we also need to validate our findings by checking the correlation between the ME scale and another scale. Validating two different measures is a form of convergent validation [6]. Refer to 4.3 for more details about validation.

2.1.3 100-level scale

In 2.1.2, we concluded that we need to validate the ME scale by comparing it against another. We will use a bounded scale that consists of 100 numerical levels to validate the ME scale for four reasons. First, it is impractical in this project, given the limited budget, to use other ME scales, such as measuring the intensity of the subjects' handgrips to express their judgements. Second, there does not exist any suitable dataset that we can use for validation that contains human ratings of different scenarios in hate speech detection. Third, we concluded in 2.1.1 that Likert scales have limited response freedom. Finally, in [19], the authors concluded that the 100-level scale provides more response freedom than course-grained Likert scales and has several advantages over ME in terms of usability and reliability. The 100-level scale is easier to understand than ME, does not require normalization, and provides more flexibility than Likert scales [19]. Therefore, we will create two separate surveys with the same scenarios where half of all subjects use the 100-level scale and the other half use the ME scale.

2.2 Design

We will list all independent, dependent, confounding, and control variables analyzed in our experiment in this section. Important to note is that we do not intend to study the confounding variables 'demographic' and 'age'. The main reason for this is that we do not focus on the effects of subject characteristics on the outcomes but only want to study whether the ME technique is suitable for retrieving opinions about hate speech detection predictions. We left the investigation of the effects of subject characteristics on the outcomes to future research. According to [7], there is no significant difference in how men and women perceive hate. Therefore, we do not consider gender as a confounding variable.

2.2.1 Independent variables

- **Scenarios**
  - **True Positive** Show a hateful post to the user and explain that SocialNet detected hate and ranked the post lower on people's feeds.

  - **True Negative** Show a non-hateful post to the user and explain that SocialNet did not detect hate and allowed the post.
  - **False Positive** Show a non-hateful post to the user and explain that SocialNet detected hate and ranked the post lower on people's feeds.
  - **False Negative** Show a hateful post to the users and explain that SocialNet did not detect hate and allowed the post.
  - **Rejection**
    * Show a hateful post to the user and explain that SocialNet was uncertain whether the post was hateful or not. An internal moderator will need to check the post within 24 hours. Meanwhile the post remains visible.
    * Or, show a non-hateful post to the user and explain that SocialNet was uncertain whether the post was hateful or not. An internal moderator will need to check the post within 24 hours. Meanwhile the post remains visible.
- **Social media posts**
  - Neutral content.
  - Hateful content.
    * Generic group target + aggressive
    * Generic group target + non-aggressive
    * Individual target + aggressive
    * Individual target + non-aggressive

2.2.2 Confounding variables

- **Demographic** People from different countries might have different perceptions and definitions of hate speech and how we should deal with it (tolerating hateful content or removing it).
- **Age** People of different ages might have different perceptions and definitions of hate speech and how we should deal with it (tolerating hateful content or removing it).

2.2.3 Control variables

- **Scales** The first group of subjects needs to answer the questions using the ME scale. The second group needs to answer the questions using the 100-level scale.
- **Content of the posts** All social media posts are sampled from existing datasets containing both hateful and non-hateful tweets.

2.2.4 Dependent variables

- **Reliability** Measured using the Krippendorff's alpha where values larger than 0.8 indicate reliable conclusions and values larger than 0.6 indicate tentative conclusions [10].
- **Validity** Convergent validity (if two different measures measure the same thing) [6]. Measured by calculating the correlation between the magnitude estimates and the response values from the 100 level scale.
- **Value ratios of TP, TN, FP, FN, and rejection scenarios** Measured by calculating the median of the normalized magnitude estimates of each scenario question and then calculating the mean over the resulting values to come up with the final value for that scenario type.

2.3 Planned sample

We will explain in this section how we will recruit the participants, give our sample size, and set our stopping and exclusion rules.

2.3.1 Participant Inclusion and Exclusion Criteria

We will use the Prolific platform for recruiting online participants. We will use the following inclusion criteria for our participants:

- 18 years of age and older since we are showing offensive language in the experiment.
- Fluent in English.
- Approval rating over 90% on the Prolific platform.
- Use one of the following social media platforms regularly (at least once a month): Facebook, Twitter, YouTube, LinkedIn, Pinterest, Google Plus, Tumblr, Instagram, Reddit, VK, Flickr, Vine.co, Meetup, ask.fm, Snapchat, TikTok, Medium.

We use the following exclusion/rejection criteria:

- Participants who fail the two attention checks. We will include two Instructional Manipulation Checks to check if the user pays attention to the survey[2].
- Participants who do not complete all questions.
- Participants who do not agree with the informed consent before the start of the survey. We are not allowed to collect and process their data if they do not consent.
- Participants who do fail the ME training phase, for example by providing random values that do not make any sense.

2.3.2 Participant Compensation

Every participant will be paid based on the hourly wage of 9.0 GBP (about 10,67 Euro), indicated as good pay by the platform[3].

2.3.3 Sample size

There are 4.55 billion active social media users[4]. We choose a 90% Confidence Interval (CI) and 10% Margin of Error (MoE) for this study. So for 90% of the time, our observations will fall within a 10% interval [18]. According to [18], we need a sample size of 68 participants per survey type to reach the desired CI and MoE values. We chose 10% MoE since we have a limiting budget. We will first conduct a pilot survey for 12 participants per scale to gather feedback and check if we need to improve things before the actual experiment. We want to determine the average workload using the pilot survey and decide whether it's possible to reduce the MoE by increasing the number of participants. For the pilot survey, we will use 24 participants. Therefore, in total we will need $2 * 12 + 2 * 68 = 160$ participants.

2.4 Materials

2.4.1 Survey tool

We use LimeSurvey as our survey tool since it supports all the features we need, and its (discounted) subscription price is 17 euros per month.

2.4.2 Data

We will use 40 different social media posts where 20 are hateful and 20 are not hateful to create 40 different scenarios. So, each subject will need to judge four scenarios for each type (TP, TN, FP, FN, and rejection) for each scale type (ME and 100-level) since we let each subject use both scales. We used the dataset from [4] that contains 13,000 English tweets. Each tweet is annotated in three categories: hate speech (yes/no), target (generic group or an individual), and aggressiveness (yes/no). Therefore, we have four different groups of hateful tweets: generic target + aggressive, individual target + aggressive, generic target + non-aggressive, and individual target + aggressive. We filtered out all tweets that contain replies and mentions since the context is not always clear in these messages. Then we preprocessed all tweets by removing the URLs and hashtags. We performed term frequency-inverse document frequency (TF-IDF), latent semantic analysis (LSA), and k-means

[2] https://researcher-help.prolific.co/hc/en-gb/articles/360009223553
[3] https://prolific.co/pricing
[4] https://datareportal.com/reports/digital-2021-october-global-statshot

clustering on each group of tweets. We calculated the silhouette coefficient to determine the optimal cluster size (k value) for the neutral tweets and the four groups of hateful tweets. The silhouette analysis indicated to set k as large as possible. We selected the tweets by taking the nearest data sample to each cluster centroid. We used a cluster size of 20 for the neutral tweets and sampled one tweet per cluster to collect 20 neutral tweets. And we used a cluster size of 5 for each group of hateful tweets to collect 20 hateful tweets.

3 Procedure

This section will explain all steps of the survey. Appendix A contains all presentation texts, the informed consent, and some scenario examples. In this experiment, we will shuffle and present the TP, TN, FP, FN, and rejection settings to the subjects. We include one additional question in each scenario after showing the social media post in which we ask the subjects whether they think the post is hateful or not. This question is binary and can be answered with yes or no. The results of these questions are purely informative and used to understand the results of the final question about the agreement with SocialNet's decision. There is no limit to the amount of time they spend on each question. All the collected data is completely anonymous. Finally, we will inform the participants not to put personal identifiers in their answers.

Step 1: provide informed consent

- Show the informed consent (with checkboxes for giving consent).
- Proceed to the next step only for the participants who give consent.

Step 2: introduction

- Show introductory text about what is expected from the subject.
- We split all participants up into two groups.
- The first group first uses the ME scale to rate all scenarios.
- The second group first uses the 100-level scale to rate all scenarios.
- Provide an explanation in the introduction about the scale.

Step 3: two attention checks

- Two simple attention checks where we ask the subject to select one option (shuffled through all scenarios).

Step 4a: ME practice phase (when ME is used)

- To let subjects learn how to use ME, we first run a practice phase where we shuffle and present 5 different line lengths.
- Each subject needs to estimate the line length using any positive value.

Step 4b: all scenarios using the scale (ME or 100-level)

- Show 40 different scenarios in random order: 8 TP, 8 TN, 8 FP, 8 FN, and 8 rejection.

Step 5: finish

- Show a thank you message and redirect the users to Prolific to complete the task.

4 Analysis

First, we calculate the value ratios of the TP, TN, FP, FN, and rejection scenarios in hate speech detection using the survey's results. Second, we analyze whether we can use the results to draw any conclusions by looking at two aspects: reliability and validity.

4.1 Value ratios

The goal of the complete experiment is to come up with the value ratios of the TP, TN, FP, FN, and rejection scenarios in the context of hate speech detection. The metric from the first part of our research takes these numerical values as its input to calculate the optimal rejection threshold. We do not need to know the absolute values but only the relative values. For example, if we set all values to 1, we retrieve the same optimal rejection

threshold as setting all values to 1000. Therefore, we need to know the ratios between all scenario types. The ME technique provides us with ratio data. We use a bipolar scale for question 3 in the survey since we ask the subjects whether they agree, disagree, or are neutral with the decision of SocialNet (tolerating, removing, or rejecting posts). For both scales, we will convert disagreement values to negative values, neutral values to 0, and agreement values to positive values. Since the data of both scales is skewed, we apply the median to the results of the individual questions first. Then we calculate the mean value over the resulting values to retrieve the final aggregated value ratios. The results should give us an understanding of how the subjects feel towards the different scenarios: TP, TN, FP, FP, and rejection.

For example, to calculate the aggregated V_{TP} values for both scales, we use:

$$V_{TP}^{ME} = \frac{1}{n}\sum_{i=1}^{n} r_i^{ME}$$ where n is the total number of TP scenarios and $r_{i,TP}^{ME}$ is the median response value of ME question number i.

$$V_{TP}^{100} = \frac{1}{n}\sum_{i=1}^{n} r_i^{100}$$ where n is the total number of TP scenarios and $r_{i,TP}^{100}$ is the median response value of 100-level question number i.

We apply the same calculations for the remaining scenario types. We define the values we need for the metric using the aggregated values of the TP, TN, FP, FN, and rejection scenarios rated with the ME scale. We will interpret disagreement values as negative values (costs) and agreement values as positive values (gains). We will not use the results of the 100-level scale in our metric since the 100-level scale does not have ratio properties.

4.2 Reliability

Reliability is about whether we can trust our results and if we get consistent results [6]. We do this by mainly looking at the inter-rater reliability. This means that different subjects should give approximately the same judgements to the same scenarios. We measure the inter-rater reliability using Krippendorff's alpha [13, 10]. We calculate the inter-rater reliability value using the mean response values as calculated in 4.1. We will use the inter-rater reliability scores to compare the ME scale with the 100-level scale. We also study the inter-rater reliability values for the different types of scenarios: TP, TN, FN, FP, and rejection. Other types of reliability, such as test-retest reliability, are not considered in this experiment. Guaranteeing test-retest reliability would require us to redo the complete experiment at a different time for the same subjects. This is infeasible for our thesis project, given the limited time and budget.

4.3 Validity

Validity is about whether we are measuring the things we want to measure [6]. The main goal of this aspect is to validate if we can use the ME technique to measure subjects' opinions about hate speech detection scenarios. There are multiple types of validity, but we focus mainly on convergent validity (part of construct validity), content validity, and face validity [6]. Construct validity checks whether there is an agreement between a theory and a measurement device or procedure [6]. Convergent validity is about the correlation between different types of measures to see if they measure the same phenomenon [6]. Content validity is about letting external experts review the proposed research questions and procedure [6]. Face validity is the subjective type of validity, and it is about why we think the questions and proposed procedures are valid [6].

We analyze convergent validity by comparing the mean scores from 4.1 between the two scales. We can verify that they measure the same phenomenon by analyzing the correlation between the scales. However, we can expect a low correlation since the ME scale is a (normalized) unbounded scale and the 100-level scale is bounded. Nevertheless, we think that both scales will give similar results, meaning that high ME responses should correspond to high 100-level scale responses and low ME responses to low 100-level scale responses. To guarantee content validity, we let external experts (the supervisors of this thesis project) check this pre-registration report. We guarantee face validity by discussing whether the ME technique gives us the expected results. We exclude other forms of validity from this experiment because they either are irrelevant or infeasible. For example, external validity is about the degree to which the findings can be generalized to other settings or groups. We would have to experiment with multiple groups with different demographic and age characteristics to guarantee external validity. We left this for future work to investigate since this is out of scope for our thesis project. Despite that, we think that people of different ages and demographic characteristics perceive hate differently since people have other norms and values in various parts of the world. We believe that if we conduct this experiment using different groups of subjects, then we might retrieve different value ratios. Therefore, we

decided not to create too many participant inclusion criteria but take a random sample of global social media users.

A Presentation texts

A.1 Consent

You are being invited to participate in a research study titled "Costs of predictions in hate speech detection". This study is being done by Philippe Lammerts from the TU Delft.

The purpose of this research study is to find out what social media users think of different scenarios of hate speech detection on social media. It will take you approximately 22 minutes to complete. These scenarios consist of two things. First, we show a specific social media post that can be either hateful or not hateful. You need to indicate if you feel that the post is hateful or not. Second, we explain how the social media platform dealt with this post. You need to indicate whether you agree/disagree/are neutral about the platform's decision. The results of the survey will be used in my thesis.

As with any online activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by making this survey completely anonymous. Therefore, please do not provide any personal information anywhere. The anonymous results might be shared publicly in the future.

Your participation in this study is entirely voluntary, and you can withdraw at any time.

Warning: some of the scenarios used in this experiment contain harmful and offensive content that may make some people feel uncomfortable.

Feel free to contact me with any questions or feedback you might have: p.m.lammerts@student.tudelft.nl

A.2 Short introduction ME

- You will be presented with a series of different scenarios.
- For each scenario, you need to answer two questions.
- We will explain the exact instructions later.
- But first, we will let you familiarize yourself with a scale called Magnitude Estimation.

A.3 Short introduction 100

- You will be presented with a series of different scenarios.
- For each scenario, you need to answer two questions.
- We will explain the exact instructions in the next page.

A.4 Introduction

You will be presented with a series of different scenarios

- Each scenario describes a situation of a social media user who wants to post a specific message on a fictional social media platform we now call SocialNet.
- These posts can be neutral or contain hateful content.
- SocialNet uses automated detection systems for detecting hate speech.
- When doing the study, you should be aware that it is expected for SocialNet to correctly classify hate speech. Wrong classifications are undesirable as they may cause harm to people.

Each scenario describes one of the following situations for a specific social media post:

1. **You are a user of the SocialNet platform and have not seen this post on your main feed because SocialNet's automated detection system is confident that it is hateful.**
    - You can still find this post when you scroll down your feed since SocialNet ranks hateful posts lower.

- If the post is not hateful after all, then the detection system was incorrect. This neutral post is now ranked lower on people's feeds with the consequence that the post cannot easily reach the author's followers.
- If the post is indeed hateful, then the detection system was correct.

2. **You are a user of the SocialNet platform and just saw this post on your main feed because SocialNet's automated detection system is confident that it is not hateful.**
   - This post remains visible on other people's main feeds as well.
   - If the post is hateful after all, then the detection system was incorrect. This hateful post is now visible on people's main feeds with the consequence that they can get harmed.
   - If the post is indeed not hateful, then the detection system was correct.
3. **You are a user of the SocialNet platform and just saw this post on your main feed because SocialNet's automated detection system was not confident enough in whether it was hateful or not.**
   - An internal human moderator at SocialNet needs to look at it within at most 24 hours.
   - Meanwhile, the post remains visible on people's main feeds.

A.5 100-level scale explanation

For each scenario, you need to answer two questions:

1. First, you need to indicate whether you feel that this post is hateful or not hateful.
2. Second, your task is to tell how you feel about SocialNet's decision.
   - If you feel neutral about SocialNet's decision, this value will be equal to 0.
   - If you (dis)agree with the decision, you need to indicate how much you (dis)agree by assigning any number between 1 and 100.
   - A large number means you (dis)agree with it a lot, while a small number means you (dis)agree with it a little.
   - Try to make each number match the intensity as you perceive it.

Don't worry, we will provide the same explanations in the questions as well.

A.6 ME scale explanation (inspired by [15])

For each scenario, you need to answer two questions:

1. First, you need to indicate whether you feel that this post is hateful or not hateful.
2. Second, your task is to tell how you feel about SocialNet's decision.
   - If you feel neutral about SocialNet's decision, this value will be equal to 0.
   - If you (dis)agree with the decision from SocialNet, you need to assign any number that is greater or equal to 0 that reflects how much you (dis)agree with the decision.
   - Assign any number that seems appropriate to you.
   - A large number means you (dis)agree a lot, while a small number means you (dis)agree a little.
   - If you (dis)agree twice as much with the current decision as with the previous one, you need to assign a number that is twice as large as the previous number.
   - Or, if you (dis)agree half as much with the current decision as with the previous one, you need to assign a number that is half as large as the previous number.
   - You can use any number or decimal you want, but make each assignment proportional to your subjective impression.

- Please try to not only use appropriate numbers but also avoid restricting your choice of numbers from 1 to 10.
- Try to remember which values you used before and to compare succeeding scenarios with each other.

Don't worry, we will provide the same explanations in the questions as well.

A.7 Attention check #1

This question is an attention check. You must select 'Blue' here.[5]

A.8 Attention check #2

This question is an attention check. You must select 'Orange' here.

A.9 Training phase ME

"As a warm-up task, to familiarize you with magnitude estimation, you will be shown a sequence of five lines, one at a time.

- For each line, enter a number into the text box below the displayed line. This number should reflect your perception of the length of the line. You may use any numbers that seem appropriate to youwhole numbers or decimals. However, you may not use negative numbers or zero.
- For each subsequent line, enter a number that reflects your perception of its length, relative to the previous line. For example, if you feel that the current line is twice as long as the previous, then you should assign a number that is twice as large as the number you used previously.

Don't worry about running out of numbersthere will always be a larger number than the largest you use, and a smaller number than the smallest you use. Note: The magnitude estimation scores are **not** intended to be an estimate of the length in any particular measurement units, such as centimeters." [13]

A.10 Example FN scenario with ME scale

WARNING: the example used in this section contain content that may make some people feel uncomfortable.

Suppose we have the following post:

"I want this stupid bitch out of my head I've never even liked her lmao I just basically want to rape her"[4]

Question 1:
Please indicate whether you feel that this post is hateful or not hateful.

Select input with options: 'hateful' or 'not hateful'

Question 2:
You are a user of the SocialNet platform and just saw this post on your main feed because SocialNet's automated detection system is confident that it is not hateful.

- This post remains visible on other people's main feeds as well.
- If the post is hateful after all, then the detection system was incorrect. This hateful post is now visible on people's main feeds with the consequence that they can get harmed.
- If the post is indeed not hateful, then the detection system was correct.
- You should be aware that it is expected for SocialNet to correctly classify hate speech. Wrong classifications are undesirable as they may cause harm to people.

[5]Based on https://researcher-help.prolific.co/hc/en-gb/articles/360009223553

Please indicate whether you agree, disagree, or are neutral about SocialNet's decision.

Select input with options: 'agree', 'neutral', or 'disagree'.

Question 3 (only visible when 'agree' or 'disagree' are selected):
Please indicate how much you agree/disagree with SocialNet's decision using any positive number.

Show ME input field that allows all positive values greater than 0.

A.11 Example FP scenario with 100-level scale

Suppose we have the following post:

"Ontario concerns were not listened at immigration meeting in Ottawa For illegal crossing"[4]

Question 1:
Please indicate whether you feel that this post is hateful or not hateful.

Select input with options: 'hateful' or 'not hateful'

Question 2:
You are a user of the SocialNet platform and have not seen this post on your main feed because SocialNet's automated detection system is confident that it is hateful.

- You can still find this post when you scroll down your feed since SocialNet ranks hateful posts lower.
- If the post is not hateful after all, then the detection system was incorrect. This neutral post is now ranked lower on people's feeds with the consequence that the post cannot easily reach the author's followers.
- If the post is indeed hateful, then the detection system was correct.
- You should be aware that it is expected for SocialNet to correctly classify hate speech. Wrong classifications are undesirable as they may cause harm to people.

Please indicate whether you agree, disagree, or are neutral about SocialNet's decision.

Select input with options: 'agree', 'neutral', or 'disagree'.

Question 3 (only visible when 'agree' or 'disagree' are selected):
Please indicate how much you agree/disagree with SocialNet's decision using any positive number from 1 to 100. If you feel neutral about SocialNet's decision, select neutral in the field above.

Show a numerical slider with values between 1 and 100.

A.12 Example rejection scenario with 100-level scale

Suppose we have the following post:

"Ever been so hungover that your stomach feels like it's eating itself"[4]

Question 1:
Please indicate whether you feel that this post is hateful or not hateful.

Select input with options: 'hateful' or 'not hateful'

Question 2:
You are a user of the SocialNet platform and just saw this post on your main feed because SocialNet's automated detection system was not confident enough in whether it was hateful or not.

- An internal human moderator at SocialNet needs to look at it within at most 24 hours.
- Meanwhile, the post remains visible on people's main feeds.

Please indicate whether you agree, disagree, or are neutral about SocialNet's decision.

Select input with options: 'agree', 'neutral', or 'disagree'.

Question 3 (only visible when 'agree' or 'disagree' are selected):
Please indicate how much you agree/disagree with SocialNet's decision using any positive number.

Show a numerical slider with values between 1 and 100.

References

[1] I. E. Allen and C. A. Seaman. Likert scales and data analyses. *Quality progress*, 40(7):64–65, 2007.

[2] A. Balayn, J. Yang, Z. Szlavik, and A. Bozzon. Automatic identification of harmful, aggressive, abusive, and offensive language on the web: A survey of technical biases informed by psychology literature. *ACM Transactions on Social Computing (TSC)*, 4(3):1–56, 2021.

[3] E. G. Bard, D. Robertson, and A. Sorace. Magnitude estimation of linguistic acceptability. *Language*, 72 (1):32–68, 1996.

[4] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics, 2019.

[5] H. N. Boone and D. A. Boone. Analyzing likert data. *Journal of extension*, 50(2):1–5, 2012.

[6] K. Fitzner. Reliability and validity a quick review. *The Diabetes Educator*, 33(5):775–780, 2007.

[7] M. W. T. H. D. Gold and T. Zesch. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. 2018.

[8] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan. All you need is "love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*, pages 2–12, 2018.

[9] K. Hendrickx, L. Perini, D. Van der Plas, W. Meert, and J. Davis. Machine learning with a reject option: A survey. *arXiv preprint arXiv:2107.11277*, 2021.

[10] K. Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433, 2004.

[11] M. Lodge and B. Tursky. Comparisons between category and magnitude scaling of political opinion employing src/cps items. *American Political Science Review*, 73(1):50–66, 1979.

[12] M. Lodge, J. Tanenhaus, D. Cross, B. Tursky, M. A. Foley, and H. Foley. The calibration and cross-modal validation of ratio scales of political opinion in survey research. *Social Science Research*, 5(4):325–347, 1976.

[13] E. Maddalena, S. Mizzaro, F. Scholer, and A. Turpin. On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Transactions on Information Systems (TOIS)*, 35(3):1–32, 2017.

[14] M. McGee. Master usability scaling: magnitude estimation and master scaling applied to usability measurement. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 335–342, 2004.

[15] H. R. Moskowitz. Magnitude estimation: notes on what, how, when, and why to use it. *Journal of Food Quality*, 1(3):195–227, 1977.

[16] J. Murray. Likert data: what to use, parametric or non-parametric? *International Journal of Business and Social Science*, 4(11), 2013.

[17] G. Norman. Likert scales, levels of measurement and the laws of statistics. *Advances in health sciences education*, 15(5):625–632, 2010.

[18] J. S. Olson and W. A. Kellogg. *Ways of Knowing in HCI*, volume 2. Springer, 2014.

[19] K. Roitero, E. Maddalena, G. Demartini, and S. Mizzaro. On fine-grained relevance scales. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 675–684, 2018.

[20] S. S. Stevens. The direct estimation of sensory magnitudes: Loudness. *The American journal of psychology*, 69(1):1–25, 1956.

[21] C. R. Sunstein. Does the clear and present danger test survive cost-benefit analysis? *Cornell Law Review*, 104:1775, Nov 2019.

[22] A. E. Van't Veer and R. Giner-Sorolla. Pre-registration in social psychologya discussion and suggested template. *Journal of experimental social psychology*, 67:2–12, 2016.