Politecnico di Torino

ISI

ISI Foundation & ISI Global Science Foundation

POLITECNICO DI TORINO

Master degree in Physics of Complex Systems

Modeling the boundaries of social norms online

Supervisor:

Dr. Francesco Bonchi
ISI Foundation

Co-Supervisor:

Prof. Luca Dall'Asta
Politecnico di Torino - DISAT

Candidate:

Sara De Candia

Academic year 2020/2021

Contents

Abstract

The topic of moral judgment and social norms has been widely addressed to understand human behaviour. Since nowadays the majority of interpersonal interactions happen on online platforms, researchers need to examine the new social spaces constituted by social platforms. Social media can be in fact precious data sources: together with a sociological interpretation of the results, they can provide new insights into the understanding of the perception of social norms. This thesis investigates social norms in the context of Reddit: a social news aggregation, web content rating, and discussion website. The Reddit community of interest for the present study is r/AITA, a community dedicated to asking and providing feedbacks about social behaviour. Given this rich dataset, we aim to investigate the determinants of social norms as expressed online, and in particular to explore their biases and boundaries when controversial subjects are addressed. What has been said so far brings us to two main research questions: what are the factors that push the collective judgement in one direction or the other? And, on the flip side, is it possible to model the response of a single member of the community given their history on Reddit? To this aim, we retrieved the data of interest from the social network and extracted the age and the gender of the poster through a regular expression and the tags given in the several comments. Then, the analysis has been performed using statistical tests and prediction algorithms. Our results show that the online perception of social norms is influenced by the age and the gender of the poster, showing a harsher judgement towards male users and people of 21-23 and 23-70. Furthermore, the judgment is also influenced by the topic of the submission and by the previous history of the user, confirming the role played by cultural and social background in the outlining of social norms and moral judging.

Chapter 1

Introduction

The topic of moral judgment and social norms has been widely addressed to understand human behaviour. Since nowadays the majority of interpersonal interactions happen on online platforms, researchers need to take into consideration the new social spaces constituted by social platforms. To this purpose, social media can be modelled as complex networks and studied using statistical analysis and machine learning algorithms. The usage of these techniques together with a sociological interpretation of the outputs can provide new insights in the understanding of the perception of social norms. This thesis investigates the previously mentioned topic in the context of the social media Reddit: a social news aggregation, web content rating, and discussion website. Posts are organized by subject into user-created boards called "communities" or "subreddits", which cover a variety of topics. Reddit ranks among the ten most popular social media sites in the United States, and it appears to be more popular among males: 63.2% of its users identify as males. The site is the most popular among users in the 25 to 29 age group which brings to the conclusion that the majority of users are young adults [24]. Even if Reddit is considered as a social media, unlike Facebook and Instagram it focuses less on interpersonal interactions (e.g., posting a photo for your friends and acquaintances) and more on dicussions about certain topics. The relevance of a particular discussion is decided by the users through a system (given by the social network) of downvotes and upvotes: the most upvoted threads appear on the homepage of the site. The Reddit community of interest for the present study is r/AITA, which is a community dedicated to asking for feedbacks about social behaviour. A user describes a situation and how they managed it and asks the community if their behaviour was morally acceptable or not. The users of the community answer with judgements, according to their perceived social norms, encoded as tags: YTA or ESH if they agree and NTA or NAH if they do not. Given this rich dataset, we aim to investigate the determinants of social norms as expressed online, and in particular to explore their boundaries when controversial subjects are addressed. What has been said so far brings us to two main research questions: what are the factors that push the collective judgement in one direction or the other? And, on the flip side, is it possible to model the response of a single member of the com-

munity given their history on Reddit? For this purpose, the work has been organized in the following way: in the beginning, the data has been collected from the social network and it has been processed and labelled in order to perform the analysis. Then, the work has been performed on the final dataset by the use of statistical tests like Chi-Squared, Binomial Regression and Multinomial Regression to investigate the correlations between the judgments given to the author of the submission and his/her age and gender. The statistical tests have also been performed on a smaller dataset which has been labelled manually with the main topic of the submission (Family, Friendships, Work, Society, Relationships). Furthermore, the investigation on the correlation between the previous history of the author of the comment and the judgment he/she gives has been carried on by means of a Random Forest classifier combined with different techniques of feature selection.

Chapter 2

State of the art and related work

The questions about moral judgment and social norms have been changing in recent years as social media has revolutionized the way people communicate and share information. Users are constituting new social spaces where the social norms can be perceived in a different way with respect to the real life. Social networks allow users from different part of the world, different culture, different age and different gender to interact as never before. This new scenario opens new research questions that have been investigated by psychologists, scientists and sociologists. Social norms are fundamental to human behavior [7, 1]. Former literature defines norms as statements "that something ought or ought not to be the case" [5], as institutionalized role expectations [11], or as becoming apparent if behavior attracts punishments [17]. In general, norms are mental representations of appropriate behavior in society and smaller groups and, consequently, guide the behavior of individuals. Norms that are characterized as social "must be shared by other people and partly sustained by their approval and disapproval" [1],[15]. Online debates about various topics give researchers the opportunity to deepen the topics of disagreement in groups, moral evaluation and judgment. The work by Yardi [8] about group polarization on Twitter highlights fundamental issues in designing socio-technical systems. First, people should engage in the exchange of ideas and views among a diverse group. This can be facilitated through cross-linking between ideologically competing groups; this can also limit isolation and social enclaves. Competing views, including within like-minded groups, should also be promoted. While not all views need to be endorsed within a group, it is important that no single majority view dominates such that members of the group are unable to promote and discuss other ideas. Voting and ranking algorithms can help control this balance. Finally, diversity of viewpoints may well be best promoted by encouraging members from diverse racial, social, and educational backgrounds to participation in discussions. As more and broader demographics use the Internet, from elderly users to rural users, there are opportunities to engage people in more diverse discussions than they did before. An important aspect to prevent is the derailing of online discussions into toxic exchanges between participants, recent studies analyze how it is possible to detect these antisocial behaviors by analyzing single

comments in isolation. The focus is the flow of the discussion, rather than properties of individual comments. Furthermore, a conversation can end or derail at any time (i.e. it has an unknown horizon). In the work by Chang and Danescu-Niculescu-Mizil a forecasting model has been implemented that learns an unsupervised representation of conversational dynamics and exploits it to predict future derailment as the conversation develops [18]. The derailment can lead to online firestorms in which it has been seen that non-anonymous individuals are more aggressive compared to anonymous individuals [15]. An expression of moral judgment is moral outrage (in our study it will coincide with the assignment of a negative judgment): moral outrage is a powerful emotion that motivates people to shame and punish wrongdoers, it can have positive outcomes like correcting negative behaviors but in the digital age the consequences can escalate into destructive feuds. The work by Crockett [16] highlights how the digital media may exacerbate the expression of moral outrage by inflating its triggering stimuli. People are more likely to be exposed to immoral acts online than in person and above all there is a higher probability of expressing moral outrage. Shaming a stranger on a street is more risky than doing it anonymously on a social network among a crowd of thousand of users. Anonymity online can play a double role: participants are under an equal condition independently from their backgrounds but, at the same time, anonymity can lead to counter social norms [9]. Crockett concludes in its work that digital media may promote the expression of moral outrage by magnifying its triggers, reducing its personal costs and amplifying its personal benefits [16]. At the same time, online social networks may diminish the social benefits of outrage by reducing the likelihood that norm enforcing messages reach their targets, and could even impose new social costs by increasing polarization. Another aspect of interest that is needed to analyze the topic of moral judgment online is the anonymity. Since posters and commenters are anonymous on Reddit (most of the times posters use throwaway accounts, i.e. accounts that are created and used only in that specific occasion) they have no barriers in sharing their sensitive stories or in expressing their moral opinion. Computational text analysis methods have been implied to understand moral sentiment in text. The work by Sagi and Dehghani [12] uses the Moral Foundation Dictionary (MFD) for the purpose of text analysis. The MFD consists of 295 words and word stems related to each of the moral intuitions of harm, fairness, authority, loyalty to in-group, and purity. The concept of morality used in the previous mentioned studies is based upon Moral Foundations Theory [10] that specifies five criteria for determining what should be considered a foundation of human morality. The five pillars can be summarized as:

1. Care/harm: Prompted by concerns about caring and protecting individuals from harm.

2. Fairness/cheating: Concerns triggered by acts of cooperation, reciprocity, and cheating.

3. Loyalty/betrayal: Related to virtues of patriotism, self-sacrifice and loyalty, and the vice of betrayal, unfaithfulness and disloyalty to the group.

4. Authority/subversion: Prompted by concerns about obedience, respect, insubordination, or subversion for authority.

5. Purity/degradation: Related to the emotion of disgust and triggered by practices related to sanctity, degradation, and pollution

Furthermore, the complex topic of human interactions on social media needs to be investigated also in terms of persuasion dynamics: this has been done in the work by Dutta and Das [19] that developed a model to extract argumentative components from discussion threads in order to study how people engage in argumentation on online discussion forums. A deeper investigation on the perception of social norms online has been done by Forbes et al. [21]: their starting dataset merged several sources of moral content such as two subreddits (including the one on which our work is based, r/AITA) and others. Their work consisted in developing a framework that can provide a new resource to teach AI models to learn people's norms, as well as to support novel interdisciplinary research across NLP, computational norms, and descriptive ethics. On this path, it is necessary to mention also the work done by Lourie et al. [25] which focused on predicting the moral outcome of a situation. Their dataset covered both real-life anecdotes with normative judgments and simple, ethical dilemmas (also in this case, the extraction has been made from r/AITA). Moving on with the subject another interesting work has been done by Botzer et al. right on the subreddit of interest of our work, r/AITA [23]. In their work they investigated the topic of moral outrage on Reddit, examining the subreddit r/AITA and others subreddits correlated to this latter. They developed a prediction model based on the dataset containing textual posts labeled with positive or negative moral judgments and used it to predict the moral content of the comments of other subreddits. Their results demonstrated that users prefer posts that have a positive moral valence rather than a negative moral valence. Furthermore, they demonstrated that age and gender have a minimal effect on whether a user is judged to be have positive or negative moral valence. On the other side, the work done by [22] highlights how two identical transgressive situations will be judged differentially based on the gender of the parties implicated. We can conclude that the topic of moral judgment and its prediction is an important topic to investigate, in view of approaching future works where the role of social norms in the context of AI will be crucial.

Chapter 3

Data collection and analysis

In this section I will describe the collection of datasets and show their properties. The data have been downloaded from Reddit by using the Pushshift Reddit Dataset, which collects all the submissions and comments posted on Reddit between June 2005 and April 2019. The total dataset consists of 651 778 198 submissions and 6 601 331 385 comments. The activity on Reddit has increased substantially over the years and by the end of the dataset the number of comments per day is 5M. The data in the Pushshift Reddit are divided into two sets: submissions and comments. For what concerns the structure of the two types of file they are both a collection of newline delimited JSON files. Each separate file is a month and each line in the file corresponds to a JSON object [20].

Not all of the fields included in each JSON object are useful for the analysis so the extracted ones are listed below:

- `id`: Submission's identifier (e.g.,"5lcgjh")
- `author`: Account name of the poster
- `subreddit`: Name of the subreddit in which the submission has been made
- `created_utc`: UNIX timestamp of the creation of the submission
- `score`: Score that the submission has accumulated, the score can be thought as the number of upvotes minus the number of downvotes
- `title`: Title associated with the submission
- `self text`: Text of the submission

While for what concerns the comments the fields of interest are:

- `id`: Comment's identifier
- `parent_id`: Identifier of the submission that this comment is in (e.g.,"t3_5lcgjh", so the id of the submission with t3_ in addition)

- `link_id`: Identifier of the parent of this comment, might be the identifier of the submission if it is a top-level comment or the identifier of another comment
- `subreddit`: Name of the subreddit in which the comment has been made
- `created_utc`: UNIX timestamp that refers to the submission's creation
- `author`: Account name of the poster of the comment
- `score`: Score that the submission has accumulated, the score is evaluated as the number of upvotes minus the number of downvotes
- `body`: Text of the comment

Because of this intrinsic difference of the two set of data, a different type of work has been made on each one, since they provided different information:

- **Submission**: the main information present in the submission are in the first place the age and gender of the author (which is often specified in the title or in the body of the submission) and the topic of the submission which in this analysis, after a deeper investigation of the subreddit and of the covered topics, have been divided manually into five categories: family, friendships, relationships, work and society. We have done this division for the purpose of the analysis based on the fact that the subreddit is a "private" one. These five categories are the ones that emerge after an investigation of the subreddit.
- **Comment**: the comment is where the post is judged, as mentioned before there can be four types of judgments and they are expressed in tags, which are abbreviations of the opinion of the user. The comments can contain the tag alone or the tag followed by a reasoning of the judgment. The comments can also contain no tag at all. In the following table a summary of the tags is presented:

Tag (acronym)	Tag	Meaning
YTA	You are The Asshole	The user thinks that the author of the submission is guilty in the mentioned situation.
NTA	Not The Asshole	The user thinks that the author of the submission is not guilty in the situation.
NAH	No Assholes Here	The user thinks that nobody in the situation is guilty.
ESH	Everybody iS the Asshole	The user thinks that everybody in the situation is guilty.
NFO	Not enough inFO	The user thinks that the author did not gave enough information on the situation.

3.1 Data collection

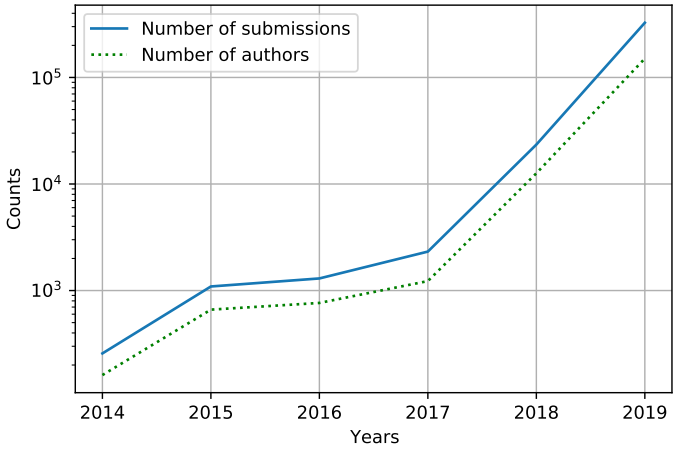
Starting from the totality of the data the first task has been that of filtering all the comments and submissions regarding our subreddit of interest: AITA (which stands for `\AmItheAsshole`). In order to do this, during the extraction of data from both sets, the field "subreddit" has been filtered by selecting only the submissions and the comments that are present in that specific subreddit.

3.1.1 Submission extraction and labeling

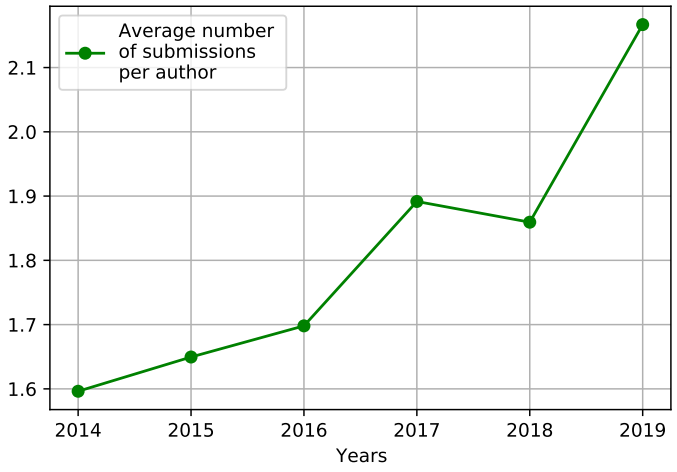
The submissions have all been saved in a unique JSON file which starts in February 2014 (date of birth of the subreddit) and ends in December 2019. The extraction of the huge amount of data has been made with Apache Spark. Once the final JSON has been obtained the following work has been done in order to extract the information needed for the study.

Submissions' authors

A first analysis is done by evaluating the average number of submissions per author, which results to be $\simeq 2$. The total number of authors that posted a submission in the subreddit is 165 931. Of these authors the 80% posted just once in the total history of the data. For this reason it's possible to consider useless for the purposes of the analysis the deepening of this aspect.



(a) Behavior of the number of submissions and of the number of unique authors in time



(b) Average number of submissions per author in time

Extraction of age and gender using regular expressions

In this section the process of extracting age and gender from the submissions using the regular expressions will be explained. The regular expression used in this case has been made with trial-and-error method in order to capture as more data as possible by looking

at the different types of expressions in the submissions. The user commonly writes their demographic data in the title or in the text, so the regular expression has been applied on both of them and then the results are merged. Examples of title and text containing demographic data are the following:

- Title: AITA for not calling my(19F) step-dad "dad"?
- Text: I (24F) live at home with my mom(56) and my brother (34), his wife (35) and their three kids (8,3 and 11 months).

The regular expression that has been constructed is:

```
|^.*((I)+|(I[\s\'']*(a?m)*(\sa)*)|(M?m?y)|([Mm]e))[\s\,\:]
+(([\(\[\s]*((?P<age1>[0-9]{2})[\s\,]*(?P<gender1>[mfMF]))
[\s\,\.]*[\)\]\s]+)|(([\(\[\s]*(?P<gender2>[mfMF]))[\s\,\.]*
(?P<age2>[0-9]{2})[\s\,\.]*[\)\]\s]+)).*|
```

It looks for expressions such as "My, me, I'm, I am, I" followed by two cases: the first one in which the gender is expressed before the age (e.g. F26) and the second one in which is done the opposite (e.g. 26F). The regex seeks only for the cases in which gender and age are expressed together, so the demographic data have not been extracted from the submissions in which:

- Age and gender are indicated in different parts (e.g."I am 21, [...], I'm a male")
- Only one of the two is specified (e.g."I'm a Belgian teenager (16)" or " I (F) have a problem with my sister")

To evaluate the regular expression that has been written, a random sample of 100 submissions has been extracted. In the sample, 12% of the submissions contains the demographic data in the format that can be captured by the regex. The samples have been classified by hand with four labels:

- Age&Gender: binary value, set to 1 if the demographic data is present and 0 else.
- Regex match: binary value, set to 1 if the demographic data is found by the regex and 0 else.
- Age&Gender data: string, if the data are present this is compiled manually with the data.
- Regex result: string, this field is compiled with the result obtained from the regex function when it's applied to the text.

From these four vectors parameters we evaluate the precision and recall of the regular expression:

- The precision is defined as: $P = \frac{T_P}{T_P+F_P}$, where T_P is the number of true positives and F_P is the number of false positives. In the case under study this measure is equal to 1, meaning that no false positive has been found.

- The recall is defined as: $R = \frac{T_P}{T_P+F_N}$, where T_P is the number of true positives as before and F_N is the number of false negatives. This measure is equal to 0.92.

The last two vectors (Age&Gender data and Regex result) are found to be equal, meaning that there is no error in the extraction of the demographic data when a suitable expression is identified by the regular expression. The number of samples containing demographic information in the whole dataset of submissions is 14126 which is 8% of the total number of submissions collected from the subreddit. The final matrix will contain 10 134 rows. The loss of about 4000 submissions is due to the fact that no comments with tags have been given to these submissions. This happens because when the matrices containing different information are merged some part of the data is lost, due to the fact that the link id is not present in both matrices.

3.1.2 Comments extraction and labeling

In an analogous way to that of the submissions, all comments present in the subreddit under analysis have been saved in a separate JSON file. From this file some preliminary information have been extracted:

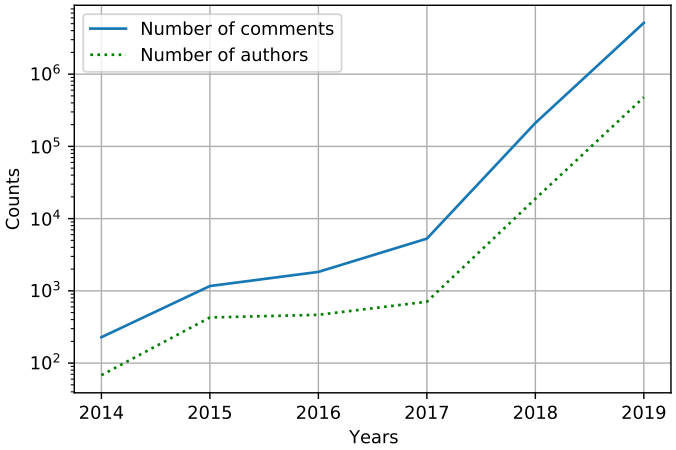
- Total number of comments present in the dataset: 5342015

- Average number of comments per author is $\simeq 10$

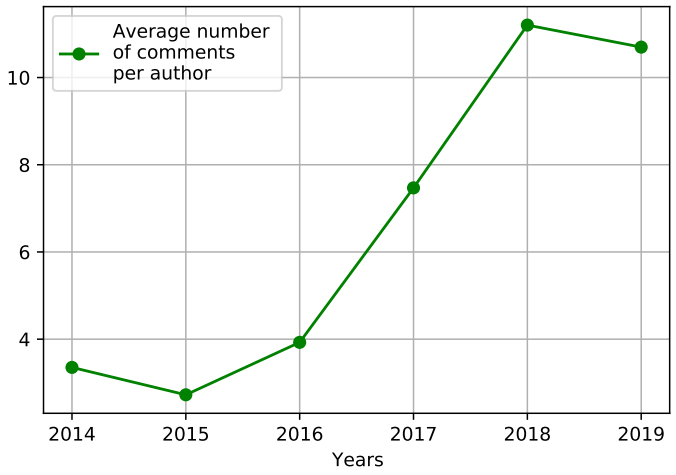
The important work that has been made on the comments is that of the extraction of tags. As it has been said before, tags concerning the judgment of the submissions are present in the comments. In order to create the dataset, the text of each comment has been converted in a tag, and this has been done by searching in the text for the expressed tag. The entire text of the comment has been tokenized by a preliminary function in order to avoid the counting of the cases in which the tag is not real but just present inside a word, e.g "nta" in "representation". Subsequently, all the text has been converted into a singular string containing only the tag uppercase: for example from this comment "That is so unbelievably cruel. Your brother sounds like a total psychopath. You are completely justified to never speak to him again, NTA" to this: "NTA". The modified data have been saved to a .txt file in order to merge everything into a unique dataframe.

Comments' authors

Unlike the submissions' authors, the comments' authors are a relevant key for the study. In order to perform the analysis all the comments done by the account "AutoModerator"



(a) Behavior of the number of comments and of the number of unique authors in time



(b) Average number of comments per author in time

have been deleted. The total number of authors that posted a comment in the subreddit is 499366. In this case only the 43% commented only once. Given that the total number of authors of the comments is very large, in order to study the behavior of the most active users it has been decided to restrict the field to the authors that have made more than 15 comments. The final number of authors under observation is 51049, being the 10% of all authors.

3.2 Topics of the submissions

The subreddit is mainly focused on private questions, and most of the topics of the submissions can be classified into five macrocategories:

- **Family**: situations related to relatives (e.g., an argument between son and mother).
- **Friendships**: situations related to to friends or a group of friends (e.g., a betrayal of one friend with respect to another).
- **Work**: situations related to work environments (e.g., troubles with the boss).
- **Society**: any situation concerning politics, racism or gender questions.
- **Romantic relationships**: situations related to the significant other (e.g., an argument concerning a jealousy question).

The topic classification is based on a subjective evaluation, and is thus performed manually and based on a sample of 200 submissions.

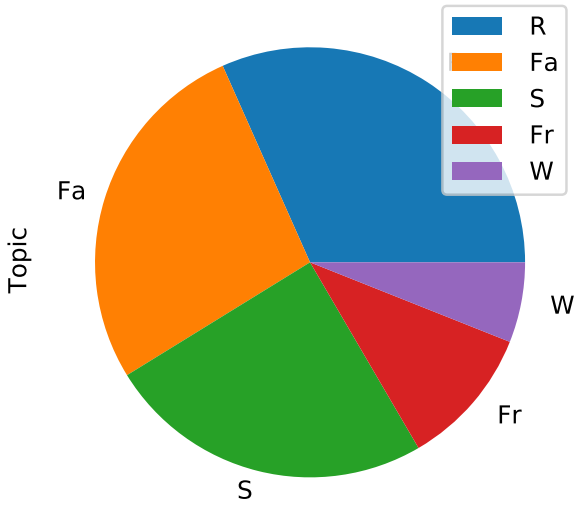


Figure 3.3: Number of submissions in each topic

3.3 Final matrix

In this section the final matrix is described: how it is done and what can be achieved from it. Then, the demographic distribution is analyzed.

3.3.1 Preliminary preparation of the matrix

Once all the relevant information have been gathered from the separate files they are all merged into a single matrix. In order to proceed with this step the starting point are two matrices: the first one containing the data of the comment (Id, Parent Id, Link Id,

Time, Author, Score and Tag), the table 3.1, and the second containing the demographic data of the user that is posting (Age and Gender), the table 3.2. Two examples of both matrices are illustrated below: Since only a restricted number of features of the first

	Id	Parent Id	Link Id	Time	Author	Score	Tag
2914565	erd3i4l	t1_erd0qqp	t3_c1dpdc	1560729792	AshleyBanksHitSingle	39	NTA
4983591	ej4w2rp	t3_b44h1g	t3_b44h1g	1553277920	egoissuffering	1	ESH
2241464	eu3q8eb	t3_ceftzf	t3_ceftzf	1563421896	theymademedarko	1	NTA
4848042	eig99q7	t3_b0prmt	t3_b0prmt	1552504101	therealgoose21	1	NTA
4826495	eib7ntx	t3_azze1g	t3_azze1g	1552343891	need2know25	4	YTA
3187813	ejxe2w8	t3_b868g6	t3_b868g6	1554179141	[deleted]	-1	YTA
2150506	etrupey	t3_cd37hn	t3_cd37hn	1563129504	buckthisnoise	1	NTA

Table 3.1: Initial matrix originated from the comments subset

	Age	Gender
e1jbiz	41	'M'
bamvzn	23	'F'
ds40tk	21	'F'
b28ck1	22	'M'
d9jb4c	19	'F'
a8q8mw	24	'F'
c1bfwa	20	'M'

Table 3.2: Matrix obtained from the submissions

matrix are necessary for the analysis, some data manipulation is done before merging: in the first place all the comments have been grouped by Link Id, obtaining all the tags (comments) that have been posted under a specific submission. Afterward, these tags are counted in order to achieve a matrix, table 3.3, in which each row represents a submission and the number of tags that users gave to this particular submission. If no one gave a certain tag, that value is represented as "NaN". At this point it is possible to proceed with the merging of the two matrices into a final one by using the"Link Id" which, as it has been said before, is the connection point between the submission and the comment, so it is possible to join the two matrices keeping only the Link Id's values that appear in both. It could happen that a submission with demographic data didn't have any comment or, more likely, that the comments were made to a post were no demographic information has been gathered. Indeed starting from the first matrix of dimensions 205740×5 and the second matrix of dimensions 14899×2, the final matrix of dimensions 10590×7 is obtained, that will be the starting point for the analysis:

	NAH	NTA	YTA	ESH	NFO
bbnv9u	1	2	NaN	NaN	NaN
bnoms2	5	35	1	1	NaN
bbe6l0	NaN	3	NaN	NaN	NaN
d30o79	NaN	17	NaN	NaN	NaN
c5bt05	16	34	17	NaN	NaN
dbzz8x	4	1	7	NaN	NaN
aktnca	NaN	1	NaN	NaN	NaN

Table 3.3: Matrix for the count of the tags

	NAH	NTA	YTA	ESH	NFO	Age	Gender
denxs7	4	14	1	NaN	NaN	28	'F'
c2fd9o	5	2	NaN	NaN	NaN	16	'F'
aapzix	1	1	2	1	NaN	22	'F'
bzlorv	3	NaN	NaN	NaN	NaN	19	'M'
cckcwt	1	32	4	NaN	NaN	17	'F'
b4oph9	NaN	5	8	NaN	NaN	17	'M'
akg10z	2	1	1	3	NaN	21	'F'

Table 3.4: Final merged matrix

3.3.2 Demographic composition of the data

First of all, in order to simplify the analysis, the age values are divided into five bins. The division has been done by a quantile-based discretization function, meaning that the function defines the bins using percentiles based on the distribution of the data and not on the actual numeric edges of the bins. This results in having the following parting in bins:

- < 18
- 18-21
- 21-23
- 23-27
- 27-70

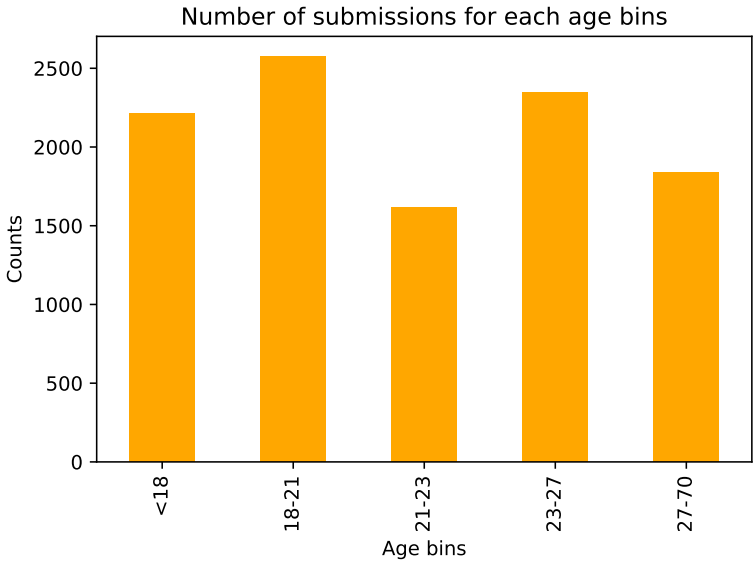


Figure 3.4: Number of submissions for each age bin

For what concerns the gender distribution in the data, the result is that 54% of the users have been identified as "Female" and 46% of the users have been identified as "Male". The gender distribution in each age bin is quite coherent with the one of the total data set:

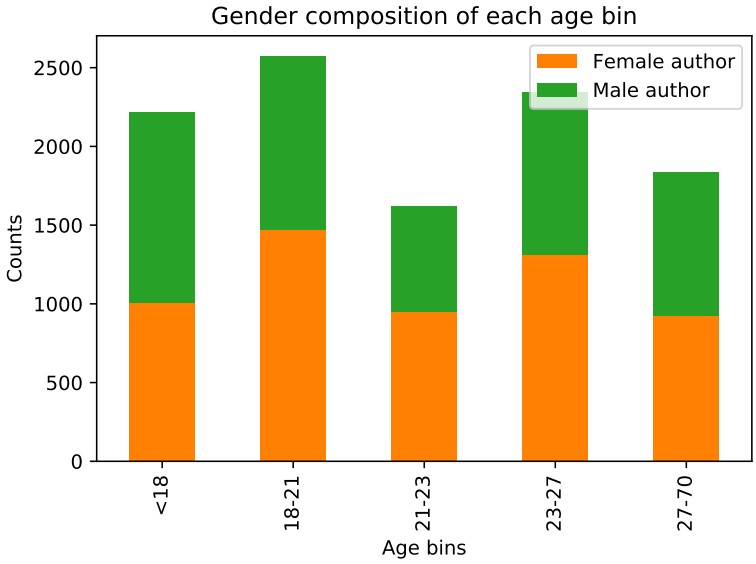


Figure 3.5: Gender composition of each age bin

3.3.3 Correlation between demographic data and tags given

The aim of this subsection is to analyze, given the demographic distribution of the data of the users that are posting, what kind of tags are given to them. To begin with this analysis, a preliminary observation is that of the counts of tags that are given to each age bin:

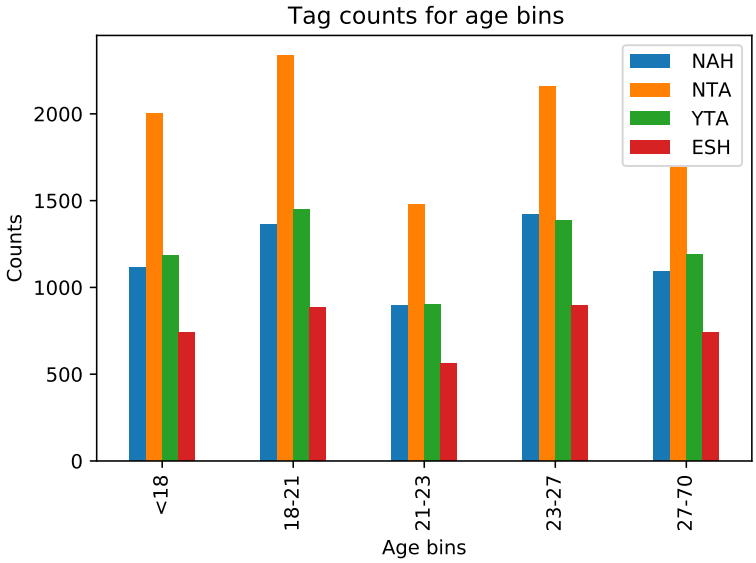


Figure 3.6: Counts for each tag in each age bin

One thing that it is worth noticing from these plots is that the distribution of tag counts is consistent in each age bin and that the counts of NTA is much higher than all the others, meaning that it is the tag that is given the most independently of the age bin.

Subsequently, the conditional probability of each tag in each single age bin has been evaluated. The conditional probability is calculated as the probability of each tag in each age bin given the number of submissions coming from a male or from a female user for that specific age and tag.

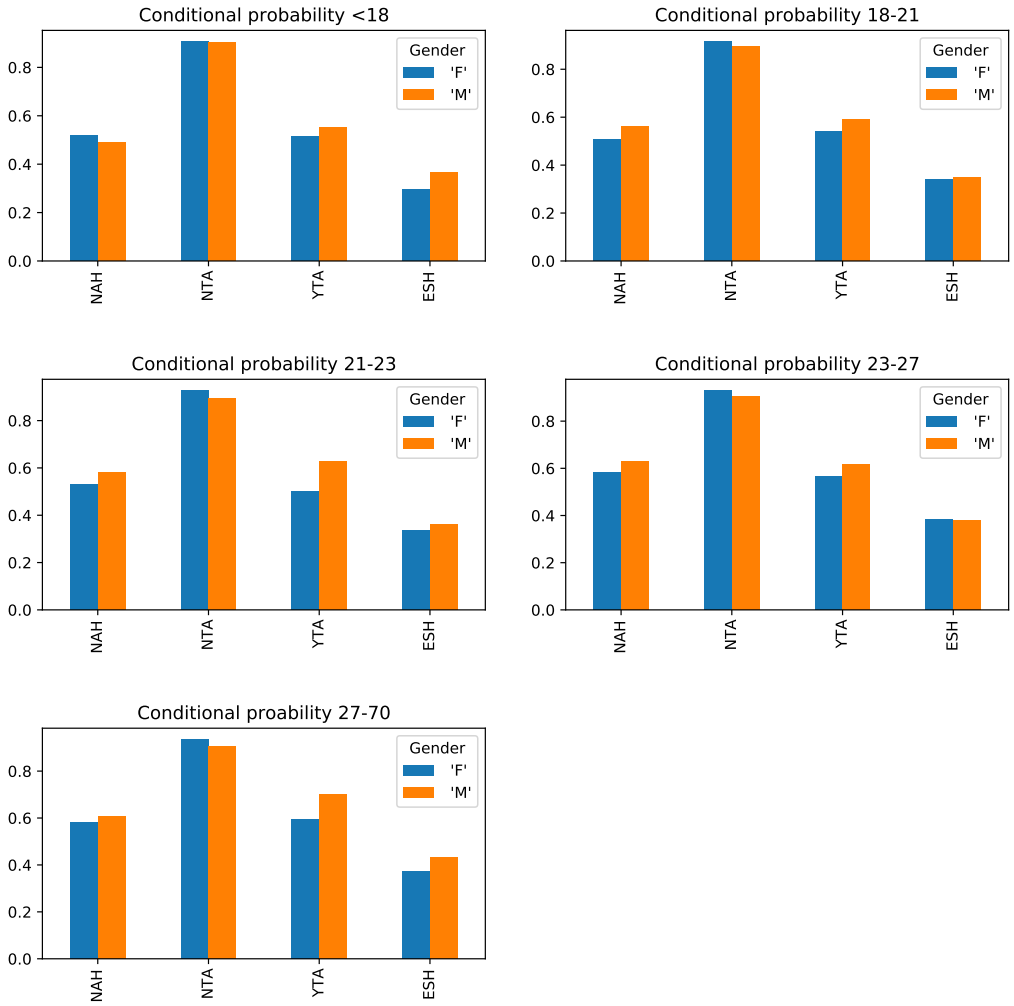


Figure 3.7: Conditional tag counts for each age bin

With regard to these plots, it is interesting to observe how the probability of the different tags is quite equal for each age bin but the bar of NTA is always higher for the women and the bar of YTA is always higher for the men. A feature that it is worth noticing is that the probability for the tag YTA, both in the age bin 21 - 23 and in the age bin 27 - 70, is significantly higher for the male users. In the following pyramidal plot it is possible to better understand the difference between the counts of each tag for male and female users:

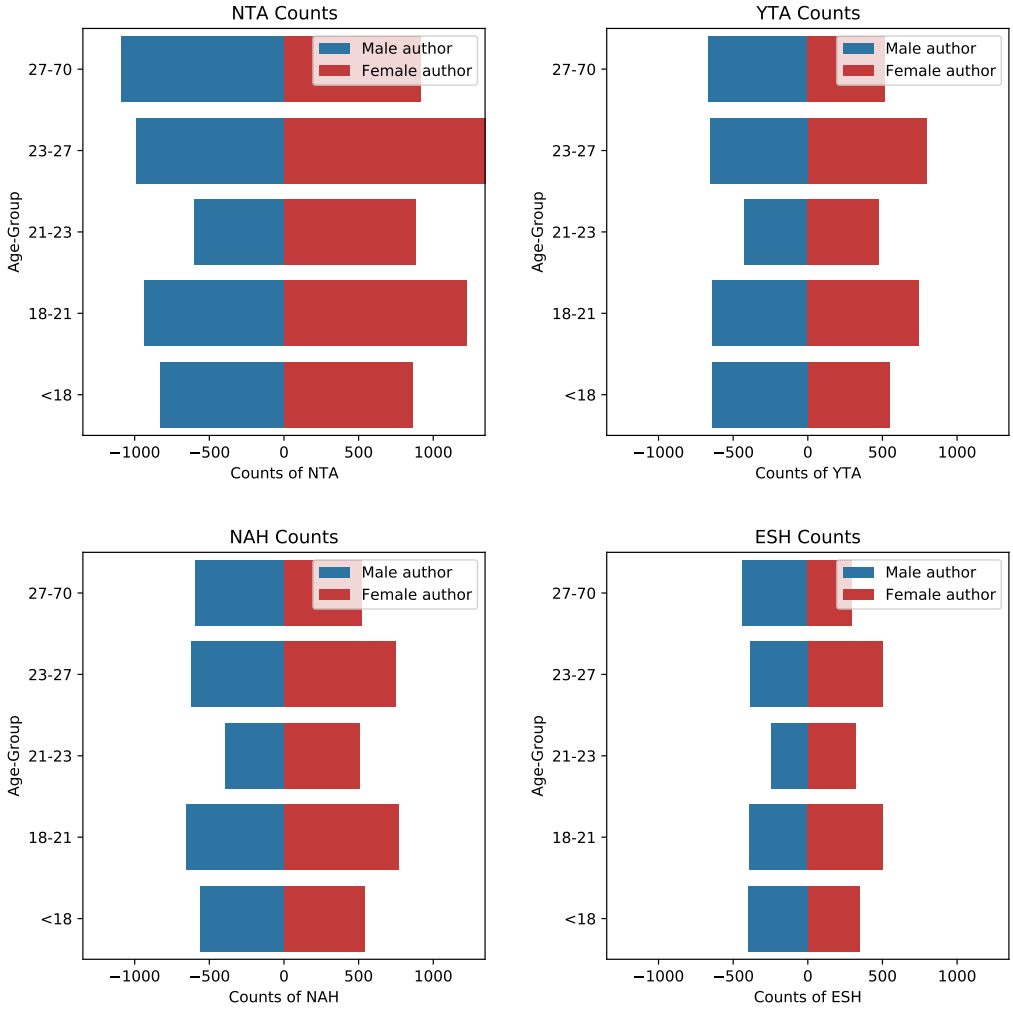


Figure 3.8: Pyramidal plot for each tag

Chapter 4

Statistical tests and results

In this chapter of the thesis four main research questions have been investigated: the correlation between the given judgment and the age and the gender of the poster, how opinions about the same submission are distributed, if topics can actually influence the moral judgment and if the previous history of the comments' authors can predict what their judgment will be. In section 4.1 we answer the first question: the moral opinion of the community will be harsher towards a young male user than towards an old female user? Is there actually a bias in the perception of social norms due to age and gender or the judgment is neutral? Then, in section 4.2 the judgments' homogeneity is studied: is the opinion of the community uniform given a certain situation or social norms can be perceived in a totally different way according to users coming from disparate backgrounds? Subsequently, in the section 4.3 we examine how the perception of social norms can be conditioned by the topic of the situation: if the user speaks about a controversial situation, the community will be influenced by the fact that this situation is about work other than a romantic relationship? In the end, in the section 4.4 we move our attention to the previous history of the authors of the comments: does the background of a user (i.e, the subreddits in which he/she is more active) tells us about his/her perception of social norms? How can social norms change upon different real-life interests?

4.1 Age and Gender

In this section we will investigate the correlation between the given tag and the age and gender of the author of the submission. In order to deepen the topic, several statistical tests have been performed on different combinations of data.

4.1.1 Kruskal-Wallis H-test

The first test performed is the **Kruskal-Wallis H-test**, that tests the null hypothesis that the population median of all of the groups are equal. In order to prepare the input

data for the test the matrix, the final matrix is grouped for age, the tags for each age value are counted and then the table is transposed, leading to a matrix of this type (this is only a sample for the age from 15-24 of the matrix that has been used):

Age	15	16	17	18	19	20	21	22	23	24
NAH	121	199	296	332	392	473	438	465	424	429
NTA	227	379	489	585	663	804	757	782	676	638
YTA	125	223	295	348	418	482	469	492	407	410
ESH	95	140	181	207	236	311	282	295	268	276

Table 4.1: Input matrix for Kruskal's test

For the aim of the test, the age is kept as a continuous variable and it is not divided in bins (as it will be done in the following). The input of the test is made of the four vectors (one for each tag). The test returns the p-value, that results to be **0.13**. The p-value obtained in this case is not significant, so the first conclusion is that there is no substantial difference in the distributions of the tag counts for age. The distributions are plotted below and it can be observed that they show all the same shape:

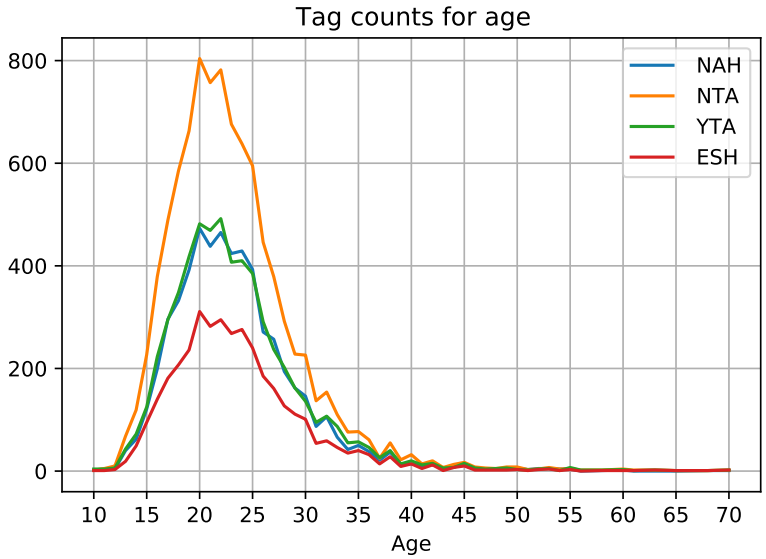


Figure 4.1: Distributions of tag counts for age

4.1.2 Chi-squared test

In the following analysis the Pearson chi-squared statistic has been used. The chi-squared test verifies the null-hypothesis H_0 that the probability of an outcome equals certain fixed value π_{ij}. For a sample of size n with cell counts n_{ij}, the values $\mu_{ij} = n\pi_{ij}$

are the expected frequencies. These frequencies represent the values of the expectations $E(n_{ij})$ when the null-hypothesis is true. For each of n observations of a variable, let π denote the probability of success. To judge whether the data contradict H_0, the values of n_{ij} and π_{ij} are compared. The larger the differences $(n_{ij} - \pi_{ij})$ the stronger will be the evidence against H_0 and then a correlation between the variables will be suggested. The analysis starts from the contingency table, a matrix that stores the frequency distribution of the variables that are taken into account. Let π_{ij} be the joint probabilities for the contigency table, the null hypothesis of statistical independence is:

$$H_0 : \pi_{ij} = \pi_{i+} pi_{+j} \quad \text{for all i and j} \tag{4.1}$$

like The marginal probabilities then determine the joint probabilities. To test H_0, it is identified $\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ as the expected frequency. μ_{ij} is the expected value of n_{ij} assuming independence.

The final formula to estimate the expected frequencies is:

$$\mu_{ij} = np_{i+}p_{+j} = n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+}n_{j+}}{n} \tag{4.2}$$

For testing independence in I x J contingency tables, the Pearson statistics equal:

$$X^2 = \sum_{ij} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \tag{4.3}$$

The function employed for the test then evaluates four parameters from the contingency table (which is given to the test as the input): the test statistics, the p-value of the test, the degrees of freedom, the expected frequencies (based on the marginal sums of the contingency table). In the following, the result that will be needed to define if the correlation is significant or not is the p-value. [2]

Chi-squared between tags and age bins

The question that we explored in this paragraph is whether the categorical variables of tags and age bins are correlated, namely if the age bin that contains the age of the author has an impact on the judgment that is given to him/her. The contingency matrix that has been used for the test is:

Age bins	NAH	NTA	YTA	ESH
<18	1060	1884	1120	696
18-21	1303	2224	1369	829
21-23	889	1458	899	563
23-27	1350	2058	1323	862
27-70	1072	1640	1159	737

Table 4.2: Contingency matrix

The resulting p-value of this test is 0.009 and consequently it is **significant** at the 0.01 level. This leads to the conclusion that there is a correlation between the age of the poster and the judgment.

Chi-squared between tags and gender

Proceeding with the investigation of the correlation of the tag counts within the data records, the chi-squared test is performed between tag counts and gender. The most interesting result that has been found in this test is the correlation between the gender of the poster and the judgments that are given to him/her. In this case the contingency table is evaluated from the matrix:

Gender	NAH	NTA	YTA	ESH
'F'	3033	5116	3025	1932
'M'	2641	4148	2845	1755

Table 4.3: Contingency matrix

The obtained p-value is $6.52e^{-05}$ which is significant at the 0.01 level. In order to better understand in which age bin the correlation between tag counts and gender is present, the same test as before is carried out in each age bin:

- Test in the age bin "< 18": The contingency matrix for this test is:

Gender	NAH	NTA	YTA	ESH
'F'	529	898	523	305
'M'	531	986	597	391

Table 4.4: Contingency matrix for the age bin < 18

The resulting p-value is 0.08 which is not significant.

- Test in the age bin "18-21": The contingency matrix for this test is:

Gender	NAH	NTA	YTA	ESH
'F'	733	1326	769	480
'M'	570	898	600	349

Table 4.5: Contingency matrix for the age bin 18-21

The resulting p-value here is 0.12, also in this case this is not significant.

- Test in the age bin "21-23": The contingency matrix for this test is:

Gender	NAH	NTA	YTA	ESH
'F'	502	864	476	314
'M'	387	594	423	249

Table 4.6: Contingency matrix for the age bin 21-23

In contrast with the other results in this case the p-value is 0.026 which is significant at the 0.05 level.

- Test in the age bin "23-27": The contingency matrix for this test is:

Gender	NAH	NTA	YTA	ESH
'F'	736	1184	720	486
'M'	614	874	603	376

Table 4.7: Contingency matrix for the age bin 23-27

In this case the p-value returns to be non significant, having a value of 0.21.

- Test in the age bin "27-70": The contingency matrix for this test is:

Gender	NAH	NTA	YTA	ESH
'F'	533	844	537	347
'M'	539	796	622	390

Table 4.8: Contingency matrix for the age bin 27-70

The p-value here is equal to 0.03, which is significant at the 0.05 level.

In conclusion, the noteworthy results are found in the age bins 21-23 and 27-70, meaning that in these age bins the judgment of the users can be influenced by the gender of the author.

4.1.3 Binomial regression model

In this subsection the correlations between the gender and each tag have been analyzed by means of the binomial regression model. The binomial regression model belongs to the family of generalized linear models. Since in this case the correlations between the gender (a binary variable, since it can take two values due to how the data set has been constructed) and the different types of judgments. This has been done both with the binomial regression model for the correlations between the gender and each single tag and, as it will be described in the following paragraph, with the multinomial regression model. The binomial regression model is generally used for the aim of predicting the possibility of observing a specific outcome. In this model the dependent variable y is a discrete random variable whose values represents the number of successes observed in m trials. In this particular case the number of trials is the number of comments under a submission and the number of successes is the number of times the comment under examination has been used. This means that the random variable follows a binomial distribution. In the following a regression model will be applied: it is assumed that the dependent variable y depends on a matrix of regression variables X. In this case the number of observations is equivalent to the number of submissions in the dataset, it will be called n subsequently. For each observation it is effective to express the probability of y_i (with i in the range [0,N]) of taking a value k as conditional upon the regression variables X taking the value x_i. The probability distribution of the Binomially distributed y in the context of a regression of y over X is the following:

$$Pr(y_i = k|\mathbf{X} = \mathbf{x_i}) = \binom{m}{k} \pi_i^k \cdot (1 - \pi_i)^{m-k} \tag{4.4}$$

The probability of observing a success π_i for a certain regression variable is expressed as some function of the regression variable:

$$\pi_i = g(x_i) \tag{4.5}$$

The important quantity for the Binomial Regression model is the expected value of the response variable. A suitable link function relates the expected value of the response variable E($\mathbf{y}|\mathbf{X}$) and $\mathbf{X}$:

$$g(E(\mathbf{y} = y_i|\mathbf{X}) = \mathbf{x_i}) = \sum_{j=0}^{p} \boldsymbol{\beta}_j \cdot \mathbf{x_{ij}} \tag{4.6}$$

In this equation g is the link function, $\boldsymbol{X}$ is the matrix of regression variables and $\boldsymbol{\beta}$ is a vector of regression coefficients. In this specific case of the Binomial regression model the Generalized Linear Model equation will be written as:

$$g(\pi_i|\boldsymbol{x_i}) = \sum_{j=0}^{p} \boldsymbol{\beta_j} \cdot \boldsymbol{x_{ij}} \tag{4.7}$$

The link function can take several forms, in this analysis the logit link function has been used:

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) \tag{4.8}$$

It expresses the log odds of success. The training of the model is done using the Iterative Reweighted Least Squares algorithm. Expanding on the detail the topic in the case of the analysis the binomial regression has been performed on each tag with gender and age as regression variables.

Binomial regression for single tag and gender

In this paragraph the results of the binomial regression will be listed, with a particular attention to the significant ones. The model in this case has only one degree of freedom: the gender of the author. The number of observations is variable since for each test have been taken into account only the submissions that have received that particular tag. In order to prepare the data for the test the probability for each tag has been evaluated for each submission by taking into account:

$$P_{tag} = \frac{\#\text{times the tag appears}}{\#\text{total comments for the submission}} \tag{4.9}$$

In the end the matrix that has been used for the test (deleting from time to time the rows with probability zero for the tag under analysis) is:

	NAH	NTA	YTA	ESH	Age	Gender	Age bins	Total comments	NAH_Prob	YTA_Prob	ESH_Prob	NTA_Prob
bspdcu	4	8	105	12	26	'M'	23-27	129	0.031	0.813	0.093	0.062
ad9rka	NaN	8	NaN	NaN	25	'F'	23-27	8	NaN	NaN	NaN	1
aqwkl3	2	120	4	NaN	25	'F'	23-27	126	0.015	0.031	NaN	0.952
b44nq8	4	4	2	2	26	'M'	23-27	12	0.333	0.166	0.166	0.333
c83666	NaN	4	NaN	NaN	21	'F'	18-21	4	NaN	NaN	NaN	1
bzh5k3	1	7	NaN	NaN	20	'M'	18-21	8	0.125	NaN	NaN	0.875
bj04aq	3	3	NaN	NaN	41	'M'	27-70	6	0.500	NaN	NaN	0.500

Table 4.9: Sample of the matrix for the binomial regression test

- **YTA & Gender** In this instance the regression coefficient is **significant** at the 0.01 level since the p-value is < 0.001 and the coefficient is positive, being 0.2921. This results in a higher probability for a male user of receiving a negative judgment. The output of the test with the discussed values is:

Dep. Variable:	YTA_Probability	**No. Observations:**	5870
Model:	GLM	**Df Residuals:**	5868
Model Family:	Binomial	**Df Model:**	1
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2943.4
Date:	Mon, 29 Mar 2021	**Deviance:**	2191.5
Time:	16:03:34	**Pearson chi2:**	1.99e+03
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-0.7392	0.039	-19.015	0.000	-0.815	-0.663
Gender[T. 'M']	0.2921	0.055	5.343	0.000	0.185	0.399

- **NTA & Gender:** Also in this instance the regression coefficient is **significant** being the p-value equal to 0.000 and the coefficient in this case is equal to -0.3371. This means that a male user has a lower probability with respect to a female user in receiving a positive judgment. This result is in agreement with the previous result, showing a higher inclination in judging negatively male users and positively female users. The output of the test is:

Dep. Variable:	NTA_Probability	**No. Observations:**	9264
Model:	GLM	**Df Residuals:**	9262
Model Family:	Binomial	**Df Model:**	1
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-4838.8
Date:	Mon, 29 Mar 2021	**Deviance:**	4394.6
Time:	16:22:04	**Pearson chi2:**	3.68e+03
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	0.6572	0.029	22.290	0.000	0.599	0.715
Gender[T. 'M']	-0.3371	0.043	-7.818	0.000	-0.422	-0.253

- **ESH & Gender:** In this case the p-value is 0.535, greater than 0.001, so definitely not significant. The result of the test is:

Dep. Variable:	ESH_Probability	**No. Observations:**	3687
Model:	GLM	**Df Residuals:**	3685
Model Family:	Binomial	**Df Model:**	1
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1366.7
Date:	Mon, 29 Mar 2021	**Deviance:**	710.67
Time:	16:46:46	**Pearson chi2:**	767.
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-1.4507	0.058	-25.010	0.000	-1.564	-1.337
Gender[T. 'M']	0.0517	0.083	0.621	0.535	-0.112	0.215

- **NAH & Gender:** Even in this case the p-value is not significant, so it has not been taken into account. The result of the test is:

Dep. Variable:	NAH_Probability	**No. Observations:**	5674
Model:	GLM	**Df Residuals:**	5672
Model Family:	Binomial	**Df Model:**	1
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2534.5
Date:	Mon, 29 Mar 2021	**Deviance:**	1612.6
Time:	16:58:32	**Pearson chi2:**	1.56e+03
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-0.9912	0.041	-24.254	0.000	-1.071	-0.911
Gender[T. 'M']	0.0396	0.060	0.664	0.506	-0.077	0.156

Binomial regression for macro-category and gender

In this paragraph, as it will be done also in the next section, an aggregation of the tags has been performed. The tags have been divided into two macro-categories which are N (that stands for negative tags) and P (that stands for positive tags). The binomial regression test has then been carried out on this data matrix:

	Age	Gender	Age bins	N	P	Tot comments	N_Prob	P_Prob
ch9nql	28	'F'	27-70	6	0	6	1	0
a6stzu	22	'F'	21-23	1	4	5	0.2	0.8
bdeak4	34	'F'	27-70	0	4	4	0	1
c8ggzd	32	'M'	27-70	0	8	8	0	1
clt6m5	29	'M'	27-70	2	4	6	0.3	0.6
a1c6l6	17	'M'	< 18	2	2	4	0.5	0.5
cyzl0a	21	'F'	18-21	1	10	11	0.1	0.9

Table 4.10: Sample of the matrix with the tag in macro-categories

From this starting point, the binomial regression test is done on both macro-categories with the same process used in the paragraph before. The independent variable is gender and the dependent one is the macro-category. The results are listed and commented on in the following:

- **Positive judgments & Gender**: The resulting p-value is **significant** at the 0.01 level with a negative coefficient for the male users. The resulting p-value is significant at the 0.01 level with a negative coefficient for the male users. This result means that there is a lower probability for the male user of receiving a positive judgment. Since there are two categories for the variable gender this is equivalent to saying that there is a higher probability for the female user to receiving a positive judgment.

Dep. Variable:	P_Probability	**No. Observations:**	10131
Model:	GLM	**Df Residuals:**	10129
Model Family:	Binomial	**Df Model:**	1
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-4916.7
Date:	Mon, 19 Apr 2021	**Deviance:**	5508.5
Time:	11:03:44	**Pearson chi2:**	4.84e+03
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	1.1368	0.031	36.270	0.000	1.075	1.198
Gender[T. 'M']	-0.3694	0.045	-8.287	0.000	-0.457	-0.282

Table 4.11: Generalized Linear Model Regression Results

- **Negative judgments & Gender**: As was expected this result is mirror-like concerning the previous one and is reported for completeness. The resulting p-

value is significant at the 0.01 level with a positive coefficient for the male users. This result means that there is a higher probability for the male user of receiving a negative judgment.

Dep. Variable:	N_Probability	**No. Observations:**	10131
Model:	GLM	**Df Residuals:**	10129
Model Family:	Binomial	**Df Model:**	1
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-4916.7
Date:	Mon, 19 Apr 2021	**Deviance:**	5508.5
Time:	10:56:28	**Pearson chi2:**	4.84e+03
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-1.1368	0.031	-36.270	0.000	-1.198	-1.075
Gender[T. 'M']	0.3694	0.045	8.287	0.000	0.282	0.457

Table 4.12: Generalized Linear Model Regression Results

Binomial regression for single tag and age bins

In this section, the correlation between the single tag and the age bins is explored. In the previous section 4.1.2 the chi-squared results reported a correlation between the distributions of tag counts and age bins. In the following, this correlation will be deepened through the binomial regression test applied on the single tag and the age bins.

- **YTA & Age bins:** In this case, none of the values is significant, meaning that there is not a particular age bin for which the outcome probability of the tag YTA is higher or lower significantly.

Dep. Variable:	YTA_Probability	**No. Observations:**	5870
Model:	GLM	**Df Residuals:**	5865
Model Family:	Binomial	**Df Model:**	4
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2956.6
Date:	Mon, 19 Apr 2021	**Deviance:**	2217.9
Time:	15:34:10	**Pearson chi2:**	2.00e+03
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-0.5747	0.062	-9.232	0.000	-0.697	-0.453
Agebins[T.18-21]	0.0035	0.084	0.042	0.967	-0.161	0.168
Agebins[T.21-23]	0.0106	0.093	0.114	0.910	-0.172	0.193
Agebins[T.23-27]	-0.0951	0.085	-1.117	0.264	-0.262	0.072
Agebins[T.27-70]	-0.0059	0.087	-0.068	0.946	-0.177	0.165

Table 4.13: Generalized Linear Model Regression Results

- **NTA & Age bins:** In this test a p-value significant at the 0.05 level is obtained for the 27-70 age bin with a coefficient equal to -0.1601. This means that there is a lower probability for the category with age 27-70 of getting a negative judgment.

Dep. Variable:	NTA_Probability	**No. Observations:**	9264
Model:	GLM	**Df Residuals:**	9259
Model Family:	Binomial	**Df Model:**	4
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-4864.9
Date:	Mon, 19 Apr 2021	**Deviance:**	4446.8
Time:	15:32:42	**Pearson chi2:**	3.71e+03
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	0.5424	0.048	11.351	0.000	0.449	0.636
Agebins[T.18-21]	0.0307	0.065	0.472	0.637	-0.097	0.158
Agebins[T.21-23]	-0.0411	0.072	-0.570	0.569	-0.182	0.100
Agebins[T.23-27]	-0.0519	0.066	-0.788	0.431	-0.181	0.077
Agebins[T.27-70]	-0.1601	0.069	-2.308	0.021	-0.296	-0.024

Table 4.14: Generalized Linear Model Regression Results

- **ESH & Age bins:** In this test there are several significant p-values. The age bin 21-23 is significant at the 0.05 level with a coefficient equal to -0.2684: this age range has a lower probability of having an ESH comment. The age bins 23-27 and 27-70 have a p-value significant at the 0.01 level and both of them present a negative coefficient, so a lower probability.

Dep. Variable:	ESH_Probability	**No. Observations:**	3687
Model:	GLM	**Df Residuals:**	3682
Model Family:	Binomial	**Df Model:**	4
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1358.6
Date:	Mon, 19 Apr 2021	**Deviance:**	694.38
Time:	15:41:55	**Pearson chi2:**	748.
No. Iterations:	5		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-1.1467	0.089	-12.940	0.000	-1.320	-0.973
Agebins[T.18-21]	-0.2296	0.124	-1.853	0.064	-0.472	0.013
Agebins[T.21-23]	-0.2684	0.138	-1.939	0.052	-0.540	0.003
Agebins[T.23-27]	-0.4282	0.127	-3.384	0.001	-0.676	-0.180
Agebins[T.27-70]	-0.4756	0.133	-3.574	0.000	-0.736	-0.215

Table 4.15: Generalized Linear Model Regression Results

- **NAH & Age bins:** In this test the p-value for the age bin 23-27 is significant at the 0.05 level with a negative coefficient and the p-value for the age bin 27-70 is significant at the 0.05 level with again a negative coefficient.

Dep. Variable:	NAH_Probability	**No. Observations:**	5674
Model:	GLM	**Df Residuals:**	5669
Model Family:	Binomial	**Df Model:**	4
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2529.7
Date:	Mon, 19 Apr 2021	**Deviance:**	1603.1
Time:	15:39:16	**Pearson chi2:**	1.55e+03
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-0.8408	0.067	-12.561	0.000	-0.972	-0.710
Agebins[T.18-21]	-0.0784	0.091	-0.864	0.388	-0.256	0.100
Agebins[T.21-23]	-0.1143	0.100	-1.138	0.255	-0.311	0.083
Agebins[T.23-27]	-0.1903	0.091	-2.088	0.037	-0.369	-0.012
Agebins[T.27-70]	-0.2789	0.098	-2.860	0.004	-0.470	-0.088

Table 4.16: Generalized Linear Model Regression Results

Binomial regression model for single tag and age bins combined with gender

In this subsection the binomial regression model is performed for each single tag and with five regression variables, which are the four age bins and the gender.

- **YTA & Age bins plus Gender**

Dep. Variable:	YTA_Probability	**No. Observations:**	5870
Model:	GLM	**Df Residuals:**	5864
Model Family:	Binomial	**Df Model:**	5
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2942.4
Date:	Mon, 19 Apr 2021	**Deviance:**	2189.5
Time:	15:46:01	**Pearson chi2:**	1.98e+03
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-0.7334	0.069	-10.568	0.000	-0.869	-0.597
Agebins[T.18-21]	0.0313	0.084	0.371	0.710	-0.134	0.196
Agebins[T.21-23]	0.0290	0.093	0.310	0.757	-0.154	0.212
Agebins[T.23-27]	-0.0730	0.085	-0.854	0.393	-0.240	0.095
Agebins[T.27-70]	-0.0070	0.088	-0.080	0.936	-0.179	0.165
Gender[T. 'M']	0.2922	0.055	5.326	0.000	0.185	0.400

Table 4.17: Generalized Linear Model Regression Results

- **NTA & Age bins plus Gender**

Dep. Variable:	NTA_Probability	**No. Observations:**	9264
Model:	GLM	**Df Residuals:**	9258
Model Family:	Binomial	**Df Model:**	5
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-4834.6
Date:	Mon, 19 Apr 2021	**Deviance:**	4386.3
Time:	15:47:01	**Pearson chi2:**	3.67e+03
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	0.7225	0.053	13.515	0.000	0.618	0.827
Agebins[T.18-21]	-0.0096	0.065	-0.147	0.883	-0.138	0.119
Agebins[T.21-23]	-0.0807	0.073	-1.112	0.266	-0.223	0.061
Agebins[T.23-27]	-0.0857	0.066	-1.293	0.196	-0.216	0.044
Agebins[T.27-70]	-0.1741	0.070	-2.500	0.012	-0.311	-0.038
Gender[T. 'M']	-0.3369	0.043	-7.775	0.000	-0.422	-0.252

Table 4.18: Generalized Linear Model Regression Results

- **ESH & Age bins plus Gender**

Dep. Variable:	ESH_Probability	**No. Observations:**	3687
Model:	GLM	**Df Residuals:**	3681
Model Family:	Binomial	**Df Model:**	5
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1358.5
Date:	Mon, 19 Apr 2021	**Deviance:**	694.17
Time:	15:47:52	**Pearson chi2:**	747.
No. Iterations:	5		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-1.1682	0.101	-11.609	0.000	-1.365	-0.971
Agebins[T.18-21]	-0.2242	0.124	-1.802	0.072	-0.468	0.020
Agebins[T.21-23]	-0.2639	0.139	-1.902	0.057	-0.536	0.008
Agebins[T.23-27]	-0.4235	0.127	-3.335	0.001	-0.672	-0.175
Agebins[T.27-70]	-0.4744	0.133	-3.564	0.000	-0.735	-0.214
Gender[T. 'M']	0.0380	0.084	0.452	0.651	-0.127	0.203

Table 4.19: Generalized Linear Model Regression Results

- **NAH & Age bins plus Gender**

Dep. Variable:	NAH_Probability	**No. Observations:**	5674
Model:	GLM	**Df Residuals:**	5668
Model Family:	Binomial	**Df Model:**	5
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2529.5
Date:	Mon, 19 Apr 2021	**Deviance:**	1602.6
Time:	15:48:58	**Pearson chi2:**	1.55e+03
No. Iterations:	4		

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-0.8616	0.073	-11.733	0.000	-1.006	-0.718
Agebins[T.18-21]	-0.0758	0.091	-0.834	0.404	-0.254	0.102
Agebins[T.21-23]	-0.1116	0.101	-1.110	0.267	-0.309	0.085
Agebins[T.23-27]	-0.1884	0.091	-2.067	0.039	-0.367	-0.010
Agebins[T.27-70]	-0.2790	0.098	-2.861	0.004	-0.470	-0.088
Gender[T. 'M']	0.0414	0.060	0.692	0.489	-0.076	0.159

Table 4.20: Generalized Linear Model Regression Results

4.1.4 Multicategory Logit Model

In this chapter, the response variable has several categories so in order to perform the analysis, the Multicategory Logit Model will be used. At each setting of the explanatory variables, the multicategory models assume the counts in the category of Y have a multinomial distribution. This generalization of the binomial distribution applies when the number of categories exceeds two. In this case, indeed, the number of categories is equal to four. By denoting with J the number of categories for Y the response probabilities are denoted with $\{\pi_1, ..., \pi_J\}$ and they satisfy sum. Considering n independent observations, the probability distribution for the number of outcomes of the J types is the multinomial. It specifies the probability for each possible way the n observations can fall in the J categories. Multicategory logit models simultaneously use all pair of categories by specifying the odds of outcome in one category instead of another. Logit models for nominal response variables pair each category with a baseline category. Considering J as the baseline, the baseline-category logits are:

$$\log\left(\frac{\pi_j}{\pi_J}\right) \quad j = 1, ..., J-1 \tag{4.10}$$

Multinomial logistic regression for Tag and Gender

Dep. Variable:	y	**No. Observations:**	1585
Model:	MNLogit	**Df Residuals:**	1579
Method:	MLE	**Df Model:**	3
Date:	Mon, 19 Apr 2021	**Pseudo R-squ.:**	-0.1903
Time:	15:53:56	**Log-Likelihood:**	-1979.8
converged:	True	**LL-Null:**	-1663.3

y=YTA_Probability	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	0.6609	0.119	5.575	0.000	0.429	0.893
Gender[T. 'M']	0.2077	0.164	1.266	0.206	-0.114	0.529
y=ESH_Probability	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-0.1155	0.140	-0.823	0.410	-0.390	0.159
Gender[T. 'M']	0.0065	0.197	0.033	0.974	-0.380	0.393
y=NTA_Probability	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	1.2464	0.109	11.408	0.000	1.032	1.460
Gender[T. 'M']	-0.1634	0.155	-1.053	0.292	-0.468	0.141

Table 4.21: MNLogit Regression Results

Multinomial regression for tag and age bins

Dep. Variable:	y	**No. Observations:**	1585
Model:	MNLogit	**Df Residuals:**	1570
Method:	MLE	**Df Model:**	12
Date:	Mon, 19 Apr 2021	**Pseudo R-squ.:**	-0.1923
Time:	15:56:05	**Log-Likelihood:**	-1983.3
converged:	True	**LL-Null:**	-1663.3

y=YTA_Probability	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	0.7699	0.212	3.640	0.000	0.355	1.185
Agebins[T.18-21]	-0.0521	0.278	-0.187	0.851	-0.597	0.493
Agebins[T.21-23]	-0.0545	0.306	-0.178	0.859	-0.655	0.546
Agebins[T.23-27]	0.0288	0.264	0.109	0.913	-0.489	0.546
Agebins[T.27-70]	0.0481	0.270	0.178	0.859	-0.481	0.577
y=ESH_Probability	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-0.0216	0.249	-0.087	0.931	-0.509	0.466
Agebins[T.18-21]	-0.0221	0.327	-0.068	0.946	-0.662	0.618
Agebins[T.21-23]	-0.0233	0.359	-0.065	0.948	-0.728	0.681
Agebins[T.23-27]	-0.1579	0.316	-0.500	0.617	-0.776	0.461
Agebins[T.27-70]	-0.1861	0.325	-0.573	0.567	-0.822	0.450
y=NTA_Probability	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	1.0740	0.203	5.302	0.000	0.677	1.471
Agebins[T.18-21]	0.0765	0.264	0.289	0.772	-0.442	0.595
Agebins[T.21-23]	0.1109	0.290	0.383	0.702	-0.457	0.679
Agebins[T.23-27]	0.1258	0.252	0.500	0.617	-0.368	0.619
Agebins[T.27-70]	0.1160	0.258	0.450	0.653	-0.389	0.622

Table 4.22: MNLogit Regression Results

4.2 Tags

In this section, the tags distributions have been analyzed. First of all a co-occurrence matrix has been evaluated in order to understand the interaction between tags referring to the same submission. Are all the users in agreement? If not, what are the tags that are more likely to compete for the final judgment? What are the tags that appear together the most? After the analysis of the co-occurrence matrix, the following subject of this section is entropy. As it will be examined entropy reports of the average level of

uncertainty of the outcomes of a random variable. In this case the random variable is the judgment given to the submission and the question is: what is the level of uncertainty of the result?

4.2.1 Co-occurrence matrix

The co-occurrence matrix is an upper triangular matrix of dimensions n x n. In this case, n is the number of tags and is equal to 5: NTA, YTA, ESH, NAH, NFO. Each row and each column represents a tag and their intersection is filled if the two appears together under a submission:

$$C(i,j) = \sum_{x=1}^{N} \sum_{y=1}^{N} \begin{cases} 1 & \text{if x contains i and y contains j} \\ 0 & \text{else} \end{cases} \tag{4.11}$$

where N is the total number of submission. In order to build the co-occurrence matrix, the starting point is the matrix containing the Link Id of the submission and its relative count of tags. From here it has been obtained a dictionary that holds as keys the 15 possible combinations of the tags (being $\binom{5}{2}$ plus the 5 couples of the tag with itself) and as values the number of times that the combination appears. The intermediate step in building the co-occurrence matrix, the dictionary, is this one:

	Value
NAH, NTA	4.87×10^9
NAH, NAH	1.50×10^{10}
NTA, NTA	1.77×10^{10}
NAH, YTA	5.00×10^9
YTA, YTA	1.63×10^{10}
NTA, YTA	4.22×10^9
ESH, NTA	4.71×10^9
ESH, ESH	1.47×10^{10}
ESH, NAH	5.21×10^9
ESH, YTA	4.87×10^9
NFO, NTA	5.58×10^6
NFO, NFO	2.26×10^7
NFO, YTA	5.60×10^6
ESH, NFO	5.55×10^6
NAH, NFO	3.22×10^6

Table 4.23: Dictionary with tag counts

Then, from this the co-occurrence matrix (after a renormalization of the values) is achieved and visualized as a heat map:

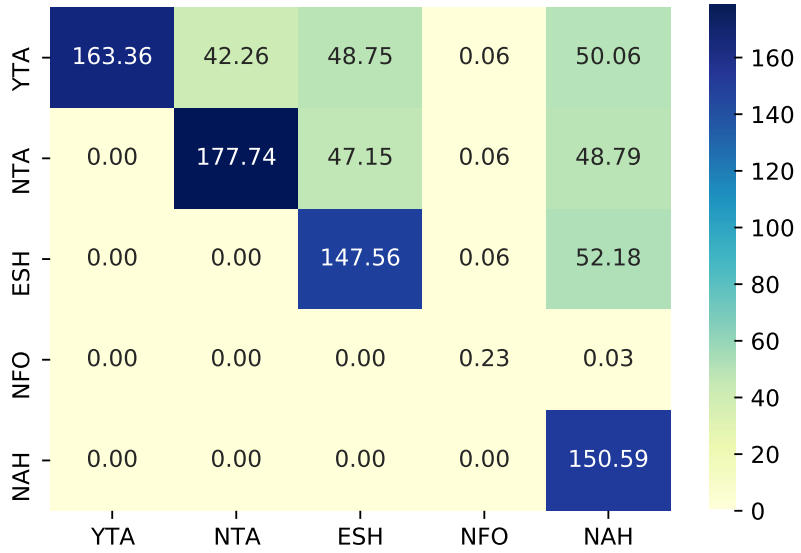


Figure 4.2: Co-occurrence matrix for tags

The highest values are the ones for ESH-NAH, YTA-NAH, YTA-ESH. This means that in the submissions where the judgment is splitted between two or more opinions these are the tags that compete the most. However in general, it is more common that the judgments is divided between a "stronger" opinion (YTA) and a "lighter" one like ESH, NAH.

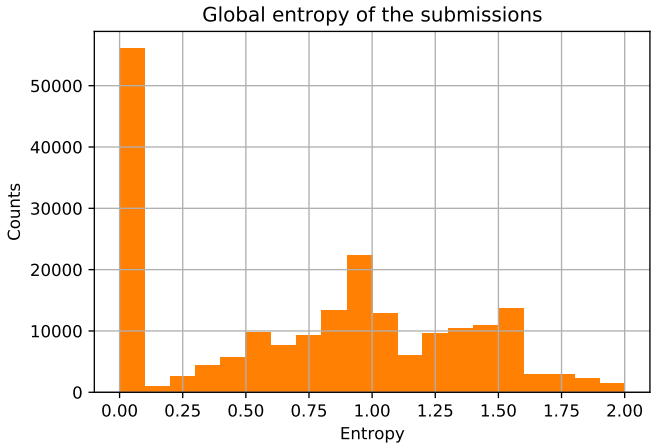
4.2.2 Entropy

After the investigation of the co-occurrence of certain tags under the submissions the entropy of the distribution of probability values has been evaluated. The entropy has been analyzed in three cases: the global one, the couple YTA-NTA and the couple ESH-NAH. These two couples have been taken into account in order to understand the competition between two opposite mirror opinions.

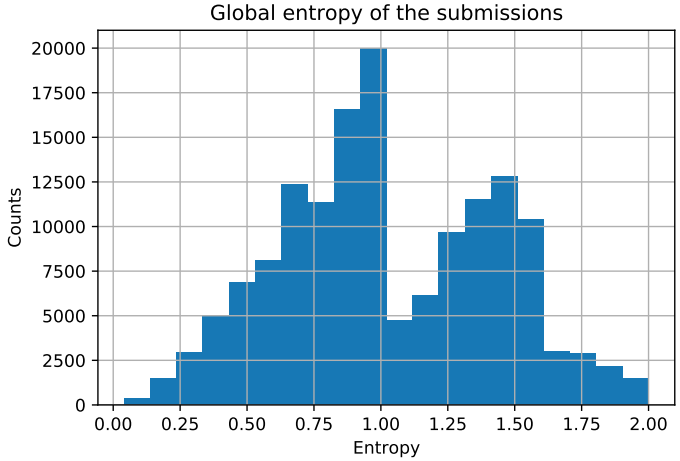
Global entropy

The entropy needs as an input the probability distribution, so from the initial matrix containing Link Id and tag the first step is that of extracting the probabilities. In order to do this, the probabilities regarding each submission have been stored in a nested dictionary-like structure. The dictionary holds in the keys the Link Id of the submission and in the values a further dictionary containing the tag and its relative probability. The probability of each tag is evaluated as the number of the times that the tag has been

used in the comments of that submission over the number of total tags/comments under that submission. The next step is computing the entropy: the entropy function has been applied, as an input the list of the probability values has been passed and the function returns as an output the entropy value of each submission. The logarithmic base that has been used is the binary one and the possible outcomes are 4, this is the reason why the entropy range goes from 0 to 2. The submissions that report an entropy value equal to 0 will be that with a certain outcome, e.g. submissions with only one tag or with only one type of tag. The other extreme case, the submissions with an entropy value equal to 2, means that the outcome is totally uncertain: e.g. submissions with two tags with the same probability and so on. The plot of all the entropy values is shown below:



(a) Global entropy

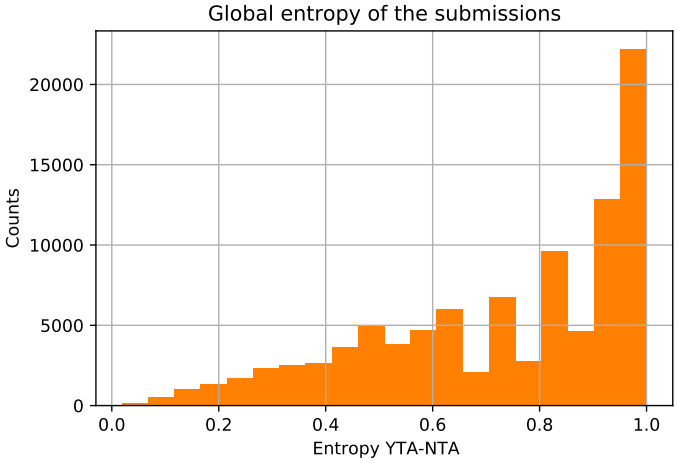


(b) Global entropy without the 0 value

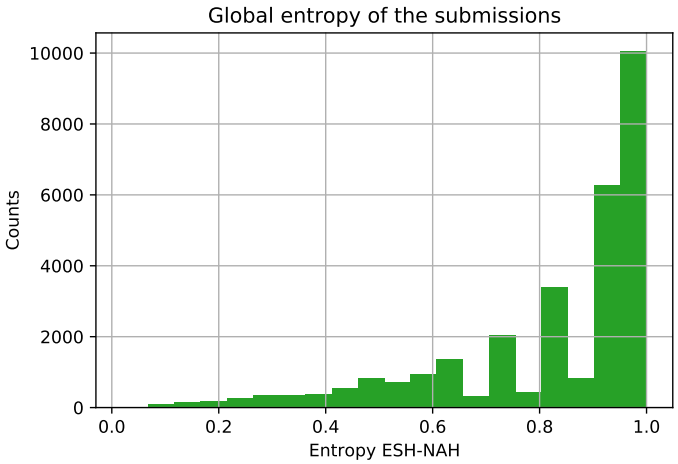
In the first plot all the values of the entropy are taken into account in order to highlight that the majority of the submissions have entropy equal to 0. In the second plot the value 0 of the entropy has been deleted, to show in detail the trend of the entropy in the intermediate range.

Entropy YTA-NTA and NAH-ESH

After the analysis of the global entropy, in this paragraph the "symmetric" judgments YTA-NTA and NAH-ESH are studied. The starting point is the same as the global entropy, with the difference that in this case in the dictionary-like structure holds only the submissions with YTA and NTA (and respectively NAH and ESH) as comments have been maintained. As in the case analyzed before, the logarithmic base has been used for the evaluation of the entropy when the function is applied to the data and the range of the entropy values is from 0 to 1 since the possible outcomes are two. In this case the range does not start from 0 since in the dictionary-like structures have been maintained only the submissions containing both tags. The two plots are shown below:



(a) NTA-YTA Entropy



(b) ESH-NAH entropy

4.3 Topics

In this section, the topics are the aspect under investigation. The reason for investigating this theme relies on understanding how the topic of the submission can influence or not the judgment. The "topic" is defined as the main subject of the submission. Since the subreddit that is the object of this thesis, as it has been already explained in the section 3.2, is mainly focused on private questions, the topics have been named after a specification of the different possible types of private questions. Given that the initial dataset is really large, the categorization of the submissions into different topics has been performed manually. Due to this limit, the number of submissions that have been taken into account is 200. The classified submissions have been extracted randomly from a subset of all the submissions with an entropy value greater than 0.8. This has been done to study the topics of the most controversial submissions, so the ones that led to a less predictable outcome of the judgment from the commenting user. Several tests have been performed in this section to deepen the possible correlations between the distinct topics and the tags. This has been made according to two methods: the binomial test and the multinomial logistic regression test. The data involved in the two examinations are different, therefore the data matrix will be specified at the beginning of the section. Moreover, a preliminary analysis of the data is done using the chi-squared test.

4.3.1 Chi-squared test

The starting dataset for the chi-squared test is the subset of 200 submissions. Each row of the matrix contains the relative count for each tag given and an encoding of the topic into an abbreviation:

- Family: Fa
- Friendship: Fr
- Work: W
- Relationships: R
- Society: S

	Link Id	NAH	NTA	YTA	ESH	Topic
95	c9zdyc	8	6	3	NaN	Fa
15	ak8t9y	NaN	2	4	NaN	W
30	avl9o2	26	78	9	8	R
159	d0go5c	449	1357	76	24	R
186	degzhj	8	878	68	122	S
115	ci551c	2	2	1	1	S
69	bpjseq	7	2	1	NaN	Fa
172	d5asyj	3	8	1	NaN	W
161	d1ctpt	NaN	7	8	NaN	R
45	bfp5a7	6	1	6	NaN	Fr

Table 4.24: Sample of the initial subset of 200 submissions

A first manipulation has been done by aggregating the matrix by topic, obtaining the final matrix used as the starting point for the examination, which is the following contingency matrix:

Topic	NAH	NTA	YTA	ESH
Fa	177	605	305	114
Fr	44	81	56	14
R	1216	3372	450	217
S	181	1332	937	197
W	19	130	22	25

Table 4.25: Contingency matrix for the test

The p-value obtained from this test is equal to 1.601e-245, which is significant at the 0.01 level. This means that there is a correlation between the topic of the submission and the judgment that is given by the user. With the following tests, this result will be deepened.

4.3.2 Binomial test

The starting point for this test is the same as the one mentioned in the table 4.25 with the addition of a column containing the total number of comments:

Topic	NAH	NTA	YTA	ESH	Total number of comments
Fa	177	605	305	114	1201
Fr	44	81	56	14	195
R	1216	3372	450	217	5255
S	181	1332	937	197	2647
W	19	130	22	25	196

Table 4.26: Starting matrix for the binomial test

The binomial test compares the deviations from an expected probability of success with the one of the actual data. This test has been performed for each tag and each topic. It takes as inputs the number of successes, which in this case are the number of comments with that particular tag, the number of trials, which in this case is the total number of comments for that topic and the expected probability. The expected probability is evaluated for each tag from the total dataset (not only the subset of 200 submissions) as the number of times the tag appears over the total number of comments. Since the possible judgments are four the sum of these probabilities is 1.

YTA	NTA	ESH	NAH
0.256	0.574	0.067	0.102

Table 4.27: Probabilities of the tags on the overall dataset

In the subsequent table the resulting p-values of the test are listed:

	Family	Friendships	Relationships	Society	Work
YTA	7.900579e-01	2.474568e-01	**1.522706e-207**	**4.355432e-32**	**0.000002**
NTA	**1.371483e-07**	**5.727778e-06**	**2.976565e-20**	**3.547731e-15**	**0.016996**
ESH	**2.637301e-04**	7.736743e-01	**1.980006e-15**	1.293917e-01	**0.002257**
NAH	**1.085956e-06**	**6.075490e-07**	**6.193387e-161**	**1.498932e-09**	0.906264

Table 4.28: P-values of the binomial test for each tag

There are several p-values (highlighted in bold) that are significant at the 0.001 level. This result highlights that for the pair topic-tag with a significant p-value, the probability of that particular judgment is significantly different from the base one evaluated on the overall data-set. To proceed with the deepening of this aspect, the following step that has been performed is to understand what is the direction of this deviation from the null hypothesis.

Binomial test on the aggregation of tags

To go on with the analysis a manipulation of the data has been done. The four tags, meaning the four types of judgments, can be seen from a more macroscopic point of view. In order to do this, two new categories have been introduced: Positive and Negative. The Positive class (abbreviated with P) and the Negative class (abbreviated with N) collects respectively the collection of NTA plus NAH comments and YTA plus ESH comments. To better investigate the direction of the test performed in the section before, it is valuable to observe the following chart. The chart represents the rate of negative judgments for each topic and compares each specific rate with the generic probability of a negative judgment outcome (marked with a dashed red line). The rate for each topic is computed as the number of comments containing a tag over the total number of comments. The interesting information that is extracted from this graph is that the topics that differ the most from the probability baseline are Relationships and Work in one direction and Society in the opposite direction.

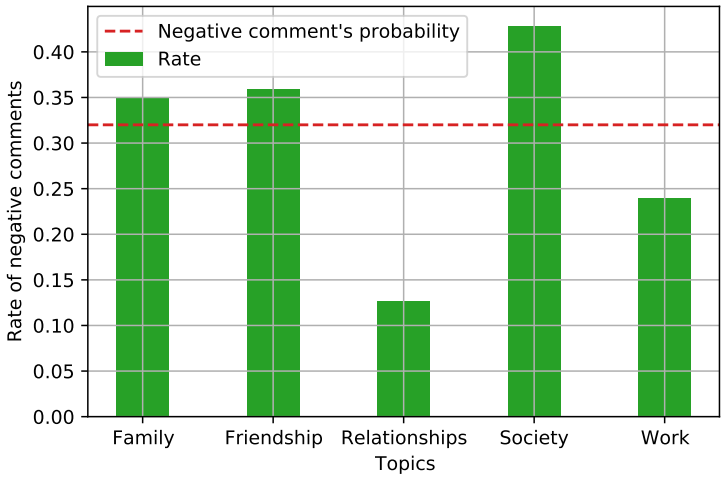


Figure 4.5: Comparison of the topics' rates with the negative probability threshold

To complete the analysis of the data through the binomial test, the latter is performed on the aggregated data. The starting matrix is:

Topic	Total number of comments	N	P
Fa	1201	419	782
Fr	195	70	125
R	5255	667	4588
S	2647	1134	1513
W	196	47	149

Table 4.29: Starting matrix with tag aggregation

The p-values obtained from the test are collected in the following table:

	Family	Friendships	Relationships	Society	Work
N - Negative	0.055	0.284	**2.039e-238**	**8.950e-30**	**0.011**
P- Positive	0.068	0.319	**2.588e-240**	**3.273e-29**	**0.011**

Table 4.30: P-values for the binomial test on aggregated tags

The results of the binomial test performed on the aggregated data are in accordance with what was already predicted by the plot 4.5. Indeed the categories that were mentioned before are the ones that present a p-value significant at the 0.001 level and the graph is used to understand if the deviation means that a certain topic with a significant p-value is more inclined to attract negative judgments or less inclined to attract negative judgments.

4.3.3 Multinomial test

To reproduce correctly the test performed in the previous section the multinomial test has been performed on the whole dataset. The submissions that are not present in the subset of 200 labelled submissions have been categorized with NT (No Topic). The input matrix for the test presents the following form:

	Link Id	Tag	Topic	YTA_Probability	NAH_Probability	ESH_Probability	NTA_Probability
303305	dmlcx3	YTA	NT	0.704	0.032	0.038	0.225
172724	cmrdmb	NTA	NT	0.019	0.013	0.059	0.907
196912	cu1tyy	NTA	NT	NaN	0.190	NaN	0.809
253179	d89vf5	NTA	NT	0.111	0.111	NaN	0.777
274301	devei3	YTA	NT	0.769	0.153	NaN	0.076
16234	apm75b	ESH	NT	0.029	0.102	0.044	0.823
201278	cuwjfz	YTA	NT	0.833	NaN	NaN	0.166

Table 4.31: Input matrix for the multinomial test

From this the Multinomial Logit Regression Test has been performed with the following results:

Dep. Variable:	y	**No. Observations:**	312633
Model:	MNLogit	**Df Residuals:**	312612
Method:	MLE	**Df Model:**	18
Date:	Tue, 03 Aug 2021	**Pseudo R-squ.:**	0.005517
Time:	19:48:09	**Log-Likelihood:**	-3.3599e+05
converged:	True	**LL-Null:**	-3.3785e+05

y=Tag[NAH]	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	2.3979	1.044	2.296	0.022	0.351	4.445
Topic[T. Fa]	-1.8498	1.052	-1.758	0.079	-3.912	0.212
Topic[T. Fr]	-1.2528	1.089	-1.151	0.250	-3.386	0.881
Topic[T. R]	-0.2843	1.047	-0.272	0.786	-2.336	1.767
Topic[T. S]	-2.6307	1.050	-2.506	0.012	-4.688	-0.573
Topic[T. W]	-1.8673	1.118	-1.670	0.095	-4.058	0.324
Topic[T.NT]	-2.0062	1.045	-1.921	0.055	-4.053	0.041
y=Tag[NTA]	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	0.6931	1.225	0.566	0.571	-1.707	3.094
Topic[T. Fa]	0.8961	1.230	0.729	0.466	-1.514	3.306
Topic[T. Fr]	1.0622	1.258	0.844	0.399	-1.404	3.529
Topic[T. R]	2.4836	1.226	2.025	0.043	0.080	4.887
Topic[T. S]	1.1496	1.227	0.937	0.349	-1.255	3.555
Topic[T. W]	1.5041	1.269	1.185	0.236	-0.984	3.992
Topic[T.NT]	1.4486	1.225	1.183	0.237	-0.952	3.849
y=Tag[YTA]	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	2.7081	1.033	2.622	0.009	0.684	4.732
Topic[T. Fa]	-1.5831	1.039	-1.523	0.128	-3.620	0.454
Topic[T. Fr]	-1.3218	1.075	-1.229	0.219	-3.429	0.786
Topic[T. R]	-1.9149	1.036	-1.849	0.064	-3.945	0.115
Topic[T. S]	-1.1938	1.036	-1.153	0.249	-3.224	0.836
Topic[T. W]	-2.0149	1.103	-1.827	0.068	-4.177	0.147
Topic[T.NT]	-1.3683	1.033	-1.325	0.185	-3.393	0.656

Table 4.32: MNLogit Regression Results

Dep. Variable:	y	**No. Observations:**	312633
Model:	MNLogit	**Df Residuals:**	312626
Method:	MLE	**Df Model:**	6
Date:	Tue, 03 Aug 2021	**Pseudo R-squ.:**	0.007436
Time:	20:33:02	**Log-Likelihood:**	-1.9446e+05
converged:	True	**LL-Null:**	-1.9591e+05

y=Tag[NA]	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	-0.207	0.373	-0.556	0.578	-0.939	0.524
Topic[T. Fa]	0.693	0.379	1.831	0.067	-0.049	1.435
Topic[T. Fr]	0.787	0.402	1.958	0.050	-0.001	1.576
Topic[T. R]	2.514	0.375	6.702	0.000	1.779	3.250
Topic[T. S]	0.455	0.375	1.213	0.225	-0.280	1.191
Topic[T. W]	1.479	0.427	3.466	0.001	0.643	2.316
Topic[T.NT]	0.937	0.373	2.510	0.012	0.205	1.669

Table 4.33: MNLogit Regression Results

4.4 Comments' authors

The main purpose of this section is to analyze and investigate the relationship between the authors of the comments (those who give the judgment), the age and the gender of the poster and, mainly, the previous history of the authors. The previous history of the authors is intended as the subreddits in which they are the most active, i.e. the subreddits in which they post the most. The analysis is then performed by means of machine learning methods such as Random Forest and Logistic Regression.

4.4.1 Selection of the authors and of the subreddits

The starting point of the analysis is the creation of the initial matrix containing the data of interest used as an input for the supervised learning algorithms. The input matrix must hold three categories of information:

- Judgment that the author gave to the poster
- Frequented subreddits of the author
- Age and gender of the poster in the initial subreddit

The principal step in the creation of the input matrix concerns the collection of the subset of the subreddits of interest. Due to the huge amount of data not all the authors have been investigated but only the most active ones: as it has been already mentioned in section 3.1.2 the average number of comments per author is 10, so in the following, the

authors with more than 15 comments will be considered active. The subset of authors with more of 15 comments is the 10% of the total number of authors that commented in the subreddit. After the derivation of the subset of the authors, their previous history is researched in the initial dataset (Pushshift Reddit Dataset) and is extracted as a list of Author - Subreddit - Number of submissions in that subreddit. The first matrix contains the Link Id (always needed in order to merge the data), the author and the tag:

	Link Id	Author	Tag
2747301	t3_by70mb	originalthaerun	NTA
2710650	t3_bxhld6	t4ctic4lc4ctus	NTA
4448976	t3_asxd52	jcmclovin	NTA
1856008	t3_dpdwko	Director_Tseng	NTA
1912723	t3_c8q960	jewishdaughter	ESH
5184957	t3_9z7ztl	areulistening	NTA
3754791	t3_d07c6b	videobrat	NTA

Table 4.34: Matrix

The second matrix contains the author, the name of the subreddit in which he/she posted and the number of submissions done by the author in that particular subreddit.

	Author	Subreddit	Number of submissions
885092	Konjonashipirate	GradSchool	1
160061	Ouma_Shu	ProjectFi	2
292722	Azothlike	Overwatch	1
845750	RubberDuckHuh	piercing	1
1333265	SarahVen1992	cats	1
596187	UpsetMuffins	Supercorped	2
884162	Dyna_Sean	betterCallSaul	2

Table 4.35: Matrix

The additional action that is done to this matrix is the selection of the most active subreddits for this set of authors. To do this the data is aggregated on the subreddits and summed to obtain the total number of submissions done by the subset of authors. The subreddits have then been arranged in descending order:

Subreddit	Number of submissions
AskReddit	159118
Frei_Donald	44256
Showerthoughts	36618
aww	25569
NoStupidQuestions	20111
funny	20029
The_Donald	19342
memes	19199
unpopularopinion	18930
teenagers	18721

Table 4.36: First seven rows of the matrix

From this final form, the 1000 most active subreddits have been extracted and this subset of subreddits will be used as part of the features in the following analysis. To proceed with the analysis it is necessary to observe that, being the subreddits categorical variables, they need to be represented by means of one-hot-encoding: each row of the matrix represents an author and each column represents one of the 1000 selected subreddits. The cell is then marked with a 1 if the author posted in that subreddit and 0 else. This results in having a matrix of dimension 109044 (number of authors) rows $\times$ 1000 (number of subreddits) columns. The final form of this matrix is:

	AskReddit	Frei_Donald	Showerthoughts	aww	NoStupidQuestions	funny	The_Donald	memes
———V———	0	0	0	0	0	0	0	0
–0mn1-Qr330005–	0	0	0	0	0	0	0	0
–BMO–	1	0	1	0	0	1	0	0
–Darkrai–	1	0	0	0	0	0	0	0
–Eug–	0	0	0	0	0	0	0	0
–Leilani–	0	0	0	0	0	0	0	0
–Replicant–	0	0	0	0	0	0	0	1
–THOT-PATROL–	1	0	0	0	0	0	0	0

Table 4.37: First 8 rows and 8 columns of the matrix

To obtain the input matrix it is still necessary to add the information concerning the author of the submission, i.e, age and gender. These information were already contained in the tables used for the previous analysis (e.g. Table 3.2). At this point the final step is to merge all the necessary data from each one of the tables mentioned into a unique matrix. This is done by merging on the author's name. The tags have been first converted into the two general judgments: positive and negative. Both on the tag feature (P or N) and on the gender feature (F or M) the one-hot encoding has been

applied. The age feature has been kept continuous and has been scaled between 0 and 1 according to the feature range. In the final matrix each row represents a comment and each column represents a feature: in conclusion, the features that will be used are the gender, the age, the given tag and the frequented subreddits. The input matrix, in the end, has dimension 185244 rows $\times$ 1005 columns and this form:

	Gender_ 'F'	Gender_ 'M'	Tag_ N	Tag_ P	Age	AskReddit	Frei_Donald	Showerthoughts	aww	NoStupidQuestions
allenidaho	1	0	0	1	20	1	0	0	0	0
LoriTheGreat1	0	1	1	0	22	1	0	0	1	1
Jumbajukiba	0	1	0	1	24	0	0	0	0	1
boringandsleepy	1	0	0	1	19	0	0	0	0	0
lionheart059	1	0	0	1	22	0	0	0	0	0
saidsatan	0	1	1	0	22	0	0	0	0	0
lmp515k	1	0	1	0	18	0	0	1	0	0
Alias_X_	0	1	0	1	15	1	0	0	0	0

In the following different methods will be applied to this input matrix. The final goal is to predict if a user, given the features that are the gender and the age of the author and his/her previous history, what will be the judgment that he/her will give. Will the judgment be positive or negative? Will the user agree or disagree with the author of the submission?

4.4.2 Random forest

In the following subsection, the random forest classifier has been implemented with several combinations by changing parameters like the number of features with different feature selection procedures. A random forest is a classifier consisting of a collection of decision trees, where each tree is constructed by applying an algorithm A on the training set S and an additional random vector θ, where θ is sampled i.i.d. from some distribution. The prediction of the random forest is obtained by a majority vote over the predictions of the individual trees. [13] Given the classifier and the training set, there are four possible outcomes, that are: *true positive* if the instance is positive and is classified as positive, *false negative* if the instance is positive and classified as negative. If the instance is negative and is classified as negative, it is a *true negative*, if the instance is negative and is classified as positive, it is a *false positive*. These four values form the confusion matrix, which is the basis for many metrics. The metrics that will be reported to evaluate the performance of each implementation are:

- **Precision**: The number of positives correctly classified over the number of true positives plus false positives.
- **Recall**: The number of positives correctly classified over the total number of positives.

- **Balanced accuracy**: Defined as

$$\text{Balanced accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{4.12}$$

  It is used in this case since the dataset is unbalanced, because it contains 70% of the samples marked as "Positive" and 30% of the samples marked as "Negative".

- **Matthews Correlation Coefficient**: in the case of binary vectors, the correlation coefficient is defined as

$$C = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \tag{4.13}$$

  It is always between -1 and +1. A value of -1 indicates total disagreement and +1 total agreement. It is equal to 0 for completely random predictions. This means that if two variables are independent, their correlation coefficient is 0. [4, 13, 6]

The random forest classifier is applied on the input matrix data, the whole dataset is used. The parameters set for the classifier are a number of estimators equal to 100 and the class weights are set to balanced, meaning that the weights are adjusted inversely proportional to class frequencies in the input data. The input data is divided into training and test set, with the test set being the 25 % of the whole dataset. The shape of the test and training set is:

- Training shape: (138750, 1003)
- Test shape: (46250, 1003)

The metrics of this classifier are:

	Precision	Recall
Positive	0.839	0.848
Negative	0.662	0.646

Balanced accuracy	0.747
MCC	0.498

Table 4.38: Parameters for the Random Forest

Due to the huge number of features that are present in the input matrix, different techniques of feature selection and dimensionality reduction have been studied.

Random forest with sequential feature selection

Sequential feature selection is one of the most common sequential search algorithms. In this analysis the Forward Sequential Selection is applied. The algorithm begins with zero attributes, evaluates all feature subsets with exactly one feature and selects the one with the best performance. It then adds to this subset the feature that yields the best performance for subsets of the next larger size. This cycle repeats until the number of selected features desired [3]. Due to computational limits, the number of samples used as an input for the selector has been reduced. The number of samples used is 1000. On this reduced set of samples, the selector has been applied and the number of features to select has been set equal to 10. The final shape of test and training set is equal to:

- Training shape: (750,10)
- Test shape: (250,10)

The metrics of the classifier are:

	Precision	Recall
Positive	0.71	0.99
Negative	0.71	0.07

Balanced accuracy	0.52

Table 4.39: Metrics of the random forest with SFS

The two following methods are part of the univariate statistics methods. It is computed whether there is a statistically significant relationship between each feature and the target. Then the features that are related with the highest confidence are selected. A key property of these tests is that they are univariate, meaning that they only consider each feature individually. Consequently, a feature will be discarded if it is only informative when combined with another feature. These methods for discarding parameters use a threshold to discard all features with too high a p-value (which means they are unlikely to be related to the target). The methods differ in how they compute this threshold [14].

Random forest with SelectKBest

The method used in this paragraph for the feature selection is the "SelectKBest". It selects a fixed k number of features according to the highest scores. As mentioned before, this test selects k features according to their p-value. The fixed number k has

been choosen equal to 10. In this case all the dataset has been used and the selected features are:

Gender F
Gender M
worldnews
medical
BigBrother
tipofmytongue
nosleep
JUSTNOMIL
entitledparents
thebachelor

The metrics of this classifier are:

	Precision	Recall
Positive	0.735	0.555
Negative	0.369	0.566

Balanced accuracy	0.561
MCC	0.113

Table 4.40: Metrics of the random forest with SelectKBest

Random forest with family-wise error rate

The method used in this paragraph for the feature selection uses univariate statistical tests for each feature: in this case the family wise error rate. The FWER is the probability of having one or more rejection of a null hypothesis. Then, a threshold value is applied and only the features with a p-value below the threshold are kept. The algorithm is applied on the totality of the dataset and 116 features have been selected (out of the 1003). The final shape of test and training is:

- Training shape: (138750, 116)
- Test shape: (46250, 116)

The metrics obtained are:

	Precision	Recall
Positive judgment	0.817	0.724
Negative judgment	0.521	0.648

Balanced accuracy	0.700
MCC	0.355

Table 4.41: Metrics of RF classifier with FWER

The Receiver Operator Characteristic curve has then been plotted. The ROC curve is a two-dimensional depiction of classifier performance. A method to reduce ROC performance to a single scalar value is to calculate the area under the ROC curve, which is the AUC value. The AUC value will always be between 0 and 1, however no realistic classifier should have a value less than 0.5. The AUC value of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. The diagonal line y = x represents the strategy of randomly guessing a class.

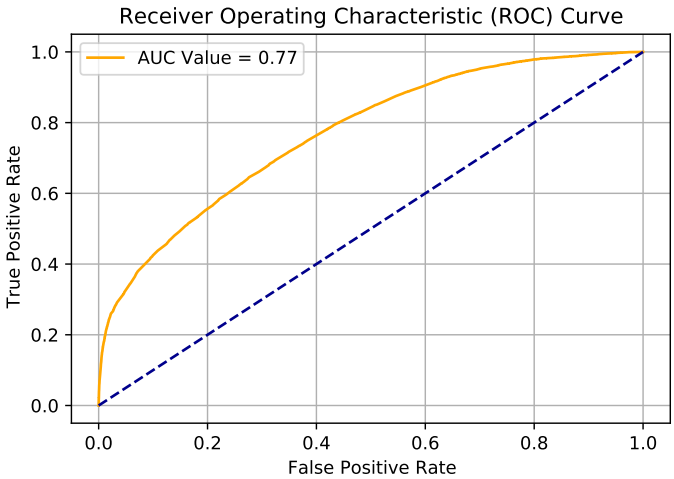


Figure 4.6: ROC plot

To better understand the contribution of each feature to the final model, e.g., how much each feature contributes to predicting whether a comment is positive or negative the Shap values have been plotted. Shapley values are used to understand how the prediction is distributed among the features, and in particular the feature importance. The SHAP Summary Plot contains both feature importance and feature effects. Each point on the summary plot is a Shapley value for a feature and an instance. The position

on the y-axis is determined by the feature and on the x-axis by the Shapley value. The color represents the value of the feature from low to high, the features are ordered according to their importance. In this case, due to computational limits the Random Forest classifier has been modified to a maximum depth equal to 8 and the SHAP values have been evaluated and plotted of the 3 % of the training set.

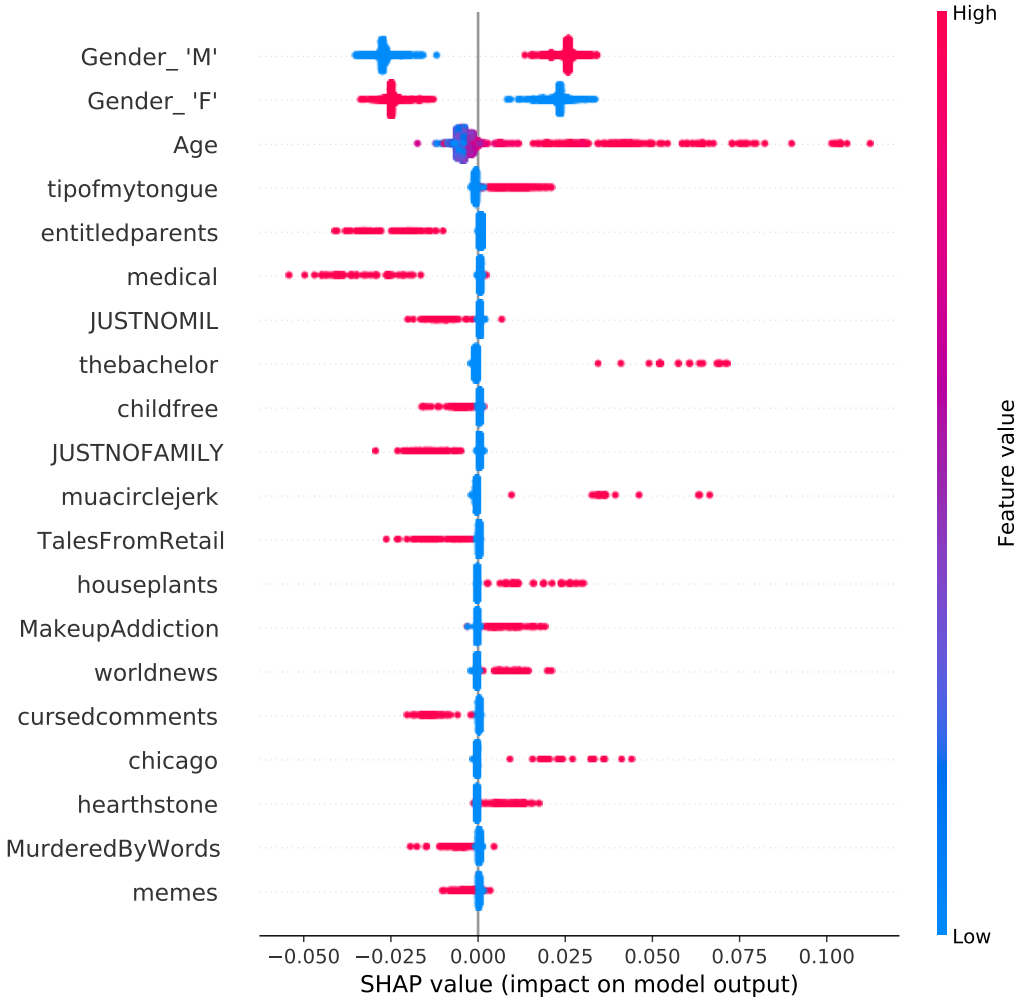


Figure 4.7: SHAP Values

Nested cross validation in Random forest

In this section an additional topic is investigated, by implementing nested cross validation on the dataset. In the nested cross validation procedure multiple splits of cross-validation are done. This process is called "nested" because an outer loop is defined over splits of the data into training and test sets. Then, for each of them, a grid search is run returning best parameters. What grid search does is tuning the parameters of the model by trying all possible combinations of the parameters of interest. In this case, for the Random

Forest classifier, the parameters were the number of estimators (number of trees in the forest), the maximum depth of the tree and the minimum number of samples required to split an internal node. At this point, for the outer split the test set score using the best settings is obtained. The aim of this procedure is to obtain a score (the best parameters set is obtained by performing only cross validation on the outer loop). This procedure does not provide a model but it is useful for evaluating the performance of a model on a given set of data [14]. This process has been applied both on the random forest with FWER and SelectKBest methods for feature selection.

- **FWER** Due to computational limits the highest number of samples that can be used is 60000. FWER feature selection has been applied and the number of selected features is 12. The cross-validation is performed with 5 splits and the obtained score difference, which is evaluated as the difference between the best score of the classifier performed on the inner loop and the cross validation performed on the outer loop with parameter optimization. The best parameters set is:

Maximum depth	None
Minimum samples split	5
Number of estimators	500

  In this case the value obtained is 0.002, meaning that we are not having overfitting. The final shape of test and training is:

  - Training shape: (45000, 12)
  - Test shape: (15000, 12)

  The parameters obtained are:

	Precision	Recall
Positive	0.753	0.592
Negative	0.397	0.581

Balanced accuracy	0.586
MCC	0.162

Table 4.42: Parameters obtained with cross-validation and FWER

- **SelectKBest**: Due to computational limits, also in this case, the highest number of samples that can be used is 60000. The number of K features to select is equal

to 10 with the chi-squared test. The cross-validation is performed with 5 splits and the obtained score difference, which is evaluated as the difference between the best score of the classifier performed on the inner loop and the cross validation performed on the outer loop with parameter optimization. The best parameters set is:

Maximum depth	5
Minimum samples split	5
Number of estimators	100

In this case the value obtained is 0.0006 so we can conclude that also in this test we are not having overfitting. The final shape of test and training is:

- Training shape: (45000, 10)
- Test shape: (15000, 10)

The parameters obtained are:

	Precision	Recall
Positive	0.739	0.559
Negative	0.376	0.574

Balanced accuracy	0.567
MCC	0.124

Table 4.43: Parameters obtained with cross-validation and SelectKBest

KFold

In this subsection, to accurately test the performance of the classifier, the K-Fold Cross-Validation has been implemented. In this method the data is split repeatedly and multiple models are trained; the data is partitioned into ten parts of equal size and then a sequence of models is trained. The first model is trained using the first fold as the test set, and the remaining folds (2-10) are used as the training set. The model is built using the data in folds 2-10 and the accuracy is evaluated on fold 1. This process is then repeated for each of the ten splits of data and each time the accuracy is computed. When using 10-fold cross-validation, the model is fitted with nine-tenths of the data, resulting in a higher accuracy [14]. In the following analysis the whole dataset has been used. It has been divided into ten folds and a Random Forest classifier with 100 estimators and

balanced class weights has been employed. Then, the feature selector with family-wise error rate has been applied to each fold, and the classifier has been fitted on the train fold modified with the selected features. The average number of selected features is 62. The balanced accuracy of the cross validation is 0.76 with a standard deviation of 0.003. Furthermore, the ROC curve has been plotted in order to visualize the performance of the classifier. In the following plot it can be seen that it is not present a substantial difference between the different ROC curves for each fold. Indeed, the mean AUC value is 0.85 with a standard deviation of 0.004.

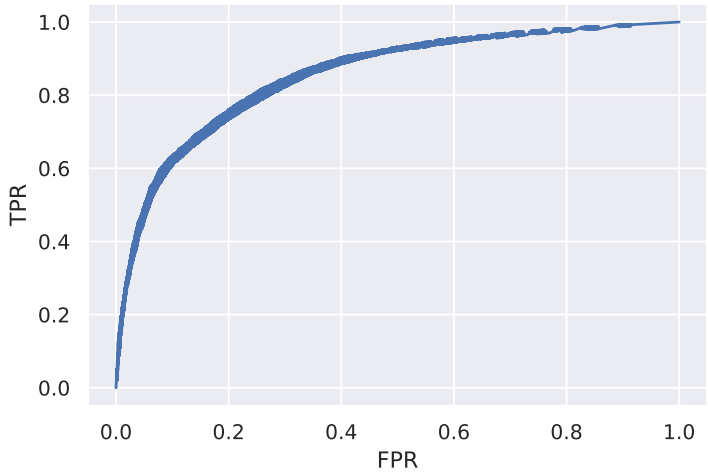


Figure 4.8: ROC plot

Principal Component Analysis (PCA)

In this section the Principal Component Analysis has been applied. The analysis did not bring significant results but they will be reported for the sake of completeness. Principal component analysis is a method that rotates the dataset in a way such that the rotated features are statistically uncorrelated. This rotation is often followed by selecting only a subset of the new features, according to how important they are for explaining the data [14]. First of all, the PCA has been applied on the dataset selecting a number of components equal to 1000 (all the features) to evaluate the cumulative sum explained variance. Then, the PCA has been performed with 300 components; the results are not significant since the first component is the gender. The plot below shows the second and the third principal component:

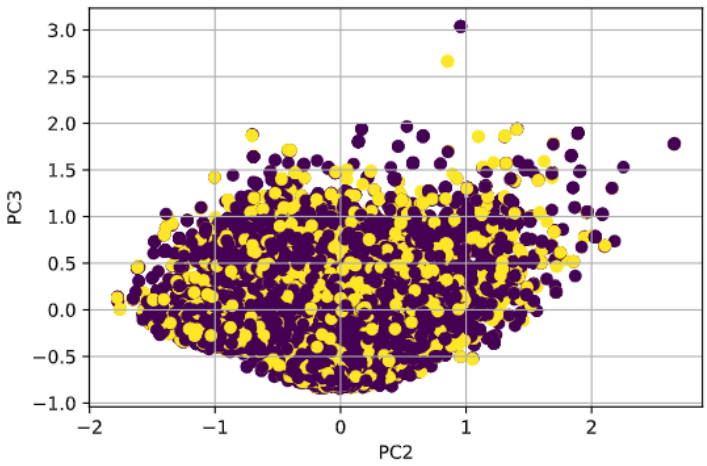


Figure 4.9: Scatter plot PCA2 vs PCA3

4.4.3 Logistic regression

Logistic regression is a linear classification algorithm, which is a special case of a Generalized Linear Model. It is introduced h(x) as the probability that the label of x is 1. The sigmoid function used in logistic regression is the logistic function which is defined as [13]:

$$\phi_{sig}(z) = \frac{1}{1 + exp(-z)} \tag{4.14}$$

The variable that will be predicted is the "Negative" one. Due to computational limits, in this analysis, the largest sample that can be used for the logistic regression is made up of 50000 rows that are extracted randomly from the input matrix. All the features are used (1003). The input data is split into training and test, with the test size choose as 25 % of the training. The final shape of test and training is:

- Training shape: (37500, 1003)
- Test shape: (12500, 1003)

The metrics of this classifier are:

	Precision	Recall
Positive	0.686	0.959
Negative	0.450	0.071

Balanced accuracy	0.515
MCC	0.064

Table 4.44: Metrics of the Logistic Regression

We can notice that in this case the performance of the classifier are worsened so it is not considered as a good prediction.

Chapter 5

Conclusions

In the end, several correlations between the given judgment and the different features are extracted, outlining some interesting behaviours of users online. The interpretation of the data reveals a negative bias in the moral judgment of a male user, meaning that the online community is more inclined to give harsher penalties to them. This first result is in accordance with what was already predicted by literature: two identical transgressive situations will be judged in a different way based on the gender of the parties implicated [22]. Furthermore, as we have investigated, the moral opinion of a user about a situation is influenced by the topic: if the main character is involved in a work situation or a romantic relationship situation and so on. In the following the single results are listed and explicitly stated.

Correlation between the received judgment and the age of the poster

The chi-squared test performed between tags and age bins reports a p-value significant at the 0.01 level leading to the conclusion that there is a correlation between the age of the poster and the judgment he/she receives. The correlation of the age is then further investigated in each age bin, where it results to be significant at the 0.05 level in the age bin 21-23 and 27-70. This leads us to the conclusion that the moral opinion about a user in this age range could be influenced by his/her age. Then, by performing the binomial test we obtained the in which way the judgment is influenced: the "older" users have a lower probability of receiving a negative judgment.

Correlation between the received judgment and the gender of the poster

In the investigation of this correlation several results have been achieved. The chi-squared test performed between tags and gender provides a p-value significant at the 0.01 level reporting that there is a correlation. These results are then confirmed and additionally investigated through the binomial regression test which reports one of the most important results: a male user has a higher probability of receiving a negative judgment.

Correlation between the topic of the submission and the judgment

The first result achieved in this section of the study is the one obtained from the preliminary chi-squared test, which reports that there is a correlation between the topic and the moral opinion aroused by the submission. Then, by performing the binomial test it has been obtained another important result that is extracted from the figure 4.5: comparing each topic with the baseline of negative comments' probability we can observe that three topics in particular differ significantly from it. The topic "Society", which covers all the submissions about political questions or racism or gender questions, has a higher probability (with respect to the baseline) of raising a negative moral opinion. On the other side, the topics like Relationships and Work tend to receive a rate of negative comments lower than the average.

Prediction algorithm: results and metrics

The last result is achieved from the prediction algorithm: the best metrics are obtained by means of K-Fold Cross Validation combined with Sequential Feature Selection resulting in a balanced accuracy of 0.76 and an average number of selected features of 62. From this, we can conclude that the features represented by the previous history of the user on the social media constitute a good basis to predict his/her judgement, meaning that, as in real-life the social norms are defined also by the cultural background of the individual.

Bibliography

[1] Jon Elster. "Social norms and economic theory". In: *Journal of economic perspectives* 3.4 (1989), pp. 99–117.

[2] Alan Agresti. *An introduction to categorical data analysis*. New York: Wiley, 1996. URL: http://www.worldcat.org/search?qt=worldcat_org_all&q=0471113387.

[3] David W. Aha and Richard L. Bankert. "A Comparative Evaluation of Sequential Feature Selection Algorithms". In: *Learning from Data: Artificial Intelligence and Statistics V*. Ed. by Doug Fisher and Hans-J. Lenz. New York, NY: Springer New York, 1996, pp. 199–206. ISBN: 978-1-4612-2404-4. DOI: 10.1007/978-1-4612-2404-4_19. URL: https://doi.org/10.1007/978-1-4612-2404-4_19.

[4] Pierre Baldi et al. "Assessing the accuracy of prediction algorithms for classification: An overview". In: *Bioinformatics (Oxford, England)* 16 (June 2000), pp. 412–24.

[5] Karl-Dieter Opp. "When do norms emerge by human design and when by the unintended consequences of human action? The example of the no-smoking norm". In: *Rationality and Society* 14.2 (2002), pp. 131–158.

[6] Tom Fawcett. "An introduction to ROC analysis." In: *Pattern Recognit. Lett.* 27.8 (2006), pp. 861–874. URL: http://dblp.uni-trier.de/db/journals/prl/prl27.html#Fawcett06.

[7] Werner Güth and Stefan Napel. "Inequality aversion in a variety of games–an indirect evolutionary analysis". In: *The Economic Journal* 116.514 (2006), pp. 1037–1056.

[8] Sarita Yardi and Danah Boyd. "Dynamic debates: An analysis of group polarization over time on twitter". In: *Bulletin of science, technology & society* 30.5 (2010), pp. 316–327.

[9] Daegon Cho, Soodong Kim, and Alessandro Acquisti. "Empirical analysis of online anonymity and user behaviors: the impact of real name policy". In: *2012 45th Hawaii international conference on system sciences*. IEEE. 2012, pp. 3041–3050.

[10] Jesse Graham et al. "Moral foundations theory: The pragmatic validity of moral pluralism". In: *Advances in experimental social psychology*. Vol. 47. Elsevier, 2013, pp. 55–130.

[11] Talcott Parsons. *The social system*. Routledge, 2013.

[12] Eyal Sagi and Morteza Dehghani. "Measuring moral rhetoric in text". In: *Social science computer review* 32.2 (2014), pp. 132–144.

[13] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. USA: Cambridge University Press, 2014. ISBN: 1107057132.

[14] A. Müller and Sarah Guido. "Introduction to Machine Learning with Python: A Guide for Data Scientists". In: 2016.

[15] Katja Rost, Lea Stahel, and Bruno S Frey. "Digital social norm enforcement: Online firestorms in social media". In: *PLoS one* 11.6 (2016), e0155923.

[16] Molly J Crockett. "Moral outrage in the digital age". In: *Nature human behaviour* 1.11 (2017), pp. 769–771.

[17] George C Homans, A Paul Hare, and Richard Brian Polley. *The human group*. Routledge, 2017.

[18] Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. "Trouble on the horizon: Forecasting the derailment of online conversations as they develop". In: *arXiv preprint arXiv:1909.01362* (2019).

[19] Subhabrata Dutta, Dipankar Das, and Tanmoy Chakraborty. "Changing Views: Persuasion Modeling and Argument Extraction from Online Discussions". In: (July 2019).

[20] Jason Baumgartner et al. "The Pushshift Reddit Dataset". In: *Proceedings of the International AAAI Conference on Web and Social Media* 14.1 (May 2020), pp. 830–839. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7347.

[21] Maxwell Forbes et al. "Social Chemistry 101: Learning to Reason about Social and Moral Norms". In: (Nov. 2020).

[22] Tania Reynolds et al. "Man up and take it: Gender bias in moral typecasting". In: *Organizational Behavior and Human Decision Processes* 161 (2020), pp. 120–141.

[23] Nicholas Botzer, Shawn Gu, and Tim Weninger. "Analysis of Moral Judgement on Reddit". In: (Jan. 2021).

[24] Pew Research Center. *Social Media Use in 2021*. 2021. URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/.

[25] Nicholas Lourie, Ronan Le Bras, and Yejin Choi. *Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-Life Anecdotes*. 2021. arXiv: 2008.09094 [cs.CL].