TU WIEN Informatics

Klassifizierung von Benutzern in Online News Foren

eine Datenanalyse der Benutzertypen und Benutzerinteraktionen des Forum einer österreichischen Tageszeitung

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Business Informatics

eingereicht von

Felix Scholz, BSc
Matrikelnummer 01126314

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner
Mitwirkung: Mag.rer.nat. Dr.techn. Julia Neidhardt

Wien, 25. September 2023

Felix Scholz	Hannes Werthner

TU WIEN Informatics

Classification of users in an online news forum

a data analysis of user types and user interaction in the online forum of an Austrian newspaper

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Business Informatics

by

Felix Scholz, BSc
Registration Number 01126314

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Hannes Werthner
Assistance: Mag.rer.nat. Dr.techn. Julia Neidhardt

Vienna, September 25, 2023

Felix Scholz | Hannes Werthner

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Felix Scholz, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 25. September 2023

Felix Scholz

Danksagung

Zum Anfang dieser Diplomarbeit möchte ich mich bei allen Personen bedanken, die mich über den langen Zeitraum der Durchführung unterstützt haben und mir durch Motivation geholfen haben diese abzuschließen.

Mein besonderer Dank gilt Univ.Ass.in Mag.a rer.nat. Dr.in techn. Julia Neidhardt, die mir als Ko-Betreuerin über einen ausgesprochen langen Zeitraum mit Rat und Enthusiasmus für das Projekt eine unschätzbare Hilfe war und dabei nie die Geduld verloren hat. Ihre Expertise in den wissenschaftlichen Bereichen der Datenanalyse und des wissenschaftlichen Arbeitens waren, neben ihren persönlichen Eigenschaften, die notwendigen Stützen um diese Arbeit zu vollenden.

Für die Möglichkeit an diesem einzigartigen Projekt forschen zu können und die Arbeit zu betreuen, bedanke ich mich bei Univ.-Prof. Dipl.-Ing. Dr.techn. Hannes Werthner. Seine Hinweise zum strukturellen Aufbau, waren ausschlaggebend um, auch über einen längeren Zeitraum, stets wieder zurück in den Schreibprozess zu finden.

Weiterer Dank gilt den Mitgliedern der Gruppe der Diplomandinnen und Diplomanden, die mich in unseren regelmäßigen Jour fixe Treffen, nicht nur mit wertvollem Feedback, sondern auch moralisch unterstützt haben und mich durch konstruktive Gruppendiskussionen zu besseren Leistungen angespornt haben.

Auch an den Standard gilt mein Dank, für die Bereitstellung der Daten, die einen beindruckenden Einblick in die Geschichte der täglichen Online-Nachrichten gewährleistet haben, die nicht vielen Diplomandinnen und Diplomanden zur Verfügung steht.

Zuletzt, mein großer Dank an Kathi und Maxi, die auf der einen Seite ausdauernd an die baldige Fertigstellung geglaubt haben und zeitgleich, auf ihre Art, den Weg mitgegangen sind. Einerseits indem sie sich selbst ausreichend Zeit zum Schreiben erlaubt haben, um mir die nötige Gesellschaft zu leisten, andererseits in dem sie mich erinnert haben, dass mein Ziel schon vor meinen Augen lag.

Acknowledgements

At the beginning of this master thesis I want to thank all people who helped me in the writing process, and who motivated me to complete this work.

I would particularly like to thank Univ.Ass.in Mag.a rer.nat. Dr.in techn. Julia Neidhardt, who assisted me as co-advisor for a decidedly long period of time, with guidance and enthusiasm for the project, and who has been an invaluable help while never losing patience. Her expertise in data analysis and the academic process were, next to her personal traits, the necessary support to finish this thesis.

I thank Univ.-Prof. Dipl.-Ing. Dr.techn. Hannes Werthner for the opportunity to carry out research on this exceptional project and for him to supervise this thesis. His pointers on the structural composition of this work, were crucial for me to find my way back into the writing process, even after longer breaks.

Further thanks are directed towards the members of the group of diploma students, who helped me with invaluable feedback and moral support in our regular jour fixe meetings, and who inspired me through constructive discussions to keep improving my work.

Additionally, I want to thank the Standard, for providing me with data, that allowed me an impressive view into the history of daily news, which not many diploma students are given.

Finally, a tremendous thanks to Kathi and Maxi, who on one side kept believing with perseverance in my imminent completion of this work, while simultaneously keeping me company. On one hand by writing alongside me for as long as possible, and additionally by reminding me of how close to my goal I already was.

Kurzfassung

Online News Foren bieten Nutzern die Möglichkeit sich mit anderen Nutzern zum aktuellen Tagesgeschehen auszutauschen. Diese, oftmals sehr großen Plattformen bilden einen Querschnitt der Bevölkerung ab, mit unterschiedlichen Meinungen und Motivationen. In dem wiederkehrende Verhaltensmuster durch Klassifizierung in einem derartigen Kontext erkennbar gemacht werden, soll die Analyse der Zusammensetzung von selbst umfangreichen Nutzerbasen ermöglicht werden. Darüber hinaus können Untersuchungen mit Langzeitdaten Einblicke in die Zusammensetzung und Entwicklung im Verlauf der Zeit bieten.

Diese Diplomarbeit erstellt ein Klassifizierungsmodell für Benutzer eines Online News Forums. Das Model entsteht durch die Kombination einer explorativen sowie einer statistischen Datenanalyse und versucht die wiederkehrenden Verhaltensmuster in solch einem Umfeld zu bestimmen. Ein derartiges Modell kann dazu benutzt werden die Entwicklung von großen Benutzerbasen zu analysieren und ermöglicht es schnell einen Überblick über die Zusammensetzung zu schaffen.

In dem Versuch das Modell allgemeingültig zu halten, ohne die Ergebnisse zu stark an die verwendeten Daten zu koppeln, werden sechs Rollen für aktiv an der Diskussion teilenehmende Benutzer, sowie eine zusätzliche Rolle für nicht teilnehmende Leser, vorgeschlagen. Die aktiven Rollen sind: Taciturn, Silent Voter, Regular, Conversationalist und Celebrity. Die inaktive Rolle heißt Lurker. Die Leistungsfähigkeit des Modells wurde getestet und seine Vorhersagekraft erreicht einen F1 Wert von 0.8632.

Auf Langzeitdaten, die von derStandard.at bereitgestellt wurden, angewandt, wurde die Rollenverteilung über einen längeren Zeitraum analysiert, welche einen kontinuierlichen Trend zu aktiv am Forum teilnehmenden Rollen beschreibt. Darüber hinaus wurde bestimmt welche Rollen gemeinsam im Langzeitverhalten von Benutzern auftreten, sowie die Frequenz von Rollenwechseln, um zu evaluieren, ob Benutzer einzelnen Rollen zugehören oder Anzeichen von unterschiedlichen Rollen zeigen, die von Faktoren wie Zeit oder Kontext abhängig sein können.

Abstract

Online news forums provide a longstanding way for the exchange of opinions with other users. These often-large user bases consist of a cross-section of all people with various opinions and motivations. By classifying recurring types of behaviors common in such a context, an effort is made to provide an understanding about the composition of even large user bases. Additionally, long-time observations could provide insights into the composition and development over time.

This thesis creates a classification model for users of an online news forum. The model is created by combining exploratory and statistical data analysis and attempts to explain recurring behaviors found in such an online community context. Such a model can be used to analyze how the user base of a large community develops and provides a quick overview of the composition of its users.

In order to keep the model generally applicable, and not too tied to the provided data, the thesis proposes six roles for actively participating users and one additional role for non-participating readers. These active roles are Taciturn, Silent Voter, Regular, Conversationalist, Power User, and Celebrity. The inactive role is called Lurker. The model performance was tested for its predictive power and achieved a macro F1 score of 0.8632.

Applying the model to a set of long-term data, provided by the online news forum of derStandard.at, role distribution over time was analyzed, showing a gradual trend towards higher activity of forum users. Additionally, co-occurrences of roles in the long term behavior of users and the frequency of role switches were measured in order to evaluate whether users have inherent roles or show signs of various roles, which could be dependent on time or context.

Contents

CHAPTER 1

Introduction

1.1 Problem definition

Online communities have long been a place where people interested in similar topics gather[Hag99]. News forums, especially, are a popular [DN11] and a direct way for readers to voice their opinions concerning events of the day. On these, they can comment on the articles and surrounding topics. Each user is interested in different topics and news categories, which might reflect in their reading and commenting behaviour. While some users may only care for politics-related information, others could only be interested in sports. In the same manner, users have varying ways and degrees of voicing their opinions and interacting with other users. Their postings can be insightful or personal, but also harassing or spam. Whereas the former can be of interest to the news publishers, in terms of pursuing and encouraging such behaviour, the latter poses a risk to the forum culture and thus the publishers. It can even require them to take actions, be it from excluding certain users from the forum to taking legal action against authors of such postings.

Besides toxic behaviour smaller scale negativism can also have an impact on users and their future communication behaviour, such as negative evaluations of a user's content. A recent study [CDNML15] suggests that negative evaluations of a user's postings can decrease the quality of future postings of the same user, creating a negative feedback loop. Moreover, the same user is more likely to evaluate other content negatively, further harming the surrounding discussion environment.

This highlights the demand for a deeper understanding of who these users are and how their behaviour, positive and negative, can shape the community of such forums. Whereas moderators can directly intervene in discussions and take cautionary measures from warning users up to excluding them from the forum altogether, they face scaling issues as forums grow larger, favouring an automated or at least semi-automated process to detect these users of special interest.

Possible changes in the behaviour of users over time are another aspect, in which automated processes have an advantage over human moderators. Behaviour and opinions can change over time. Mood, which varies by the time of the day [GM11], can also have an impact on current behaviour [CBDNML17]. Negative behaviour such as trolling can further be induced by seeing negative comments of other users [CBDNML17], creating an incentive for providers of a forum to stop such demeanour early or even prevent it altogether.

News forum providers have an interest in understanding who their users are, not only in terms of avoiding negative impacts but also in supporting and encouraging beneficial discussions. As stated above negative behaviour can be contagious. The solution does not solely include repressing negative behaviour, but understanding what makes good behaviour and emphasizing it. The providers want to tie the users to their forum and encourage fruitful discussions because it gives them more opportunities to interact with their users and take better actions related to their own business goals.

1.2 Research questions

The goal of this thesis is to create a robust classification model for users of an online news forum, and to research changes in the classification over time. The main objective is to suggest a generally applicable typology that captures common behavioural patterns of such news forum users by characterizing their behaviour and interactions without regarding textual information. The model aims to answer the questions:

- To what degree, based on standard performance measures, can users of an online news forum be classified based on known and latent characteristics, as well as interaction data?
- Furthermore, does the classification of users change over time?

1.3 Methodological approach

The methodological approach begins with literature research covering research about user typologies in the context of online social communities. Furthermore, the actual model creation begins with an exploratory approach aiming to create a qualitative understanding of user behaviour and, by defining the appropriate variables for a quantitative analysis. A statistical analysis will be conducted to detect cluster solutions within the data. Combining the results of both approaches will then create a model, which further has to be validated in terms of its goodness of fit by statistical measures. The complete methodological approach is described in Chapter 4.

1.4 Structure of the work

Chapter 2 provides information about news forum usage and the difficulties and chances forum providers are facing with large user bases. It covers the background of user categorization in the context of news forums and the reasons for its necessity. Finally, impacts of positive and negative user behaviour within such communities are introduced to further demonstrate the potential impact various behaviours can have on such a community.

Chapter 3 shows the current state-of-the-art in classifying users in the context of social communication platforms. While the work specializes in understanding users of online news forums, we can use approaches used by studies examining general forums and social community platforms as well. The focus of the literature research is on understanding the behaviour of users in such settings and classifying them based on their interactions.

Chapter 4 explains the methodology used to conduct this research in detail. It will provide information on the approach chosen to create the classification model, qualitative and statistical, as well as the steps involved in validating this model. It will also cover the data used for this analysis.

Chapter 5 describes the two parts of the data analysis: exploratory and statistical. The exploratory analysis focuses on the initial model creation process, especially the detection of patterns and evaluation of potential variables for the statistical analysis. The latter will create the classification model based on the findings of the exploratory analysis.

Chapter 6 presents the results obtained during our model-finding process. This part of the thesis explains the classification model and its roles. It details the differences between the classes and analyses class distributions among the user base over time.

Chapter 7 shows how the retrieved model was evaluated. The statistical evaluation will be based on the criteria defined in Chapter 4 and measure how well the resulting classification model explains the observed data.

Chapter 8 summarizes the results obtained by the thesis. It gives a summary of the most important observations, the contributions of this thesis, explains its limitations, and suggests possibilities for future work.

CHAPTER 2

Background

Consumption of news has changed with the rise of social networking platforms. Users are no longer limited to only reading what journalists wrote. Instead, they can now publish their own opinions, as well as comment publicly on daily news. In traditional printed news the consumer was limited to reading the newspaper, with a few exceptions that could have their comment printed in special sections of the paper. Social networking sites and open community forums on the other hand depend on the contributions of users in the form of user generated content (UGC) [GLZ15]. It is not only possible but encouraged for consumers to voice their opinions.

News consumption is becoming more social and interactive [DN11]. A survey from 2010 found that 25% of the surveyed have commented on a news story. Moreover, 37% of online news consumers said that commenting on news stories is an important feature with even higher percentages among younger users (51% of 18-29 year olds) [PRM+10]. While these numbers show promising possibilities from the viewpoint of publishers, in terms of customer connectivity and affinity to their user base, they also bring challenges with them. Newspapers often struggle with the quality and volume of postings. A solution in managing such amounts is often the (not exclusive) use of crowd-based moderation features [DN11]. These features often consist of reporting capabilities but also in providing direct feedback, in the form of approving or disapproving evaluations, or votes.

Understanding the behaviour of user interactions, as well as their motivations for reading and writing news comments, can assist forum operators in their task of creating an engaged user community while maintaining a favourable discussion environment.

2.1 News forums as important meeting places of online social interactions

Online news forums provide an opportunity for the news outlets that operate them [Bin12]. A forum with a thriving community can drive traffic towards the publication itself. It can provide the company with valuable user-generated content, as comments to articles but often also in the form of longer user commentaries, which can in turn fuel further discussions. It can give the publisher insights into public opinion, upcoming trends or satisfaction with its services [CC03], covering a variety of topics, from daily news to recurring sporting events and more activities of everyday life. In the end, an involved community, and the accompanying regular visitors, also mean revenue to the media house [Bin12].

Online social interactions are increasingly popular [PRM+10]. People want to discuss news and voice their opinions. Publications that utilize this circumstance can profit from their willingness, by providing the platform that enables readers to do so.

Along with these possibilities comes the responsibility to nurture and foster the community, including the labour-intensive task of moderating discussions. This includes removing disturbers on one side, without limiting or even censoring the discussion for a majority of users on the other side, but instead promoting active discussions among the users.

2.1.1 User-generated comments in online news forums

Not only does a general news forum provide a discussion platform for daily news, but sub-forums, often separated by categories or topic clusters, provide a more detailed insight into the discussions within specific subgroups of the user base. The degree and form in which users discuss and communicate might vary between sub-forums and so does the requirement to moderate the community.

Depending on the context of discussions, users take up different social roles [GD04]. Maybe the topical experts in the science sub-forum prefer to share their in-depth knowledge, while the politics resort is predominantly filled with strong opinions about varying policies. Other parts such as the sports category might be filled with rivalries, between fans of certain athletes and teams, with a recurring back-and-forth each time these athletes face each other. Overlaps in the form of communication are also likely, as in politics and sports, where hardened loyalties are often cultivated over years and decades, and the related discussions are accordingly intense. Even though these types can likely be found in each category, the degree to which one type dominates the discussion culture over the others might vary.

However, these distinct patterns in discussion must be considered by the moderators of a forum. What appears exuberant in one category could be the base level within another section. To provide the necessary conditions for a thriving community, the forum operators must be aware of how the community is composed. They have to know

and identify these behavioural patterns to augment and encourage discussions, without sacrificing the general discussion atmosphere.

One way to generate insights into these discussions is by detecting these behavioural communication patterns and subsequently categorizing users based on them. These categorizations can then support the moderators and operators, in their task of cultivating fruitful discussions but also help them notice escalating discussion environments.

2.2 User categorization in online communities

Whittaker et al. [WTHC03] already discovered in 1998 that online discussions on Usenet newsgroups had a massive *participation inequality* in effect. They analysed over two million messages by over 650,000 users. In their study 27% percent of the users only posted once, whereas a small fraction of 2.9% contributed 25% of all messages posted. This drastic split strongly suggests that users do not participate equally, but these different levels of participation do not allow conclusions about the behaviour in such communities.

Referencing this study, Golder et. al [GD04] put it as follows: "*unequal levels of participation underscore the fact that participants are not all the same.*" They further propose that the scope of participation has to be considered as well. Some users might be interested in single topics, while others may have an interest in many. In the same way, some users might want to contribute a lot to the community spending large portions of their day in there, while others use it only scarcely. Golder et. al [GD04] argue that combining these activity factors, along with other information, can tell us a lot about a user. For example, about the profoundness of their connection to the community or about their passion for specific topics.

2.2.1 Necessity for classification

Even though examples of the benefit that comes from understanding the user base have been given, the question of why the classification of users is necessary remains open to some extent. While understanding some users and their motivations is also feasible on a smaller scale, classification is the key to analysing and understanding the larger scale. As Lerner [Ler05] concludes "*Classification is the key to understand large and complex systems that are made up of many individual parts.*".

By classifying the relations and behaviours of the users, the complexity of the system can be reduced to a smaller set of types, which allows for a, as Gleave et. al put it, "*[...] comparative study of populations across time and setting*" [GWLS09]. Mapping users to roles or types simplifies the structures, which allows for better management on a higher level. For instance, user-type distributions can be monitored closely over time, to detect unwanted deviations. Such a deviation could be an increase in lurking users at the expense of the participating users' ratio, which could hint at technical or structural problems that hinder or demotivate users from participating further. At the same time,

these ratios could assist the forum operators as benchmarks when trying to convert users from a lower participating category to a more active one.

As noted above different people communicate in different ways. Consolidating these ways of communication into fitting user types could help the forum operators tailor their offered services towards more favourable types, in a community sense, while detecting increases in unwanted or malicious behaviours early. The goal has to be the generation of insights that can assist the operators in creating a flourishing discussion culture.

Marett et al [MJ09] for instance emphasize the importance of *lurkers* as potential future participators in a community. By seeing this group as a recruiting pool for future posters, the community administrators can take targeted actions to convert these users from purely reading behaviour to active participation. This group makes up the largest portion of a community, as stated above, therefore finding the right motivation that leads them into active roles, can be a key lever in the process of cultivating the community.

The following section provides further information on how the negativity and positivity of user postings can influence the behaviour of other users and, on a larger scale, influence the discussion culture as a whole within news forum communities.

2.3 Impact of discussion polarity on the course of online discussions

Humans are highly social beings and their interactions carry over into social networks and online communities. However, communication in online public forums does not necessarily follow the same rules as in offline environments. Cheng et al. [CDNML14] examined the effects positive and negative feedback have in the context of social forums. They wanted to show how evaluations by their peers impact the future contributions of news forum participants.

In their introductory remarks, they hint at expected behaviours that are based on classic behavioural psychology, such as the *operant conditioning* framework by Skinner, which aims to explain *"'tendencies' or 'predispositions' to behave in particular ways"*[Ski65]. Cheng et al. argue that if positive evaluations (e.g. positive ratings) should therefore act as *reward stimuli*, while negative evaluations (e.g. negative ratings) should act as *punishment stimuli*, the consequences should follow behavioural psychology. That is, rewarded authors should be encouraged to create high-quality content in the future and punished authors will decrease their contributions compared to rewarded authors.

2.3.1 Negative discussion evaluations

Yet, Cheng et al. [CDNML14] observe an almost opposite effect by studying the interactions of four large online news communities. In contrast to the deducted assumption that negatively evaluated users will reduce their contributions, the study showed an increase in postings, while also showing a decline in the quality of their postings. In addition,

the researchers found that these users would further evaluate other community members more negatively, resulting in a negative feedback loop. These observations are further supported by a separate study conducted by Cheng et al. [CBDNML17] that analyses how the behaviour of users can be negatively influenced by being exposed to negative "*troll*" postings.

"*Troll*" is a popular term given to users of social forums and networks that deliberately attract negative attention, by using controversial opinions, abusive language and more subtle tactics to disrupt conversations. According to Binns [Bin12] the term does not come from the mythical creature, but originates "*[...] from a kind of angling where a lure is dragged through the water to provoke a feeding frenzy amongst the fish.*" Further Binns concludes that the purpose of a troll is to "*be subtly or blatantly offensive*" to disturb and disrupt a conversation and to annoy fellow users.

The study by Cheng et al. [CBDNML17] examines the effect that being exposed to troll postings, as well as the current mood of the user, has on the likelihood of a user posting trolling comments themselves. It aims to disprove the general narrative that such behaviour comes from a small group of people, with specific motivation, personality or biological traits [Bak01][BTP14][HJSSB02], but rather show that the potential for such behaviour lies within everybody and that it can be reinforced by bad mood and exposure to such negative behaviour.

The researchers found that the negative mood increases the likelihood of a user trolling by 89%, while the exposure to such postings increases the odds by 68%. Both factors combined double the baseline rates of the participating user's likelihood to troll.

The effects of negative experiences are not limited to the causing event, though, but instead can persist for some time, and thus influence how future events are perceived, even though they are unrelated to the triggering event, that produced these negative emotions [KEE93]. Thus, negative mood can have a prolonged effect, making the negativity spill from one discussion into the next, unrelated conversation, where its influence spreads to a potentially new subset of users. A suggested approach by Chang et al. [CBDNML17] is to let users take a "time-out" to calm down after such exposure or even participation, which effectively reduces the probability of posting a comment that will later be flagged. Even a short break of five to ten minutes reduces such risk significantly.

2.3.2 Positive discussion evaluations

In contrast to the negative effects, the researchers [CDNML14] could not detect a cascading positive effect for positively evaluated users. They did not mirror their incoming evaluations onto their peers. Additionally, their being well received had no further impact on their posting frequency or future posting quality, disagreeing with their initial expectations based on Skinner's [Ski65] framework.

The study authors also looked at the users who did not receive feedback, neither positive nor negative, and found that these users showed a tendency to leave the community. Golder et al [GD04] consider this feedback of ignoring a user to be also negative.

Despite the missing expected result of passing on positive effects, as negative postings do, one might argue that the sole fact of a positively evaluated user continuing as a member of the community is the positive effect. While negative feedback reduces the posting quality and no feedback leads to dropping out, positive evaluations keep the community engaged. Golder et al [GD04] suspect that positive feedback encourages the users to keep participating, leading to a positive feedback cycle that manifests in a higher posting frequency.

They also argue that perceived positive feedback can differ between users. Some might consider intense arguments negative when they are joining a conversation, but could also perceive them as positive if they are actively looking for a discussion. As to them, the polarity of feedback does not depend on the tone of a response but on whether the interaction followed community guidelines.

This underlines the necessity for forum operators to keep a positive and, equally important, socially engaged community. Based on these results fulfilling only one of these two aspects will either lead to an increasingly hostile or potentially extinct community.

CHAPTER 3

State of the art

The previous chapter highlighted the necessity of user categorization and pointed out the consequences of various user behaviors, in terms of community development and its influence on other users' behavior. In this chapter, several user classification models for social communities are presented that have a high similarity to the news community we analyze. It will also be concerned with the methods used to develop these models and the various approaches selected by the analyzed literature.

3.1 Class definition and detection

Users get split into roles, types, or classes to group them into manageable clusters and to reduce the complexity of underlying social structures. The different terminology is often important for the nuances of a research task, but across research, they are often used interchangeably. In this thesis, classes describe a high-level placement of users into better manageable subsets with common behavioral traits. These patterns of behavior form the basis of our classes. They are the result of people who consistently adopt distinct behaviors in social settings [GWLS09]

3.2 Methods for analyzing user behavior

Analyzing user behavior is a well-researched field, resulting in a variety of approaches to conducting such analyses. Depending on the research target, methods included but are not limited to textual analysis [PT11, DSR14], network visualizations [WGFS07, VS04, TSFW05, WCK+11], regression analysis [SV04, NP00], social graph analysis [AA15, HCH16], clustering [BH11, EMG07, Hor07, BHK11] and social studies [BH11, Nie06, LB07, GD04].

3.2.1 Large Language Models (LLMs)

In recent years the availability of large-scale language models, pre-trained with unlabelled data, opened up new ways for textual classification[ZZH21]. Instead of using supervised learning approaches to train one's model, it is possible to make use of existing LLMs, which incorporate a huge quantity of existing texts, and apply these general-purpose models to one's research questions by extracting their pre-trained neural network (NN) layers and applying a new layer on top built for the selected task [GMS+20]. It is important to note that training LLMs is in general computationally expensive.

Zhao et al. use such a setup to analyze toxic comments in user-generated content (UGC) to compare results between the LLMs BERT, RoBERTa, and XLM [ZZH21]. To do so, they use already labeled datasets to conduct their analysis. For tasks requiring a textual analysis where pre-labeled examples already exist, fine-tuning an LLM seems a reasonable method to analyze the given data.

A recent study by Viswanathan et al. [VGL+23] attempts to apply LLMs to the clustering stage, to improve the quality of clustering results according to the requirements of the domain expert's needs. They propose three possible stages for LLM-enhanced clustering: before clustering, during clustering, and after clustering. These respectively have the task of improving the textual representation, adding cluster constraints, and correcting cluster assignments with low confidence.

Another study by Zhang et al. [ZWS23] proposes a framework called ClusterLLM to improve clustering results using LLMs. It utilizes triplet questions with a structure of 'does A better correspond to B than C', where A, B, and C are points of data belonging to different clusters, to predict which of the two options is closer to the data point. The results of these triplet questions are then used to fine-tune small embedders, to improve the quality of the clustering results.

LLM-based approaches are best used with content-related problems, as their advantage lies in textual assessment and being able to better extract the meaning of text passages compared to previous keyword-based content analysis. They can determine the proximity of two pieces of text, and allow for better analysis of content similarity.

3.2.2 Selecting the methodological approach based on available data

As discussed, typologies and classification models are developed by utilizing various methodologies. Selecting the right approach is related to the data and dimensions available for the research. In the context of user typologies for social forums and networks, this data generally comes in the form of user questionnaires, scraped data such as forum entries, data retrieved via public APIs, or data provided by the organization behind the online community. The data can include textual dimensions, such as forum postings or tweets, user relationships, and networks, for example, data about which user follows another, interaction data between users, such as replies, direct messages, or votes

but also non-public organization-dependent information, such as internal labels or flags for hostile users and spammers.

It is important to distinguish between the limitations of available data and the types of analysis that are possible to conduct with it. As such, the results of typologies also have to be regarded in the context of the data available for the analysis of its limitations. Approaches with special interest for this thesis are described alongside their roles in Section 3.5.

3.3 Feature and measure selection

Users can be classified with various measures. Gleave et al. [GWLS09] distinguish two methodological approaches when performing social role analysis: interpretative and structural. Interpretive methods help researchers identify roles and understanding users. They include the context in which roles develop and deal with the individual's motives. However, according to Gleave et al., their weakness lies in omitting a view of the macro scale of these social structures. As a result, the consequent roles are too context-specific and not applicable across social settings.

Structural methods on the other hand use social network analysis, building upon network data and looking for *structural signatures*[FS02] to identify roles. Yet, Gleave et al. argue that even though structural methods are capable of revealing roles, they lack sufficient capabilities to explain them. Therefore they propose a mixed approach to combine the detection and explanation of user roles.

Gleave et al. [GWLS09] limit the measures to two general categories: interpretive and structural. While the interpretive approaches rely on social studies, surveys, and analysis of the content, structural approaches depend on metrics and measures that allow for quantitative analysis. In contrast, Chan et al. [CHD10] distinguish five types of features in their analysis: structural, reciprocity, persistence, popularity, and initialization features.

- **Structural Features** measure the interactions of users with their neighbors in an unweighted, directed graph.
- **Reciprocity Features** measure the degree to which users are having conversations with others, by analyzing the degree of incoming and outgoing edges.
- **Persistence Features** group the mean and standard deviation of posts per thread.
- **Popularity Features** measures the popularity of a user, as the degree of incoming replies to a user compared to all replies and the degree of posts of a user that receives any replies.
- **Initialisation Features** are defined as the percentage of threads initialized by a user. It is used to differentiate between initializing users and those that just reply.

Their research focuses on the interactions between users, with a special focus on the directed interactions depicted by graphs. Additionally, they also take the way users communicate (e.g. initialization, replies) or the amount of their contributions into account.

This thesis will focus on what Chan et al. define as *Persistence Features*, *Initialisation Features* and *Popularity Features*, although they will not necessarily be defined and implemented in the same manner.

3.4 Participation inequality of users in online communities

One of the most important basic conditions that have to be taken into account is that online social communities usually have massive participation inequality [GD04, GLZ15, MJ09, WTHC03, Nie06, BRCA09]. Because of this, the types of roles within typologies have to be seen in the right context. Various active roles are often only a subset of the participating minority, as opposed to the silent majority of a community. Therefore this circumstance has to be considered during the analysis in chapter 5. The central statement of all these inequality observations is that there prevails a drastic split between actively contributing users and those that are only consuming within online social communities. Although the percentages and definitions of these roles differ between each study, they all share the view that most of the users are not actively participating and a majority of contributions in return come from a minority of users.

Whittaker et al. [WTHC03] already discovered in 1998 a strong participation inequality within their analysis of 500 Usenet newsgroups with over two million postings. In their study, they noticed that a share of only 2.9% of users contributed a disproportionately large amount of postings (25%). Golder et al. [GD04] noted that their most vocal role, the *Celebrity*, showed even higher percentages. Their study discovered that the 14% with the highest amount of postings were responsible for 75% of messages (albeit the sample size was much smaller with 137 individuals).

In 2006 Nielsen [Nie06] proposed the *90-9-1 Rule for Participation Inequality*. Its core statement claims that in most online communities 90% of the users are lurkers, or users who do not contribute, 9% of users contribute a little and 1% of users are responsible for the majority of contributions. While these percentages form the general rule, the study mentions that these numbers are not set in stone. Instead, the magnitude of the numbers is the important takeaway. They give two examples with similar distributions. First blogs, for which they suggest a 95-5-0.1 rule as an estimate, and Wikipedia which follows an even more drastic split of 99.8-0.2-0.003, whereas the highest contributing group only consists of 1000 members.

In 2020 Gasparini et al. [GCBC20] put the distribution proposed by Nielsen to the test on open source contribution data of GitHub. They wanted to analyze if what they call the '*volunteer's dilemma*', which states that participants can benefit from the work of a group without participating themselves, and was coined by Andreas Diekmann [Die85],

also applies in open source development. They could partially confirm the results for large projects, while they could also disprove it in parts when they regarded that users were generally active in their projects. Their results highlight the importance of looking at the context in which user activity is measured, and how different contexts can lead to different activity levels.

Another study, conducted in 2019 by Antelmi et al. [AMS19], attempted to analyze participation inequality in Online Social Networks (OSNs) by using Twitter data and focusing on the same proposed distribution by Nielsen. They proposed a typology consisting of four groups: high-activity (5.18%), medium-activity (11.11%), low-activity (33.74%), and no-activity (49.94%). While their results confirm a strong inequality in the distributions, their results indicate that a 3 out of 4 distribution regarding inactive users appears more accurate in the social networking context than the 9 out of 10 distribution by Nielsen.

Benevenuto et al. [BRCA09] demonstrate with clickstream data that 92% of all user activities are browsing, which directly relates to the lurking majority. Yet, this data also includes browsing activities of otherwise participating users so it does not apply exactly to the above rule. Still, the study finds that users who are only browsing, are 13 times more than those who also post messages, highlighting that the difference is still enormous. The authors observe that the more user engagement an activity requires and the more time-consuming it is, the less users are performing it.

3.5 Important classification typologies for this thesis

The following sections describe typologies of importance to the context of this work. These classifications focus especially on social computing, social networks, and interactions between users. They use different methodologies and approaches to define their roles, but their results provide a promising basis for our research and offer the possibility of comparing the results of this work with similar typologies.

These roles in particular shall provide a foundation to compare results gathered during the exploration phase and serve as guidance when defining and naming roles discovered within this work's analysis.

3.5.1 Typology of social networking sites users by Brandtzaeg and Heim

Context of typology

This study [BH11] was conducted with users of four Norwegian social networking sites. The data consisted of answers to an online questionnaire filled out by 5,233 users with a median age of 16 years, as well as socio-demographic measures (residence, gender, age, and education). The main question the researchers wanted to answer was how users can

be classified based on their participation in an online social community and their mode of communication.

Brandtzaeg et al. used a k-means cluster analysis for the identification of the classes. They validated their solution by repeating the analysis with smaller subsets, aiming for the repeated creation of the same clusters, as well as a qualitative validation by extracting representative users for each cluster.

Roles

The typology distinguishes five roles that represent different levels of participation and motives of usage. These motives are informational or recreational. Whereas the level of participation can be high or low. Both of these measurements are on a scale, resulting in the following roles:

Sporadics (19%) visit the community scarcely and have a low rate of participation. They read other's postings infrequently and do not wish to contribute content themselves. They check if other people reached out to them.

Lurkers (27%) are the largest category and very passive. They do not make much contact with other users. Yet, they show interest in many activities but only from a non-involved passive viewpoint.

Socializers (25%) show high levels of interaction with other users. They are eager to reach out to other users and have generally high participation levels. Their interactions are limited to small talk, though.

Debaters (11%) are similar to Socializers in terms of participation but show more appreciation for longer discussions. They are very involved and contribute actively to the community.

Actives (18%) are engaged in all kinds of activities. They are very social and contribute in all kinds of activity to the community, engaging with other users and sharing user-generated content.

3.5.2 Participation typology of Social Media and Online Community users by Nielsen

Context of typology

An analysis by Nielsen [Nie06] investigated the usage and user participation of users in online social communities. Although it is not an academic publication, it had a great impact on related research [BH11][Bra10]. The study looked at large social networks and communities to classify users based on their involvement and activities.

Roles

Lurkers (90%) are defined as users that only read or observe but do not contribute themselves. In the study, examples of large social communities like blogs, Wikipedia, and Facebook are given with even higher percentages (95-99%).

Intermittent Contributors (9%) are actively participating users but only from "*time to time*". As Nielsen puts it, "*other priorities dominate their time*" [Nie06].

Heavy Contributors (1%) are those users, who contribute very actively. They are responsible for a majority of contributions within the social community with up to 90% of postings coming from them.

3.5.3 Adopter categorization based on innovativeness by Rogers

Context of typology

An early and perhaps one of the most important theories on technology use is Rogers' diffusion of innovations model [Rog62]. It explains the process by which innovations are adopted and categorizes the people depending on when they begin to adopt such innovations. Although conducted in a time before computers were the electronic all-purpose machines they are today, the categorization depends on the adoption rates of technology, which is related to the problem we face.

While this theory explains the general innovation and adoption rate, we are more interested in usage and media behavior. Still, the model might be useful to determine patterns along the adoption and usage of online news forum features, describing the behavior of users. It could be argued that users also follow specific progress from lurking to posting. Whereas adoption is describing the process of participating in activities that demand more user engagement than previous contributions.

Roles

The model categorizes users, based on the time at which an individual adopts an innovation, into the following categories:

Innovators (2.5%) are the first group to adopt an innovation and make up the smallest percentage. They are described as venturesome.

Early Adopters (13.5%) is the second-smallest group but still very early in adopting innovation. They are further described with the attribute "*respect*".

Early Majority (34%) form one of the two largest groups. They are still considered as early adopters and are described as deliberate.

Late Majority (34%) is the second large group, following the *Early Majority*. Their dominant attribute as stated in the model is being sceptical.

Laggards (16%) are the last people to adopt an innovation. People within this group are described as traditional.

3.5.4 Taxonomy of social roles by Golder et al.

Context of typology

This research by Golder et al. [GD04] aims to characterize the social roles of users in newsgroup communities (Usenet). They combine data obtained from observing sixteen different newsgroups with traits defined by research in the fields of sociolinguistics, ethnography of communication, and social psychology. The taxonomy attempts to explain typical observable behaviors found in newsgroups with the following six roles.

Roles

Celebrities are highly active users who are known within the community. They are central figures who spend lots of time contributing. Celebrities represent participation inequality in that they are a small percentage of users, yet are responsible for a disproportionally large amount of contributions.

Lurkers are users who only read but do not participate. *Lurking* is a suggested strategy for *Newbies* to familiarize themselves with a community. Yet, lurking is not a phase but can be the whole strategy for some users. There can be multiple reasons for not contributing. Very infrequently they do participate making differentiation difficult to other roles.

Newbies are characterized by low communicative competence and often lacking targeted subject-specific knowledge. They are often instructed to study the community before contributing on their own, to learn the expectations of their respective group.

Flamers show aggressive behavior, adopting intimidation as the primary means of communication. They do not wish to become members of a community but show hostile language from the beginning.

Trolls try to trick a community into thinking they are somebody they are not. Trolls attempt to pass as part of the community to create disturbance. They try to waste the time of members, stir up controversy, and provoke arguments.

Ranters aim to stir up unnecessary debate, but have, unlike the *Troll*, a clear goal or mission in mind. They are characterized by a high amount and extent of postings. Their agenda is very important to them.

3.5.5 Common user roles in discussion forums and boards by Chan et al.

Context of typology

Chan et al. [CHD10] conducted a study analyzing users' communication interactions in one of the largest general topic discussion boards in Ireland (boards.ie). The interactions were modeled as weighted, directed graphs portraying the relationships of users, threads, and postings. The model included information about the reply amounts two users had

for each respective thread they posted in so that the amount of interaction between sets of users is depicted.

Roles

Joining Conversationalists show high levels of communication, but only with a limited amount of users. They do not initiate conversations but rather join existing ones.

Popular Initiators show a high level of popularity, defined as a high amount of incoming responses by other users. They often initiate threads, opening the discussion for other users.

Taciturns have limited conversations with a few users. In general, they show a low reciprocity and volume of communication.

Supporters seem to be the most average participants of the forums. They participate but not in exceptionally high or low volumes and are described as the *backbone* of the community.

Elitists are users that prefer conversations but only with a limited amount of users. This is characterized by a low amount of neighbors with bi-directional communication but a high amount of bi-directional threads.

Popular Participants are involved with a large amount of users on the forum. They like to participate in the conversation but do not open many threads themselves. They are suggested to be a combination of *Joining Conversationalists* and *Popular Initiators*.

Grunts are defined as users with low volumes of communication to only a few users. They differ from *Taciturns* by a higher amount of reciprocity.

Ignored participate by posting, but their postings tend to not get replied to.

3.5.6 Types of members in virtual communities of consumption by Kozinets

Context of typology

Kozinets' [Koz99] types are the result of a theory based on theoretical assumptions about online social communities that center around consumer interests. People sharing these interests exchange in online communities, suggesting the development of different roles based on the ties towards the interest itself and towards other members of the community.

The two, largely independent factors, contributing to the model are consumption activity and intensity of social relationships. The first factor, consumption activity, measures the importance of the relationship towards the activity and thus its value to pursue membership in a community circling this activity. The second factor describes the intensity of relationships between members within the community.

Roles

Tourists do not feel strongly connected to the community. They lack social ties and only show superficial interest in it, joining only every so often.

Minglers have strong social ties to the community. But they only show a low interest in the consumption activity.

Devotees show strong enthusiasm for consumption but only have little social ties to the community.

Insiders have strong social ties in the community and a strong interest in consumption activity.

3.5.7 Social Technographics typology by Li et al.

Context of typology

The typology suggested by Li et al. [LB07] is based on data from two Forrester surveys about technology adaption. The goal is to explain social computing adaption, whereas the levels of participation of defined social computing behaviors serve as the criteria. The study relies on current social computing technologies, such as blogs and social networks, but advocates its applicability to future social computing technologies. The typology attempts to explain how consumers approach technologies and not just which one they adopt.

Roles

The following six roles are ordered by their level of participation. Users can be represented by multiple roles. The study authors depict the topology as a ladder, starting with the highest participating role (*Creator*) and leading down to the lowest active one (*Inactive*).

Creators (13%) are the most participating users. They consist of online consumers publishing blogs and maintaining web pages or video uploaders. Although all of these are options for *Creators*, they usually don't do all of them but only select activities. The study concludes that *Creators* are generally young and have an equal gender split.

Critics (19 %) participate in much lower volume than *Creators*. Their activities are commenting (e.g. on blogs) and reviewing (e.g. web shops). *Critics* tend to use other creations (as aforementioned blogs or webshop products) to post their contributions. According to the study *Critics* are several years older than *Creators* and that 40% of them are also *Creators*.

Collectors (15%) are aggregating and collecting resources such as bookmarks or RSS feeds that are inherently shared. They organize the content *Creators* and *Critics* create. The study discovered that *Collectors* were the most male-dominated of the roles.

Joiners (19%) are using social networking sites as their most defining characteristic. They are the youngest of the six roles and are very likely to also engage in other social computing activities.

Spectators (33%) are the primary consumers of the content-creating roles. They read, listen to, and view the contributions of other users. The study finds that this group has the lowest household income and generally a larger percentage of women than men. Some members of this group are also *Creators*, but 31% of *Spectators* do not additionally fall into one of the groups with higher participation levels.

Inactives (52%) do not participate at all in social computing activities. They are affected when the contributions of others are featured outside of social computing, e.g. in the news. On average these users are 50 years old, with a higher percentage of women.

3.5.8 Social roles in online communities by Gleave et al.

Context of typology

This paper by Gleave et al. [GWLS09] investigates the standardization of the term *social role* as a combination of social psychological, social structural, and behavioral attributes. It does not provide a full user typology per se but rather suggests an approach for creating typologies in the context of social communities.

It combines interpretive and structural approaches, as already discussed in the beginning of this chapter, aiming to apply a holistic mindset in role creation, instead of solely relying on one branch of methods. The practical approach chosen combines exploratory qualitative analysis (role discovery) with quantitative analysis (role verification).

The examples feature two distinct datasets of Usenet and Wikipedia, in which the superiority of a combined approach shall be highlighted. Detecting roles is only the first step for the authors, the main goal is to investigate role ecologies. It is described as the interplay between roles to understand the social world. The possibility this strategy poses is to simplify the otherwise complex world of social communities and provide a better means of comparison between varying social settings.

Roles (Usenet)

Answer People provide the majority of content in newsgroups by responding to the questions asked by other users. In terms of role ecologies *Answer People* are dependent on *Question People* because without them they could not exist. They require a larger amount of users to ask relevant questions to elevate themselves into this role.

Question People ask the questions that *Answer People* respond to. They provide the base on which *Answer People* exist, but at the same time are often only attracted to locations where the latter are present.

Discussion People are characterized by reciprocal exchanges with other users. They tend to have these connections with a large amount of other *Discussion People*. This

role provides the majority of discussion content in a newsgroup, resulting in longer conversation threads.

Discussion Catalysts are triggers of long threaded discussions. They provide the messages that spark longer conversations led by *Discussion People*. Their key characteristic is the creation of new threads, often by use of external resources. Despite creating these longer conversations *Discussion Catalysts* often do not further participate in the discussion.

Roles (Wikipedia)

Substantive Experts possess extensive knowledge about specific topics, in which they contribute large portions of content and help resolve disputes on the same. *Substantive Experts* have a higher likelihood of engaging with other *Substantive Experts* in longer conversations, to improve pages around their area of expertise.

Technical Editors are enforcers of formatting standards on Wikipedia. They also concern themselves with grammar, style, spelling, link accuracy, and other smaller issues. Their task is very repetitive and independent of the actual content on a page.

3.5.9 A typology of Internet users in Europe by Brandtzaeg et al.

Context of typology

This study, by Brandtzaeg et al. [BHK11], is concerned with creating a typology of Internet users in Europe. With data from five European countries (Norway, Sweden, Austria, Spain, and the UK) it attempts to classify users based on a questionnaire about their Internet usage and structured data in the form of age, gender, Internet access as well and household size. In addition to the typology, it attempts to explain differences in user types across the countries.

Roles

Non-Users (42%) do not regularly use the Internet.

Sporadic Users (18%) use Internet services infrequently and only for specific tasks. Email is an example given by the study.

Entertainment Users (10%) show increased usage of entertainment-related services, such as radio, TV, games, or music.

Instrumental Users(18%) have high scores for goal-oriented activities. The study mentions tasks such as net banking, e-commerce, and travel. Half of the members of this group use the Internet daily.

Advanced Users (12%) show the highest scores for Internet activities throughout the study with a predisposition for utility or instrumental tasks compared to leisure activities.

3.5.10 Commonalities within regarded typologies

This section aims to provide an incomplete comparison of the above typologies. It looks at the observed commonalities of various classes within these typologies, to establish a set of recurring behaviors. These reappearing behaviors will aid in the qualitative analysis, as they can point towards common traits within a social community context.

The following behaviors appear regularly throughout the analyzed typologies, suggesting a broader applicability.

Lurker - very low to no activity

The first behavior describes very inactive users. They often show no activity at all, except regularly reading, listening, or viewing the content of a community. This group usually makes up the largest percentage of the user base [GLZ15]. This role is called *Lurker* most of the time. They are often described as users that do not actively contribute at all, although some typologies distinguish further between the more inactive user types, e.g. Li et al. [LB07] or Brandtzaeg et al. [BH11], in which low contributions are differentiated from not participating at all.

The following roles most relate to this behaviour: *Lurkers & Sporadics* (Brandtzaeg et al.[BH11]), *Lurker* (Nielsen [Nie06]), *Lurker* (Golder et al. [GD04]), *Grunts* (Chan et al. [CHD10]), *Tourists* (Kozinets et al. [Koz99]), *Spectators & Inactives* (Li et al. [LB07]), *Non-Users* (Brandtzaeg et al. [BHK11]).

Debaters - active participation especially in conversations

This user behavior describes users' inclination to actively participate in discussions. These users are often very active in the surrounding community, their primary contribution is leading discussions with other users, which often show the same tendencies.

The following roles most relate to this behavior: *Debaters* (Brandtzaeg et al.[BH11]), *Intermittent Contributors* (Nielsen [Nie06]), *Elitists, Popular Participants & Joining Conversationalists* (Chan et al. [CHD10]), *Minglers* (Kozinets et al. [Koz99]), *Critics* (Li et al. [LB07]), *Discussion People & Substantive Experts* (Gleave et al. [GWLS09]).

Actives - high activity and popularity

This type depicts usually the most active group within a community. Users who show this behavior often make major contributions to the community, by posting, creating videos, voting on others' postings, opening and answering threads, and leading discussions with other users. These users are often known, regardless of whether this is because of their contribution frequency or content, and linked to being popular within the community.

The following roles most relate to this behaviour: *Actives* (Brandtzaeg et al.[BH11]), *Heavey Contributors* (Nielsen [Nie06]), *Celebrities* (Golder et al. [GD04]), *Popular Initiators & Popular Participants* (Chan et al. [CHD10]), *Insiders* (Kozinets et al.

[Koz99]), *Creators* (Li et al. [LB07]), *Answer People* (Gleave et al. [GWLS09]), *Advanced Users* (Brandtzaeg et al. [BHK11]).

Socializer - very interactive but only small talk

The last recurring behavior describes active participants without remarkable contribution amounts. These users are often seen as the core of the active user base. They contribute by posting, discussing, and evaluating others. Their contributions range from everyday interactions to being regulars, but they don't reach the same level of activity as the *Actives*.

The following roles most relate to this behaviour: *Socializers* (Brandtzaeg et al.[BH11]), *Supporters* (Chan et al. [CHD10]), *Devotees* (Kozinets et al. [Koz99]), *Joiners* (Li et al. [LB07]), *Question People* (Gleave et al. [GWLS09]).

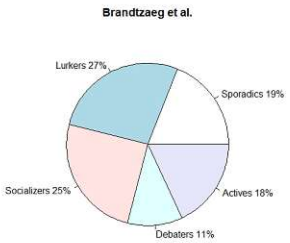
3.5.11 Typologies of special interest

This section concludes which of the previously discussed typologies provide special interest to the work of this thesis. The first important classification is by Brandtzaeg et al. [BH11], because it looks at its user base from a similar point of view, aiming to distinguish them based on their participation. The research conducted by Nielsen [Nie06] aids as an orientation guide when the hugely disproportionate amounts of contributions have to be considered. It serves as an example of the skewness of distribution among user contributions. The typology by Golder et al. [GD04] targets a similar discussion context, even though the discussion capabilities were more limited than they are in this research. Yet, their differentiation between various negative roles and their motivations could aid in a further distinction of negative patterns. The most similar setting for analysis was provided in the research of Chan et al. [CHD10]. Despite attributing greater importance to the connections between users in graphs and the lack of evaluations by users, their research comes from a similar context of discussion forums.

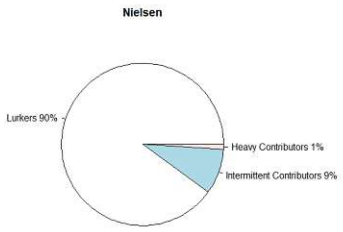
3.6 Role distributions

An important measure in evaluating the user base of a social forum is looking at the composition of roles. Various factors can influence the composition within the same forum, such as discussion context, time, or topics. For example, certain topics might be more discussion-prone than others. Chan et al. [CHD10] applied their classification model to 20 subforums of varying topics and showed that role distribution can vary greatly between them. Another dimension that might have an influence is time. The user might change their behavior with time, eg. by getting more accustomed to a forum and opening up over time or the inverse by losing interest with time.

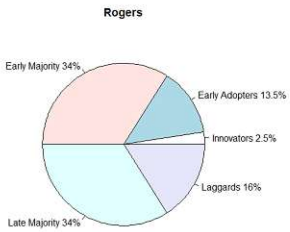
Figure 3.1 shows the distribution for selected studies of importance which provided overall distribution percentages for their classifications.



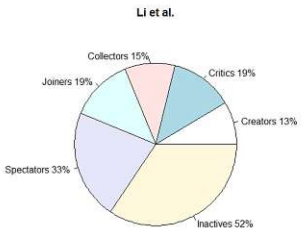
(a) Brandtzaeg et al. [BH11].



(b) Nielsen [Nie06].



(c) Rogers [Rog62].



(d) Li et al. [LB07].

Figure 3.1: Role distribution for selected typologies.

3.6.1 Context dependent differences

Chan et al. [CHD10] demonstrated that the overall distributions of roles, do not have to be similar throughout smaller sections of the community. While they do not provide a general distribution, they depict various sub-forums and highlight their most dominant roles. The context and theme of a sub-forum appear to have a large influence on the way users communicate and the collocation of roles can be subject to immense variations.

The importance of context for role assignment and discovery is also highlighted by Gleave et al. [GWLS09]. One of their examples mentions that users might inherit implicit roles such as *experts*, which gives them a say when sanctioning others, but that status is very loose and context-dependent. It is not clearly defined what makes them an *expert* in a specific part of the community, often users are not even aware of this implicit role. Some roles also only exist in the presence of others, as the authors note about the *Question People* and *Answer People*. In such a case one of the roles is dependent on the other.

3.6.2 Changes over time

A final aspect that needs to be investigated more closely is the factor of time in the classification of users. As Cheng et al. [CDNML14] suggested in their research about the impacts of user behavior, a potential negative feedback loop can lead to a decrease in

discussion positivity. If users start to discuss more hostile over time, this change should be observable over time. As a result, the attribution of users to certain role should change with time. These potential influences on roles in the context of time are therefore an important subject for discussion.

Changes could potentially span over a short period of time, resetting to their original state after a while. For example: a user could fall into a more negative role after a heated discussion and then, after a cool-down period, be reassigned to the original role. Another option, though, could also be a long-term change in discussion behavior over the lifetime of a user [Gar18]. An indication for such a shift is given by the circumstance that many users start within new communities as *Lurkers*, or other shyer user roles, before eventually starting to actively participate. These users change their way of contributing, effectively having a long-term impact on their role.

CHAPTER 4

Methodology

4.1 Data

4.1.1 Data collection & description

The data used to conduct this research was provided by the Austrian newspaper *derStandard*[1]. They operate an online news forum since 1999[2], in which users can leave comments below articles. These comments can be written as a response to the article or as a response to another user's comment, which creates the possibility for deeply nested conversations.

Due to the naming chosen by derStandard and for further coherence, we will continue to refer to these comments as a posting. Complementary to postings users also can vote on the postings of other users, giving them the power to express approval or disapproval for each posting. These forms of interaction facilitate different types of communication while allowing varying levels of conversations. Figure 4.1 shows a nested posting of the derStandard.at website, including a top-level posting with two nested replies and user votes.

To conduct the analysis, access to the majority of existing data was given, including all types of tables and data regarding the online news forum.

As this work focuses on quantitative user information only, in contrast to qualitative information, the contents of the user postings were not taken into account. Instead, the analysis focuses on calculated and aggregated data that describes a user's engagement within a set period of time. These variables include the amounts of postings written by a user, their voting behaviour (upvotes and downvotes), their perception by other users

[1] https://www.derstandard.at

[2] https://about.derstandard.at/unternehmen/die-chronologie-des-standard-und-derstandard-at

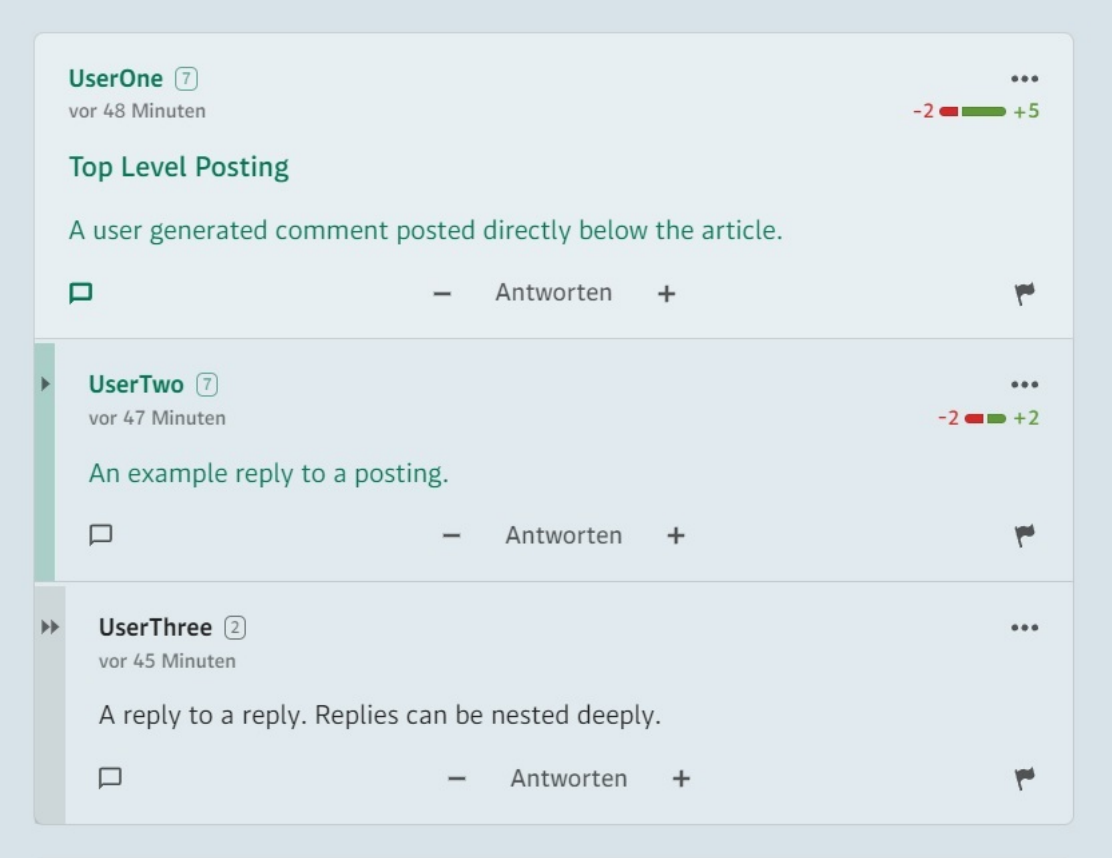


Figure 4.1: Posting on derStandard.at with nested replies and votes.

(measured as incoming votes) but also fixed values such as the gender of the user or their account age.

The full list of contemplated variables can be seen in Table 4.1. These data categories will be used in the initial exploratory qualitative approach to discover trends for the later statistical analysis.

Furthermore, the goal will be to further reduce the amount of variables to the smallest subset of meaningful variables which can successfully categorize users and explain their behaviour in this social forum context.

4.1.2 Data breakdown

As mentioned above the data was provided to us for the full duration of the forum's existence. The amount of data can be seen in Table 4.2. Due to the data size, the recommended approach was to use a smaller subset and to begin the analysis with an exploratory approach. The size of the dataset was gradually reduced to a month, a

name	description
communityIdentity	A unique identifier of a forum user.
gender	The self-reported gender of the user (not required).
postings	The number of written postings for the given week.
articlesCommentedOn	The number of unique articles the user commented on.
channels	The number of different categories a user commented on.
topLevel	The number of postings directed at the article.
otherLevel	The number of postings written as a reply.
firstLevelReplies	The number of replies directed at top-level postings.
averageTimeDifference	The average time passing between postings.
votesGivenTotal	The total amount of votes given by the user.
votesGivenPositive	The number of times the user upvoted another posting.
votesGivenNegative	The number of times the user downvoted other postings.
votesGivenRatio	The ratio of positive to negative votes given.
votesReceivedTotal	The total amount of votes received by the user.
votesReceivedPositive	The number of times the user received an upvote.
votesReceivedNegative	The number of times the user received a downvote.
votesReceivedRatio	The ratio of positive to negative votes received
votesGivenReceivedRatio	The ratio of incoming to outgoing votes.
positiveVotesGivenReceivedRatio	The ratio of positive incoming and outgoing votes.
negativeVotesGivenReceivedRatio	The ratio of negative incoming and outgoing votes.
weekNumber	The number of the calendar week.

Table 4.1: Variables used to conduct the exploratory analysis.

week and eventually a day, before realizing that analysing too granular does not give the desired clarity.

Based on the assessment of a domain expert of derStandard, we concluded that the smallest period for our initial exploratory analysis should be set to a full calendar week, to account for fluctuations in behaviour between different days of the week[ZLW+21], e.g. workday vs. weekends. Additionally, single large events occurring on specific days might skew the results by not representing an average day.

Following the further assessment of the expert, a conclusion was that the minimal period for the statistical analysis should include at least 4 whole weeks. Relying on weeks instead of using a full calendar month mitigates the aforementioned shortcomings of varying weekday behaviour and removes the disparity of the lengths of different months.

During the exploratory analysis (see Section 4.2) single day data was used as well, as the manual discovery of patterns was not affected by these shortcomings, due to the usage of a fixed discrete scale (see Subsection 5.1.3). During the statistical analysis periods between 1 day, used primarily as a control group for the exploratory analysis, and 6 months were chosen, with a focus on the aforementioned 4-week groupings (see

type	amount
users	858.926
articles	1.789.210
postings	119.884.043
votes	484.347.541

Table 4.2: Amount of data provided (Nov. 2002 - Nov. 2021).

Subsection 5.2.3).

4.1.3 Scaling the data

As discussed earlier, participation in online forums is not equally distributed (see Section 3.4), resulting in heavily right-skewed data when it comes to quantifying user interactions, as can be seen in Figure 4.2. Not only does this observed inequality affect the general usage, splitting the user base into interacting and non-interacting users. It can also be observed within the user classes or categories themselves. The reason behind this is that classes most notably describe the relationship between the selected interaction variables, but not their overall intensity. Two users with similar posting behaviour might have very different amounts of time they spend in the news forum, and as such contribute in a similar style but with a vast difference in their actual observable contributions.

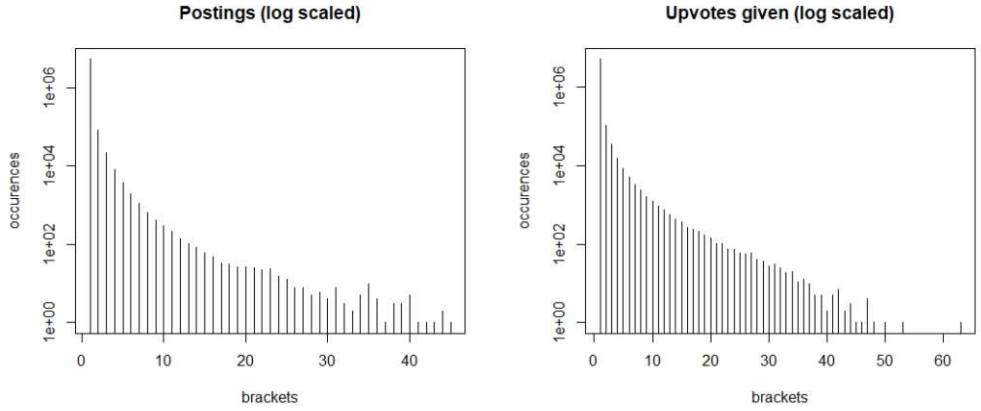
As an example: a user who only ever posts will be assigned a different category as someone who only votes on others' postings, but that does not convey any information about the amounts of postings or votes these users contribute. There will be a higher similarity in behaviour between two users who only post, even if their posting frequency is much further apart than to a user who only contributes by voting.

Knowing about these power-law distributions when it comes to social networking contexts makes it necessary to account for them during the analysis. In this work, all the interaction data will therefore be *log scaled* for the statistical analysis, to focus on the relationship between the variables, instead of the pure amount of interactions.

4.2 Exploratory approach

To detect classes of behaviour among the users a combined approach of exploratory qualitative analysis and quantitative analysis was chosen. It follows a similar approach to Gleave et al. [GWLS09], which aims to find initial roles by using a qualitative understanding of the context and exploratory data analysis and then proceeds to verify these results with a quantitative analysis of the behaviours.

The initial starting point is an exploratory approach in which salient patterns of behaviour are to be found. By looking into the data and moving between content and structure we improve our understanding of common communication patterns. The qualitative analysis



(a) Log scaled postings.

(b) Log scaled given upvotes.

Figure 4.2: Visual representation of power law distributions within the analysed dataset.

has the goal of surfacing these patterns and potential classes as a starting point for our further quantitative analysis.

The subsequent quantitative analysis aims to verify the assumptions made during the exploratory phase. It should reliably detect these behavioural patterns, not only to confirm (at least parts of) our empirical classes but also to create a solid foundation for our classification model.

4.2.1 Empirical model creation

The initial qualitative analysis starts manually by extracting recurring similar behaviours. These are patterns in the data which attract attention. Such a pattern can be high amounts within one of the variables, whereas the others remain low or a repeating co-occurrence of a pattern consisting of more than one variable. An example of this behaviour could consist of three variables, whereas two of them have high amounts and the third one low amounts. These behaviours and emerging patterns are all noted in a matrix consisting of the selected variables and the users that had contributions within the chosen timeframe. At the beginning of the analysis, a period of one full day appears manageable while not limiting.

These qualitatively detected behaviours are then extracted and quantified. Because the patterns can have arbitrary sizes in terms of variables, further co-occurrences of detected patterns will be analysed. If for example two or more of these behaviours only occur alongside each other, we have a strong indication that these behaviours could belong together.

The result shall comprise a list of assumed recurring behaviours. These behaviours will further be analyzed alongside classes of researched literature, to form a further opinion about their validity. Additionally, this early model will have to be complemented with

literature-based classes, that are *undetectable* for the scope of this research, such as potential *Lurkers*. These roles have to be validated separately in the following analysis.

4.2.2 Defining the variables for the analysis

Based on the insights about seemingly important patterns in the interaction of users, which was gained by the exploratory analysis, the set of important variables needs to be evaluated for the subsequent statistical analysis.

A large number of variables to describe the characteristics of communication within the forum proved to be helpful during the empirical model building because they provided different angles to look at the data. Whereas for the quantitative part, the analysis should only include variables conveying information that is helpful for class detection and model creation. Because many of the variables from the data matrix depended on each other or contain similar information in different shapes, they were to distort the results in the quantitative analysis.

As such the goal was to reduce the amount of variables to the smallest required subset that still contains all the information necessary to generate the model. These variables were defined by continuously reducing the variable scope to only the most important features.

4.3 Statistical analysis of the data

The statistical analysis aims to consolidate and strengthen the assumptions drawn during the exploratory analysis. Its goal is to replicate the obtained behaviours from the qualitative approach by analysing different settings quantitatively. Similar to the approaches of Brandtzaeg et al. [BH11] and Gleave et al. [GWLS09] the target is to combine a qualitative exploratory approach with a quantitative statistical one, such as clustering, to define the underlying roles and behaviours.

Using the defined variables from the previous section, a cluster analysis will be conducted for various periods and subsets of users. The analysis will compare results obtained by regular scaled data in contrast to log-transformed data, in an attempt to distinguish the necessity for transformations due to the high existing participation inequality. It will analyse different subsets of users, from all contributing users to only high-performing ones, to differentiate between various active user groups and the large inert majority.

4.4 Merging findings of empirical and statistical approaches

Finally, the findings of the exploratory and qualitative analysis have to be combined with the results of the quantitative analysis. The detected behaviours and roles of these approaches have to be compared, to extract overlaps and further examine differences.

Obvious patterns such as highly active users, as well as those without or hardly any contributions, will help merge the results of both approaches.

Whatever findings or roles are revealed in this part, they are still subject to validation. This will be done in the following step. Successfully discovered roles will be stabilized, whereas unsteady ones will have to be examined further when refining the model.

4.5 Model evaluation

4.5.1 Statistical evaluation - predictive power of the model

To quantitatively evaluate the model, subsets of users will be used to recreate the analysis results, aiming to form a robust classification model. This step will be conducted by repeating the analysis with smaller randomized subsamples to reproduce the cluster solutions, following the approach of Brandtzaeg et al. [BH11].

After extracting the confirmed roles from the analysis, a machine-learning approach will be used to examine its reliability. To evaluate and compare the solutions *precision*, *recall* and *F1 score* will be used to determine the predictive power of the model, as they are commonly used standard measurements for evaluating classifier effectiveness [YL99].

F1, or the harmonic mean of precision and recall, combines the results of these two metrics into one single value between 0 (one of the metrics being zero) and 1 (perfect precision and recall) and was introduced by van Rijsbergen in 1979 [VR79]. It will serve as the main decision criterion for how reliable the model explains the behaviours observed in the data. However, it should not be used as a standalone measurement, instead, all adjacent measures such as precision, recall or accuracy should be taken into consideration as well during evaluation[F$^+$03].

Precision and recall are measures for the quality of the retrieved solutions. Precision can be defined as the portion of retrieved results that is relevant, whereas recall is the portion of the relevant results that was retrieved [LS68].

These measures are defined by the following values commonly used in classifier-related analysis [Faw06], especially binary classification. They can be displayed as a 2x2 confusion matrix resulting from the combinations of the classifier's predictions (positive and negative) and whether the statement is in agreement with the actual observation (true and false) [OD08]:

- true positive: positive observation that is classified as positive
- false positive: negative observation that is classified as positive
- true negative: negative observation that is classified as negative
- false negative: positive observation that is classified as negative

Based on these values, the evaluation measures can be calculated as follows:

Precision (or positive predictive value) describes the fraction of relevant instances among the retrieved instances. It is calculated by dividing the correctly positively identified values by the number of all positively identified values (correct and wrong positives). In short, it provides a measure of how accurately the model identifies positive values without regard for the total percentage of positive values it finds.

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \tag{4.1}$$

Recall (or sensitivity) describes the number of positives the model finds out of all the positives that exist. It does not regard false positives as precision does, instead it only cares about the total amount of positive values identified.

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{4.2}$$

F1 combines the above into one metric, describing how many of the total amount of positives the model identifies but also if these positives are identified precisely or include a lot of false positives.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4.3}$$

By looking at the F1 scores of the individual classes as well as the macro and micro scores for all classes, we can determine the fit of the model and how well it describes the underlying behaviour of the users for the data we analysed.

4.5.2 Iterative refining

To provide satisfying results in the validation, it will be necessary to adhere to an iterative approach of refining the model based on the results of the validation. Therefore findings of the validation steps will be incorporated into improved versions of the model, which in turn will then be validated again until the model accomplishes satisfying results on both statistical and qualitative levels.

Due to the nature of the exploratory phase, iterative refinement will also play an important role during the initial role discovery phase, as the results of the exploratory findings will directly feed into the decisions of the model creation phase.

The resulting model will be applied to various time frames of varying lengths to test its usefulness in explaining the behaviour of the user base over time. The goal will be to determine its applicability in providing long-term user classification and if it can detect changes in behaviour over time.

CHAPTER 5

Data analysis

5.1 Explorative data analysis

5.1.1 Initial findings

The first step consisted of becoming acquainted with the structure of the data and the data itself. It was done by looking into different interaction possibilities users have inside the system and searching for common modes of contributing within various sections of the forum. Although obvious common patterns were noted, at this stage of the analysis it was more important to determine promising and potential variables.

5.1.2 Specification and limitation of variables

As a result of this information gathering step a list of twenty variables was created, as listed in Table 4.1 in Subsection 4.1.1. The goal was to identify potentially important variables, or characteristics, which should later be used in the statistical analysis to classify the users. These variables included information about interactions such as amounts of postings and conversation depth, voting behaviour in the form of incoming and outgoing votes, as well as ratios between these variables.

Whereas the posting and voting behaviour accounted for quick empirical findings, the ratios showed too much interdependence to the original variables, which resulted in no additional information gain when including them in the analysis. Therefore these variables were emitted from the subsequent analyses and the variable set was reduced to a minimal number of important ones, which appeared to be the most promising for the statistical analysis.

The remaining variables are the following:

Postings The amount of postings a user has submitted for the given period. Postings include direct comments on the article as well as responses to other users' comments, so-called *nested comments*.

Number of articles commented The number of different articles a user has placed comments on, either as a comment to the article or as a response to other users' comments, for the given period.

Number of top-level postings The amount of postings that were written directly below the articles, a so-called *top level comment*, opposed to a posting that was written in reply to another user's posting.

Positive votes given by user The number of positive votes the user has given to postings of other users for the given period.

Negative votes given by user The amount of negative votes the user has given to postings of other users for the given period.

Positive votes received by user The amount of positive votes the user has received for his postings by other users for the given period.

Negative votes received by user The amount of negative votes the user has received for his postings by other users for the given period.

5.1.3 Empirical findings of common behaviours

Users in the researched context have two options for interaction. First, they can write postings, either just as a standalone comment or as a reply to another user's posting. Secondly, they can vote on postings of other users. They can give them a positive or a negative evaluation. While the official description of the voting options reads "*worth reading*" and "*not worth reading*", users are likely to also use them for different reasons, such as agreeing/disagreeing or liking/disliking content.

Due to these options of interaction, the above set of variables was selected to be used for the analysis. As a first step of the initial qualitative approach, these variables were further reduced and their values were transformed into discrete values on a fixed scale. The variables were the *amount of postings*, *amount of distinct articles commented on*, *type of postings* (whether they were replies or not), *amount & polarity of given votes* and *amount & polarity of received votes*. The values respectively were on a scale from 1 to 5, representing the absolute amount in comparison to all the other users. The values of the votes were additionally transformed into their polarity, highlighting the amount and direction of the votes, e.g. +++ (average positive), −−−−− (high negative), +− (low mixed).

This manual analysis was conducted for a subset of the data consisting of one day of moderately to very active users. Its purpose was the examination of obvious patterns

within the data and the early detection of potentially promising behaviours. It was conducted because it allowed for rapid insight into the bigger picture during this exploratory phase while hinting at potential behavioural patterns.

All these smaller indications were noted and named, resulting in the following batch of observed recurring behaviours.

- high posting amount
- very low posting amount
- low amount of different articles
- high percentage of reply postings
- high outgoing vote amount
- very low outgoing vote amount
- high positive incoming vote amount
- high negative incoming vote amount
- high amount of incoming votes (positive and negative)

These were the salient behaviours that insinuated specific patterns, besides these users showed all kinds of combinations of the variables in their communication. But the above behaviours appeared much more regularly, suggesting an underlying pattern.

Additionally, these patterns were analysed on co-occurrences, which resulted in a significant amount of co-occurrences of various combinations. These frequent appearances were used alongside the lone behaviours to create the first qualitative classification.

5.1.4 First qualitative model

The seven resulting classes, which were named based on their characteristic behaviour and if possible alongside literature counterparts, are displayed in Table 5.1. Two of these classes, the **Opinion Leader** and the **Polarizer** were only defined by one variable, the amount of votes received. The difference between the two lies in the polarity of the votes. **Opinion Leaders** generate almost exclusively positive votes, while **Polarizers** have a relatively even split between positive and negative received votes. All the variables that are not explicitly named for the classes can have arbitrary values, but the highlighted variables are the ones indicating the pattern.

Power Users and **Posters** appear very similar in the sense that they have a high amount of postings. Their main difference lies in the voting behaviour, that is the outgoing votes these users provide. While **Power Users** also vote a lot, **Posters** vote very scarcely, if at

class	characteristics
Power User	high posting amount, many outgoing votes
Opinion Leader	high amount of positive incoming votes
Polarizer	high amount of positive and negative incoming votes
Poster	high posting amount, very low amount of outgoing votes
Disturber	high posting amount, low amount of different articles posted on, high percentage of postings are replies, high amount of negative incoming votes
Silent Voter	very low posting amount, high amount of outgoing votes
Discussion Leader	high posting amount, low amount of different articles posted on, high percentage of postings are replies

Table 5.1: Initial qualitative classification model.

all. **Silent Voters** are the exact opposite of **Posters**. They hardly write any postings, most do not write any at all, but they are actively voting on postings of other users.

The remaining two classes are once again quite similar, as they both show tendencies to write many postings, but only on a small number of different articles, with many of their postings being replies to other users' postings. These classes are called **Discussion Leader** and **Disturber**. Their main difference lies in their perception by other users. While **Discussion Leaders** do not show tendencies of being well-liked or disliked per se, **Disturbers** attract mostly negative evaluations by their peers.

While these seven classes already include combined behaviours, there are still co-occurrences of some of these classes within the same user. Popular groupings are for example **Power Users** and **Opinion Leaders**, **Discussion Leader** and **Opinion Leader** or **Discussion Leader** and **Polarizer**.

Observed similarities to literature

The above behaviours only reflect observations within this dataset. Yet, some of the behaviours show strong similarities to classes of the researched literature. Users who focus on discussions, while others do not actively contribute with words, but consume and evaluate others' content, as well as very interactive users, all these show strong ties to roles found in literature.

Additionally, low-impact users have implicitly been detected. Users with low contributions were deliberately excluded during this exploratory analysis to concentrate on the more obvious behaviours. Yet, by omitting these users they were indirectly also grouped into one large role of scarcely contributing users. These users are also backed up by research, as seen by Chan et al. [CHD10] in the form of *Grunts* and *Ignored*.

class	characteristics
Lurkers	no postings, no votes
Taciturns	very low amount of postings and votes
Disturbers	high posting amount, low amount of different articles posted on, high percentage of postings are replies, high amount of negative incoming votes
Provokers	high posting amount, low amount of different articles posted on, a high percentage of postings are replies, high amount of positive and negative incoming votes
Flamers	low posting amount, very high amount of negative incoming votes

Table 5.2: Additional classes for the qualitative classification model.

5.1.5 Updated qualitative model

By including the learnings of researched literature, the model was adapted to include low-activity users. This mostly silent majority was divided into two groups: **Lurkers** and **Taciturns**. While **Lurkers** are not interacting at all (no postings, no votes), **Taciturns** have at least some contributions (votes and/or postings). The added classes are displayed in Table 5.2.

Besides including the low-activity users, a segmentation of **Disturbers** was conducted. Following the distinction of Golder et al. [GD04] these negatively evaluated members were further grouped into **Disturbers**, **Provokers** and **Flamers**. **Disturbers** remain characterised the same way as before and resemble closely what Golder et al. considered to be a *Troll*. **Provokers** show the same characteristics in terms of communication volume, limiting the conversation to a few articles and having a high reply percentage but differ in the way they are evaluated. Instead of receiving only negative votes, they show a more even split of receiving votes, suggesting that while their tone or opinion is not popular they also have support from other users. **Flamers** on the other hand are not interested in a dialogue, they only post a limited amount of content but receive very high amounts of negative votes for their comments.

The updated qualitative model contains 12 roles. This model will serve as the basis for the following statistical analysis.

5.1.6 Shortcomings and potential pitfalls

Due to the selected approach for this initial model, it lacks explanatory power and has various shortcomings. First of all the model is not validated. Even though it is based on objective criteria, it is still heavily influenced by subjective role definitions. Additionally, its focus is on more active users, as users with low contributions do not provide sufficient characteristics to observe necessary behaviours. This early stage deliberately inspects for obvious behaviours. Finally, it also lacks the necessary differentiation between what is considered a high or a low value, e.g. in the posting amount of **Power Users** and

Posters. While both classes show this behaviour as a typical trait, the extent can still differ by great margins.

5.2 Statistical analysis

The statistical analysis aims to consolidate and prove the empirical findings of the qualitative analysis. It will be conducted by performing a hierarchical cluster analysis. Blashfield and Aldenderfer describe clustering as dividing extensive samples of comprehensive datasets into smaller groups with similar characteristics [BA78].

The resulting dendrograms of the hierarchical clustering then have to be interpreted, to detect classes with good fit. A good fit is achieved when a further splitting of the group does not increase the information gained about a group of users.

The expected result of the statistical analysis should be a cluster solution that explains various general user behaviours in the context of a news forum community. A good solution for explaining such behaviours lies around four to six classes according to [BH11, p. 37]. To find this ideal amount, several cluster solutions have to be analyzed, and multiple cluster solutions have to be considered in each iteration. It is important to note that these boundaries should act as guidance towards finding an expressive number of roles, but are not meant as hard limits. Horrigan et al. formulate the ideal cluster size as creating *"[...] cohesive groups that were distinct from one another, large enough in size to be analytically useful, and substantively meaningful."*[Hor07, p. 7]

The cluster analysis may produce solutions which are similar in behaviour or fit the same literature-based roles. In such cases it is important to recognize these similarities and when applicable merge these clusters into one role. This approach was also selected by Chan et al. [CHD10, p. 216] when attempting to discover roles based on varying cluster sizes.

5.2.1 Hierarchical clustering

Clustering is the process of allocating similar points of data into distinct groups. Cluster analysis describes various procedures to discover classifications within complex data sets [Gor00]. It is the method of finding or detecting these clusters within a heterogeneous set of data. Breaking this data into uniform and homogeneous segments can be achieved with various techniques, depending on the data provided and the output sought after [KR09].

In this thesis a hierarchical cluster analysis was employed to discover groups of similar behaviour, attempting to evaluate the roles from the exploratory analysis, to confirm, discard and augment the already existing roles with the results of the clustering.

Hierarchical clustering is a method of cluster analysis that follows an iterative approach to finding suitable cluster solutions. The result of the analysis is a dendrogram, which gradually divides the data into more and more cluster solutions. Comparing it to other

solutions, such as the k-means cluster analysis, hierarchical clustering provides a multitude of potential cluster solutions from which the researcher has to determine the correct one.

In contrast, k-means requires the cluster size upfront and then allocates the data points into the specified amount of groups. Thus, hierarchical clustering provides the right solution for the exploratory nature of this classification work. In addition, hierarchical clustering provides information about the clustering solution it combines with each step, making it easier to comprehend the distinct characteristics of each role and aids in deciding which granularity of roles is required.

The hierarchical clustering in this work was conducted in R by using the `hclust` method. The agglomeration method used was `ward.D2`, as it is regarded as one of the popular suggested solutions alongside `ward` and produced the most meaningful results during the analysis. Figure 5.1 shows an exemplary dendrogram for a subsample of one week of data.

As such the clustering follows a bottom-up approach, starting with individual cluster solutions for each data point and continuously merging them until only one cluster remains. It is then that the researcher has to determine the correct cluster solution, by defining the appropriate cut-off point on the dendrogram, which corresponds to the desired amount of cluster solutions.

The cut-off point for the resulting dendrograms in this work was selected based on the guiding principles by Brandtzaeg et al.[BH11], as discussed above, in an attempt to maximize the meaningful roles without dividing similar patterns into too many subgroups. It was performed using the `cutree` method on the resulting dendrograms in R.

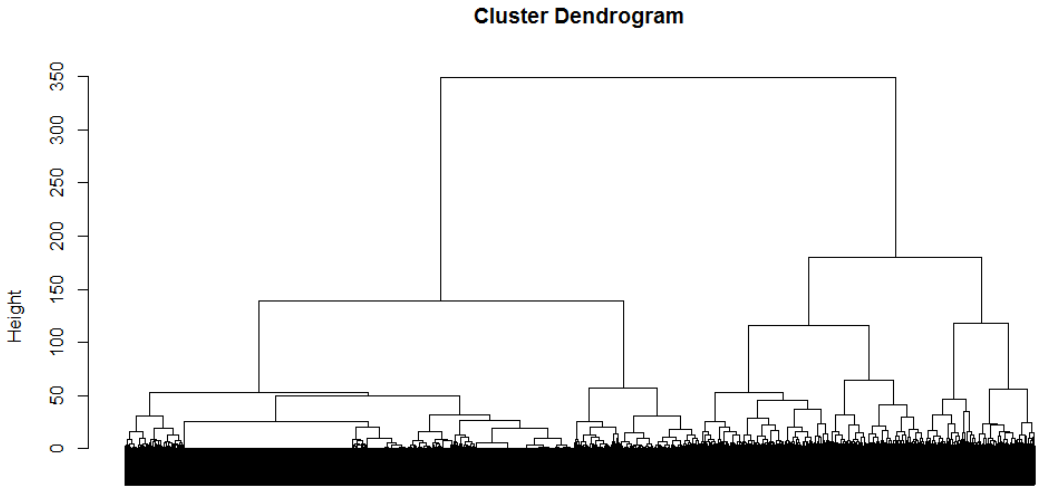


Figure 5.1: Hierarchical clustering dendrogram for a subsample of one week.

postings	articles	top level	given (+)	given (-)	received (+)	received (-)
13	7	3	0	0	102	11
3	3	2	156	42	6	0
183	19	99	87	0	233	10
37	30	21	7	2	158	57
65	33	26	154	30	429	32
165	34	8	213	154	825	131
...	...	...	...	...	...	...

Table 5.3: Example of weekly aggregated user data used for the analysis.

Handling outliers

One caveat of these clustering solutions is that outliers can, due to their behavioural specifics, form their own clusters. While these outlier roles do portray unique behaviours, they do not provide generally applicable roles, which is what this work is after. Due to this, it is not enough to rely on the solutions provided by the clustering, it is a necessity to analyse the solutions afterwards and make sure that the roles contribute in unison to the desired model.

Distance matrix

A requirement for hierarchical clustering is to first build a distance matrix of the values that should be analysed. A distance matrix is a square matrix containing the pairwise distances between all elements from the set. Based on these distances the users are then allocated into groups with the smallest possible distance during the clustering.

The distance matrix was calculated with the `dist` function in R, using the log scaled and subsampled data from the user matrix as described in Subsection 5.1.2. To find robust clusters, the dendrograms were built for aggregated user data based on different periods between one day and six months.

5.2.2 Data used for analysis

The data consisted of user metrics that were aggregated for the respective selected timeframe (see following section). Each row of the data consisted of metrics for one user, as seen in table 5.3. The variables used were: *number of postings*, *number of distinct articles posted on*, *number of 'top level' postings (= not reply to other user's posting)*, *number of positive votes given*, *number of negative votes given*, *number of positive votes received* and *number of negative votes received*. The data was log-scaled to account for the skewed distributions within the data.

5.2.3 Timespan and period decision

In consultation with experts of derStandard.at periods of at least one week were chosen as good time frames in terms of meaningfulness. The reasoning was the natural variations

in activity during the week. Such differences are also notable during different daytimes, but they also occur on different weekdays or the weekend. To analyze various extents of users several timespans were chosen to be examined.

- **1 day** as comparison for qualitative analysis
- **1 week**
- **2 weeks**
- **4 weeks** (approx. 1 month)
- **8 weeks** (approx. 2 months)
- **12 weeks** (approx. 3 months)
- **26 weeks** (approx. 6 months)

These periods aim to solidify the cluster solutions, as well as determine potential discrepancies between short to medium terms (1 week to 4 weeks) and longer periods (2 to 6 months).

5.2.4 Cluster results

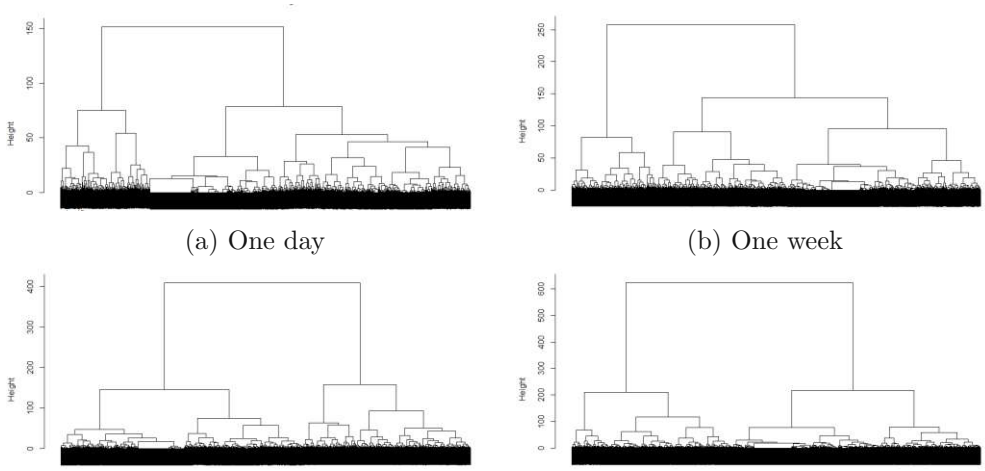
The above seven time periods were used to conduct the hierarchical cluster analysis with both absolute and log-scaled data, as well as once more with only *active* users, which are users who had at least one activity during each week of the respective period. In total, this resulted in 21 cluster dendrograms based on which the appropriate roles should be defined.

The most promising results were shown in the solutions of log-scaled data with shorter periods (1 to 4 weeks). Figure 5.2 shows four exemplary dendrograms of log-scaled data for different periods. All three structures start at height zero with individual clusters for all data points in the respective sample. With each step in height, clusters are merged based on their distances in the distance matrix, grouping results with the least distance.

As mentioned above in Subsection 5.2.1 the cut-off point then has to be selected to contain the desired amount of cluster solutions. This was done in R by using the `cutree` method on the existing dendrograms. The height must be individually chosen based on each dendrogram so that it fulfils the required number of clusters.

Cluster recognition

The result of cutting the dendrogram is a table consisting of the cluster solutions and their values for each of the variables of the analysis, as seen in Table 5.4. This table can then be used to interpret the solutions based on the variables and their corresponding forum behaviour.



(a) One day (b) One week

(c) Four weeks (d) Six months

Figure 5.2: Hierarchical clustering results compared.

	postings	articles	top level	given -	given +	received +	received -	role
1	-0.8449	-0.8425	-0.6688	-0.849	-0.6813	-0.8638	-0.6292	**Taciturn**
2	0.1009	0.1263	0.0498	-0.3443	-0.5757	0.2652	-0.0632	**Regular**
3	1.2709	1.4211	1.3827	-0.4401	-0.5533	1.1381	1.3467	**Celebrity**
4	-0.6897	-0.6928	-0.6678	0.5991	0.5657	-0.6496	-0.6231	**Silent Voter**
5	1.5137	1.4088	1.5185	1.2749	1.2298	1.332	1.3714	**Power User**
6	0.5053	0.4959	-0.0379	0.9622	1.1154	0.4962	0.2338	**Conversationalist**

Table 5.4: Cluster solutions for log scaled data of one week period (Figure 5.2b).

These tables combined with the results of the initial exploratory model creation are the basis for the final classification model. Based on the proposed clusters and the earlier determined classes the solutions of the exploratory and statistical analysis were merged and roles were defined based on these behaviours.

The focus during this role determination was on repetitive and distinct behaviours found throughout all, or most, of the analysed cluster solutions. As the analysis was conducted for numerous periods, recurring behaviours were of more significance than one-off cluster solutions.

Combining the early model of Subsection 5.1.4 with the results of the clustering solutions, leads to the roles as listed in the last column of Table 5.4. Some roles were confirmed as they were, for example **Power User** and **Silent Voter**. Others were partially confirmed or complimented, and due to better matching with existing literature renamed, such as **Opinion Leader** to **Celebrity** and **Discussion Leader** to **Conversationalist**. **Taciturns** and **Regulars** were added to the model following the literature, and are additions to the exploratory model. The reason they were not found during the manual classification analysis was due to their unobtrusive behaviour, whereas the other roles were defined by prominent behavioural characteristics.

Finally, some roles could not be confirmed as generally applicable, such as **Polarizers** or **Disturbers**. These roles were sometimes part of clustering solutions but appeared to be outliers, which hints at them as outlier roles, but not at them being part of the general classification model.

Considerations about solution

The main reason for analysing different timeframes is to stabilize the resulting clusters. Based on the observations during the analysis two things have to be considered with this approach. Under the premise that looking at a longer period makes the clusters more accurate, we expect one of two things to happen. First, an affirmation that users have a class that represents their behaviour and observing a longer period solidifies this result. Or, on the other hand, users do not take up one single, distinct role but show varying patterns in their behaviour. A longer period would then potentially average out the results of the cluster solutions.

The second case could hint at two phenomena in behaviour. First, it could mean that a user's behaviour changes over time, converging towards an inherent role that includes several other roles along the way. It could also mean that over time, users will engage in different contexts, in which they will show different kinds of behaviours, resulting in several distinct behaviours shown by each user. Both of these effects would have an impact on the cluster solutions for too lengthy periods.

A big uncertainty of longer periods is without doubt the unsteady participation of average users. Most users do not participate every single day in such an environment, not even every week (as defined as the smallest viable period for this thesis). Therefore, the overall user contributions are most likely rather low for each single user when observed over a long period. Even users that were extremely active for four weeks, would only show moderate numbers when included in a six-month time frame if they had not posted anymore after the initial four weeks.

In an attempt to reduce the latter effects, the analysis was additionally also conducted for only users that were active over six months. *Active* describing that they had at least one contribution in each week of this timespan.

5.3 Model creation

As described in Section 5.2.4 the classification model was built by combining the results of the exploratory approach and the clustering solutions. The roles were then compared with literature suggestions to produce a robust set of classes that describe the general behaviour of the analysed users. The final roles are Lurker, Taciturn, Silent Voter, Regular, Conversationalist, Power User and Celebrity. They are detailed in Section 6.1.

In order to make long-term analyses of the roles a classification model was trained. This model was then used to assign all users of the long-term data set a new role for each week of data, as this time frame was defined as the minimum amount.

5.3.1 Manual role assignment and supervised learning

To train the model one full week of data was manually classified, consisting of 2570 users, using the roles defined in the previous section. This set was subsequently used to train the model using the `k nearest neighbours` method. Finally, the model was applied to the whole user matrix, as described in Table 5.3, consisting of weekly data for each user for the whole duration of data provided.

The result was a matrix consisting of the variables defined in Subsection 5.1.2 for each week and each user with their respective roles for the behaviour shown during this week. This matrix is the basis for the long-term analysis of Section 6.2.

CHAPTER 6

Results

6.1 News forum typology - the classification model

Based on the analysis in Chapter 5, the results of the combined, exploratory, and statistical, approach are the following seven classes

- Lurker
- Taciturn
- Silent Voter
- Regular
- Conversationalist
- Power User
- Celebrity

that will be described in more detail in the following sub-sections.

These classes depict the smallest subset of reliably detectable behavior, given the parameters of the analysis. These roles do not claim to explain every possible behavior displayed by users. Instead, they attempt to explain the larger, overarching behavioral patterns that can be encountered in similarly designed systems.

All classes were selected and named following existing literature, but will, due to the differences in analysis setup, differ from their literature counterparts in details.

6.1.1 Role distribution

As stated in Section 3.4 participation among users is not distributed equally, and their role assignment follows this behavior. Figure 6.1 shows the average distribution of the six active roles (see list above without Lurker). A deeper insight into the relative changes over time is given in Section 6.2, which looks at the role distribution throughout the lifetime of the news forum.

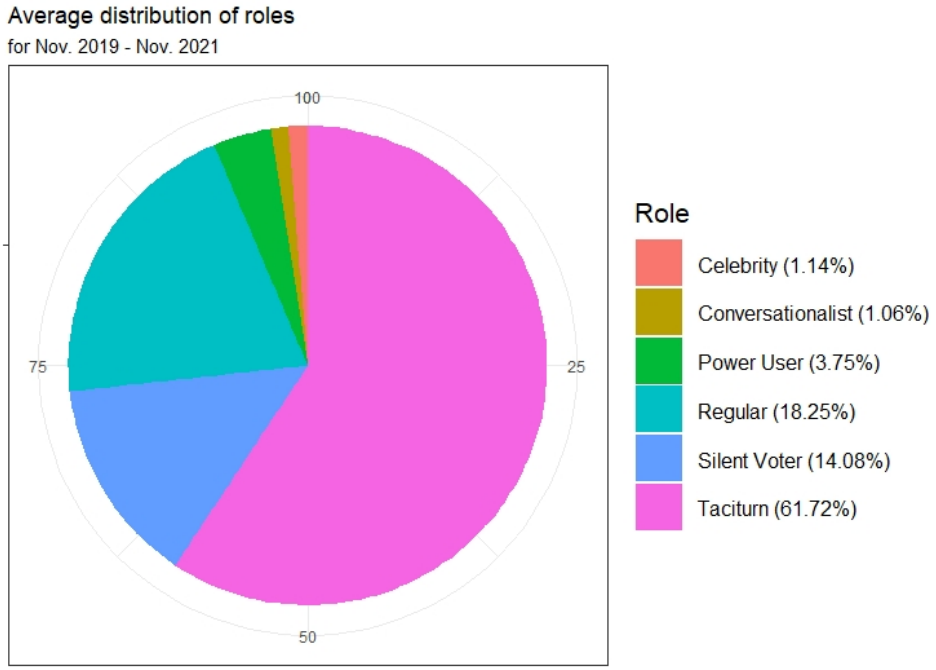


Figure 6.1: Average role distribution (Nov. 2002 - Nov. 2021).

6.1.2 Lurker

Percentage: -

Characteristics: **no interaction, only reads**

Lurkers are the only inferred role in this analysis. They are well defined in literature and despite not being included in the statistical analysis they have to be included for the sake of completeness.

Lurking describes the behavior of following the discussions in the forum without making themselves known. They comprise all readers of the news forum who do not actively participate in any means.

They make by far the biggest group but their numbers are hard to measure, as their non-existent behavior requires other means of user tracking. By broadening the definition

of Lurkers to not only readers of the forum but to all readers of the news article, a simple approach to measuring them would be to count all readers through view tracking and subtract the number of active users from that. This approach would also include readers who do not read the news forum postings. A stricter measuring scheme would require measuring how far users scroll on the page and whether they scroll onto the news forum section of the page.

Literature counterparts: Lurker [BH11, Nie06, GD04], Inactive [LB07]

6.1.3 Taciturn

Percentage: 61.72%

Characteristics: **low interaction (posting and voting)**

Taciturns are similar to Lurkers in that they are a low-activity role. The main difference between them is that Lurkers have zero activity in the news forum, whereas Taciturns have little to almost no activity, yet they do participate.

As Subsection 6.2.1 highlights later, Taciturn is the most common role to appear alongside other roles. While the analysis did not take a look at the reasons for these role switches, it makes sense that the lowest activity role is most often observable alongside others. Because ultimately reducing any behavioral pattern long enough will result in the low activity representing this role.

While shifts in behavior can be observed, albeit seldom, it is easier to imagine a person to reduce the volume of their activity than to follow a different interaction pattern.

Literature counterparts: Sporadic [BH11], Newbies [GD04], Taciturn [CHD10], Tourist [Koz99], Spectator [LB07]

6.1.4 Silent Voter

Percentage: 14.08%

Characteristics: **exclusively votes on postings, does not write postings**

Silent Voters are one of the biggest classes in the context of this news forum. They are silent judges of the postings other users write. Some members of this class do occasionally post to the forum, but the large majority of Silent Voters never contribute any postings. Instead, their social activities are limited to reading and voting on the postings of others.

A case can be made that Silent Voters are in between Taciturns and Regulars when it comes to participation behavior. Voting on other people's posts takes less effort than sharing one's viewpoint or contributing to a discussion. It is a low-friction means of participating compared to writing postings. They actively contribute to the forum with their votes, showing agreement or disapproval of opinions through the content of others, but do not go as far as voicing their own opinion.

It is important to note that even if voting is considered a less involved form of communication, it can not be inferred that Silent Voters are less active regarding interaction volume. On the contrary, many Silent Voters seem to be highly active within the forum.

Literature counterparts: Devotee [Koz99], Collectors [LB07]

6.1.5 Regular

Percentage: 18.25%

Characteristics: **writes few postings, votes occasionally, low overall volume, few incoming votes**

Regulars make up the largest share of the active user base. They are what can be described as the average participating member. They do write postings and vote on postings of others, but do it with low volume. Their content receives few incoming votes, positive and negative.

In regards to their behavior, they are most similar to Power Users. Their main difference lies in the volume of interactions.

Literature counterparts: Intermittent Contributor [Nie06], Supporter [CHD10], Mingler [Koz99], Critics [LB07], Question People [GWLS09]

6.1.6 Conversationalist

Percentage: 1.06%

Characteristics: **many postings, nested conversations, little outgoing votes**

Conversationalists are very engaged users of the forum. Their main means of interaction is by posting replies to the postings of others. While they do post top-level postings (or direct comments to the article) as well, most of their postings are part of nested conversations. Their behavior is very reciprocal and requires other users with similar posting behavior to flourish.

A Conversationalist can be regarded as the counterpart to Silent Voters. While their postings frequency is high, they do not engage as much in voting on the postings of others. As such their behaviour is mirror-inverted compared to that of Silent Voters.

Without looking further at the context of the postings, it is not possible to conclude the nature of these deeply nested conversations, whether they are held to convince other people or because they are inherently interested in the exchange of opinions. It can be observed, though, that the majority of the, albeit low, incoming votes are positive, which serves as a hint that these discussions might be positive in general.

Literature counterparts: Debater [BH11], Joining Conversationalists [CHD10], Elitists [CHD10], Joiners [LB07], Discussion People [GWLS09]

6.1.7 Power User

Percentage: 3.75%

Characteristics: **writes postings and votes on others' postings, high overall volume**

Power Users are highly interactive users of the forum. They are very engaged with other users participating in conversations and voting on postings of others. Their behavior is similar to that of Regulars in the sense that they both post and vote regularly. The main difference between the two classes is the volume. Power Users have very high volumes both in the amount of postings and the number of outgoing votes. In return Power Users also receive a higher number of votes, which corresponds to the frequencies with which they post.

Compared to the Celebrity the Power User shows two main differences. First, the overall amount of incoming votes is the main differentiator. Celebrities tend to accumulate a much higher number of incoming votes compared to the amounts of postings they write. Secondly, Power Users vote more on the postings of other users compared to Celebrity.

Power Users contribute a large portion of the content to the forum. While they are much fewer in number compared to the Regulars, their posting frequency more than makes up for their small share.

Literature counterparts: Actives [BH11], Heavy Contributor [Nie06], Popular Participants [CHD10], Insider [Koz99], Creator [LB07], Answer People [GWLS09]

6.1.8 Celebrity

Percentage: 1.14%

Characteristics: **writes a lot of postings, receives a lot of votes**

Celebrities are a tiny fraction of the active user base but receive a disproportional share of the votes on postings. They are characterized by a high posting frequency and by lots of incoming votes. The majority of votes for this role are positive, which distinguishes them from negative outlier roles such as trolls (which are not part of this analysis).

They are similar to Power Users but tend to vote less frequently on the content of others. Instead, it is their own content that attracts most of the votes of other users. Their posting frequency is similar to that of Power Users, though.

Literature counterparts: Celebrities [GD04], Popular Initiator [CHD10], Discussion Catalysts [GWLS09]

6.2 Classification of users over time

Besides asking for a classification model for users of an online news forum, this work aims to explore the long-term development of the roles these users inherit. Looking at

1 role	2 roles	3 roles	4 roles	5 roles	6 roles
367051	72535	37517	14677	4617	1522

Table 6.1: Number of users and their amount of roles (Nov. 2002 - Nov. 2021).

the classification over time might surface large-scale trends in forum user behavior or give insight into the role progression of average forum users.

Over time, not only has the analyzed news forum developed itself further, but the adoption of such forums in general has increased. Due to the extent of the data provided for this analysis, it is possible to apply the above classification model to the past 20 years of the news forum.

The main questions that are attempted to be answered are:

- Do users have one role or does their classification change over time?
- If users inherit more than one role, which roles do they occupy most together?
- Is there a long-term shift in communication behavior? And does therefore the overall distribution of roles change with time?
- Are roles inherent to users or does their behavior, and thus their role assignment, change over time?

6.2.1 Inherent roles vs. role switching

The first question looks at whether users are classified with the same role over time or if they show changes in behavior, resulting in different assigned roles. To answer this question the amount of role switches will be measured.

The first step is applying the model to the long-term forum data. For each week users will be assigned a role based on the classification model and their behavior. Next, the number of different roles will be counted for each user (see Table 6.1), as well as the number of role switches for each user.

A high number of different roles per user might be an indicator that communication behavior varies greatly among users. This could hint at context-dependent roles, and depending on where in the forum the users engage their behavior differs, eg. varying topics. A low number of roles per user, on the other hand, might indicate that users have a generally applicable underlying role that explains most of their behavior.

As Figure 6.2 shows role changes are common, but the majority of users only inherit one single role. Less than 12% of users show behavior that puts them into more than two classes. Based on these numbers one must assume that users tend to have a relatively fixed style of communication behavior.

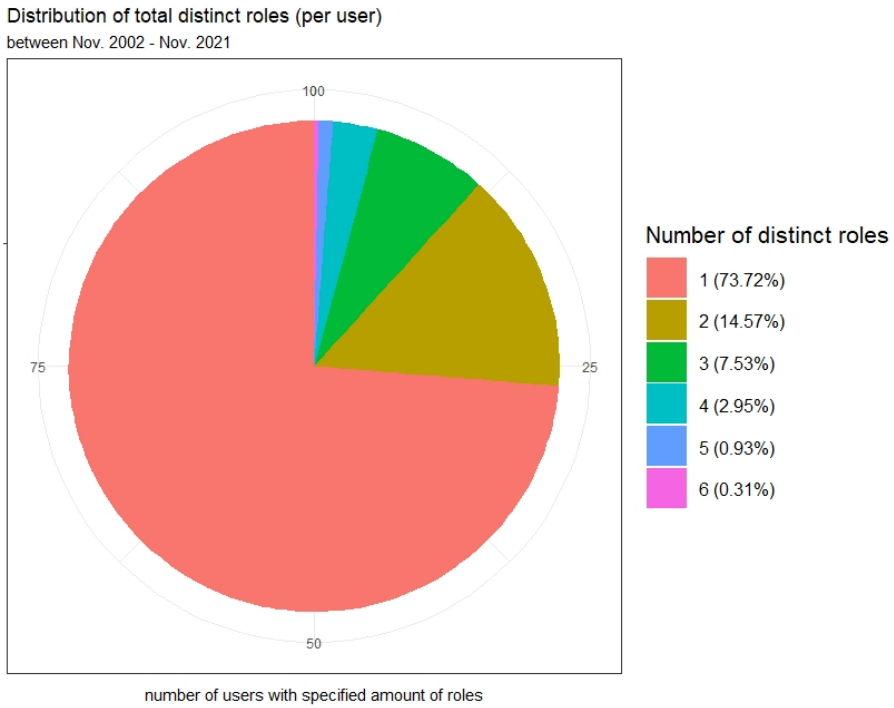


Figure 6.2: Distribution of distinct roles per user (Nov. 2002 - Nov. 2021).

Another perspective is to look at the distribution of distinct roles for only active roles. Because we already know from Subsection 6.1.1 that Taciturns make up the largest percentage of users and that they are highly inactive, we can look at the distribution without them. As Taciturns are defined by the absence of behavior, instead of by a different type of behavior excluding them from the analysis might prove valuable.

Figure 6.3 shows the distribution without Taciturns, as such the maximum number of distinct roles a user can have dropped from 6 to 5 roles. It is important to note that for this analysis, all rows in which users were assigned the Taciturn role were filtered. It is not just filtering users that only inherit the Taciturn role, while users with Taciturn assignments and other role assignments would be included.

By comparing the two results, we can observe that the distribution among the active classes appears more balanced. It still shows a clear trend towards fewer roles, but almost half the users show behavior of more than one distinct role. These results indicate that role changes are more common among active roles compared to inactive roles. An explanation might be that the inertia required to become an active participator is higher than switching between different forms of activity.

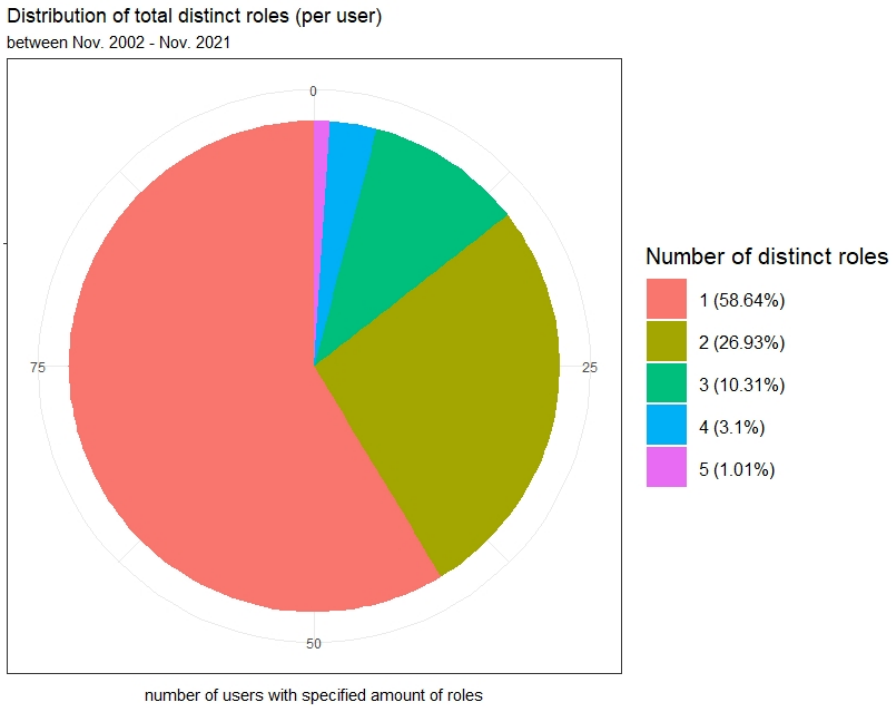


Figure 6.3: Distribution of distinct roles per user, without Taciturns (Nov. 2002 - Nov. 2021).

Absolute role changes

Finally, looking at the total amounts of role changes per user visualized in Figure 6.4, we get a clear picture that the majority of users remain with one inherent role. All the users that do in fact change roles, appear to do so very infrequently. Instead of hopping from one role to another, the majority of role-switching users tend to do so only occasionally. A small number of users show a high number of role switches, though, depicted by the long drawn leg of the chart to the right.

It is important to note that a high number of role switches could occur for a user who cannot be perfectly identified by the model. If their behavior borders two similar classes, it could also happen that the classifier repeatedly assigns it different roles for each week. In such a case, the user would still show a recurring behavior but be labeled with different roles.

6.2.2 Co-occurrences in roles

The second question is concerned with patterns in mutually occurring roles. The classification model includes roles that are similar in behavior but vary greatly in volume, but also roles that show very different means of interaction.

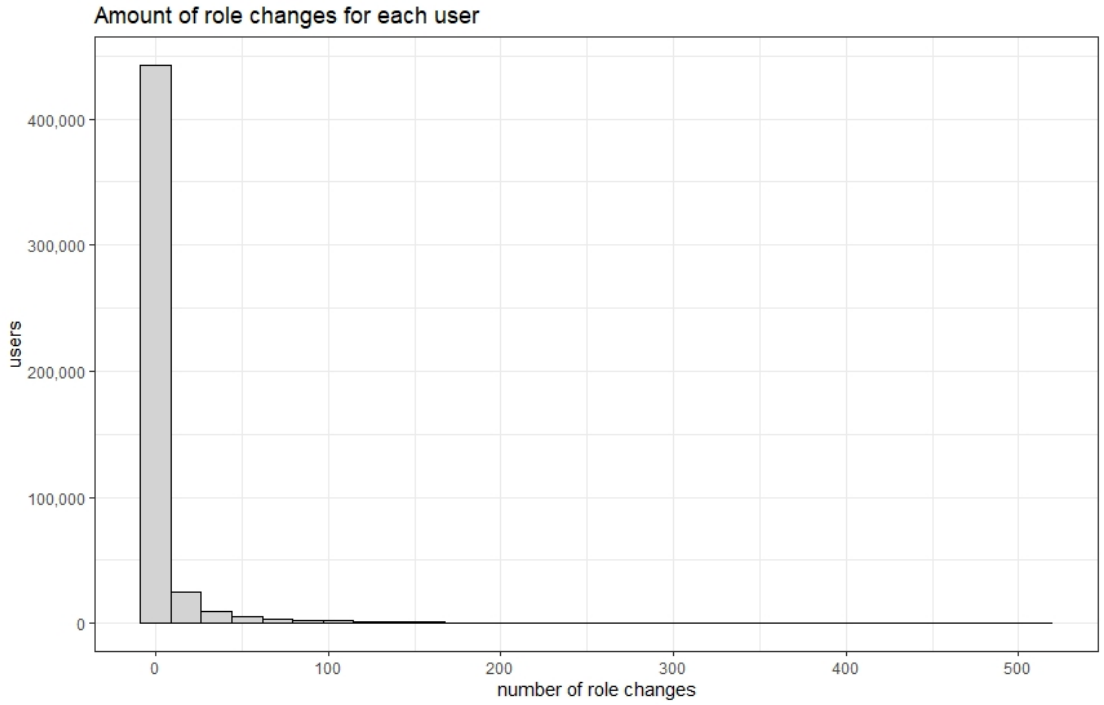


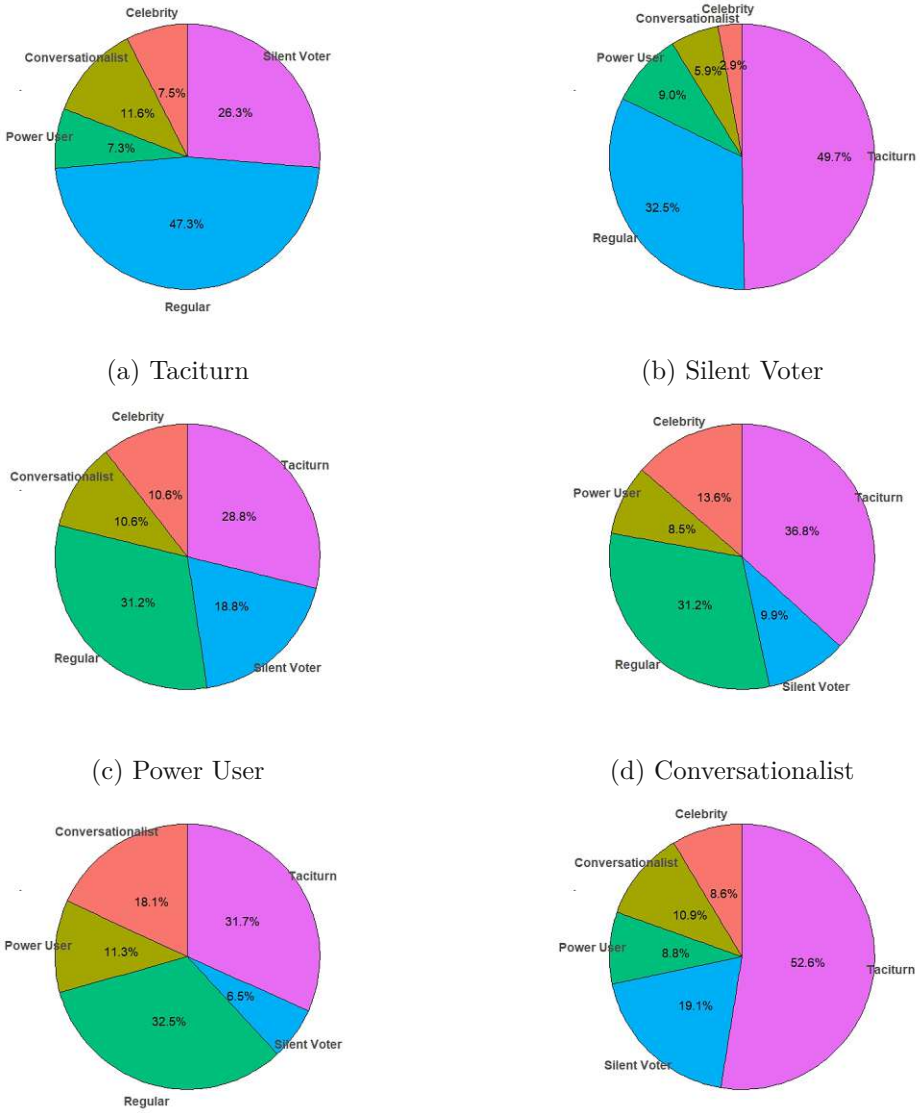
Figure 6.4: Total amount of role changes for each user (Nov. 2002 - Nov. 2021).

One assumption might be that given users tend to have an inherent underlying class, that co-occurrences in roles should mostly occur among classes with similar behavioral patterns. Similar types of roles include Power Users & Celebrities, Regulars & Power Users, or Taciturns & Regulars.

Figure 6.5 displays the distribution of pairwise co-occurrences for each role. The higher the percentage for a role, the more co-occurrences could be observed for a specific combination. As we can see for Taciturns in Figure 6.5a the two biggest co-occurrences, which make up almost 75% are Regulars and Silent Voters. Both of these classes were considered related in behavior (see Subsection 6.1.4), in regards to them being lower effort social activity roles in this news forum context.

The most equal distribution is shown by Power Users, as seen in Figure 6.5c. As discussed in Subsection 6.1.7 the main characteristic of Power Users is their engagement volume. Looking at the distribution in this graph, it might hint at the fact that Power Users are less defined by the way they engage, but by volume only. As such, the co-occurrences could be explained by lower or higher activity in some weeks than in others.

Conversationalists appear to confirm the assumption that they form the anti-role compared to Silent Voters, see Subsection 6.1.6. Despite Silent Voters being one of the roles with the most assigned users, their share among Conversationalists is very low in comparison. The lowest co-occurring role is Power Users, which show high numbers of outgoing votes



(a) Taciturn

(b) Silent Voter

(c) Power User

(d) Conversationalist

(e) Celebrity

(f) Regular

Figure 6.5: Distribution of pairwise role combinations for each role.

as well, thus further confirming these behavioral patterns as opposing.

In summary, it is noticeable that the most co-occurrences happen between any roles and Taciturns, which confirms that users of all roles have low activity phases, which places them into the Taciturn role. Additionally, the role with the least co-occurrences is Power User, which is the exact opposite in regards to the interaction volume.

Secondly, it appears confirmed that while many users occupy more than one role, they are

more likely to branch out into roles with similar behavioral traits, than to communicate in completely different styles.

6.2.3 Long-term role distribution

After discussing the individual role assignment, it is necessary to expand this analysis to the overall distribution of roles over time. This long-term analysis spans more than 20 years, beginning in 1999. As such, it illustrates not just the various changes the news forum itself implemented, including various redesigns of the news forum, but also the superincumbent shift in digitalization. From a time when the internet was accessible, but still limited in daily use and often limited to home computers, to a time in which it plays an integral part of everyday life for many people with constant and mobile access to it.

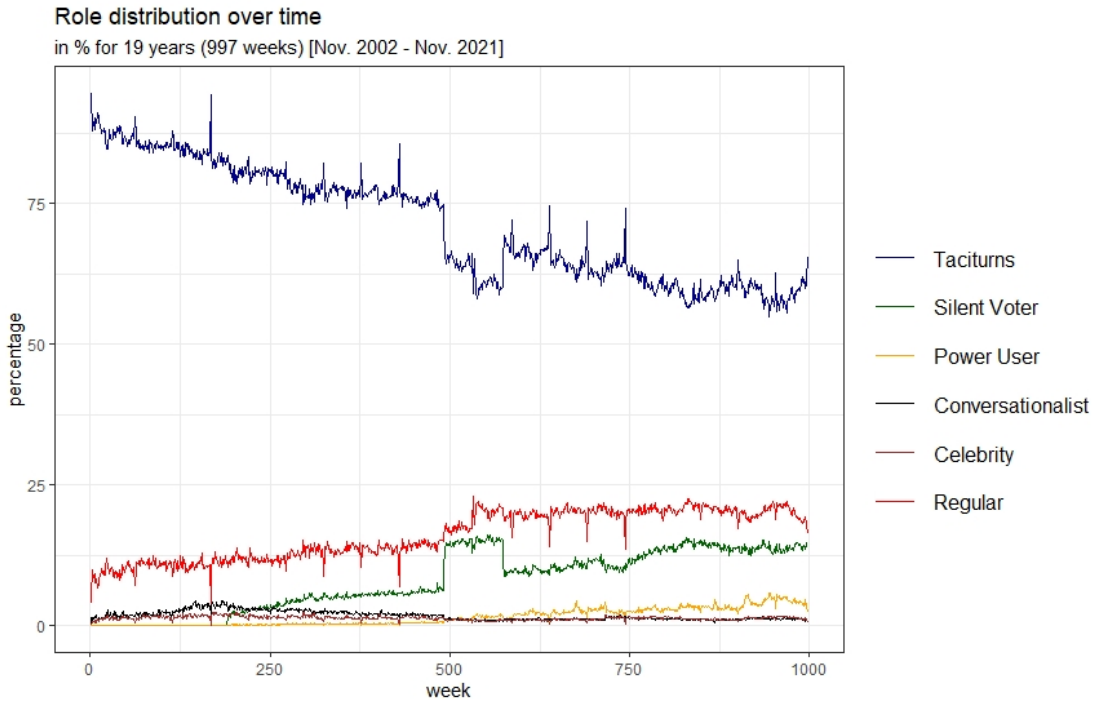


Figure 6.6: Long-term distribution of roles (Nov. 2002 - Nov. 2021).

Figure 6.6 shows the overall distribution of roles over time. The following will go into detail about two important observations:

- First, the introduction of the posting feedback system, which created two new roles.
- Second, the trend of inactive (Taciturn) and active roles harmonizing in regards to distribution.

The following two sections attempt to provide a possible explanation for these two effects. The first is the result of an action by the providers of the forum, whereas the second effect is rooted in an overall societal change.

Sidenote: With regards to the major temporary shift in distribution right before week 500, it appears to be an anomaly with the observed data. Silent Voters increased their shares on account of Taciturns, but both returned to the values that their long-term trend indicates. Based on the data provided and the timeline of derStandard.at the only major event, that would explain the initial shift was the introduction of their Android app in June, 2012[1]. But there is no explicit event in December of 2013, which would explain the second shift back to the original trend line.

New voting system

On June 7th, 2006 derStandard.at introduced a new voting system to their news forum. Before that users could only post comments and reply to others' comments. The new feedback system allowed users to *agree* and *disagree* with each posting (upvotes and downvotes).

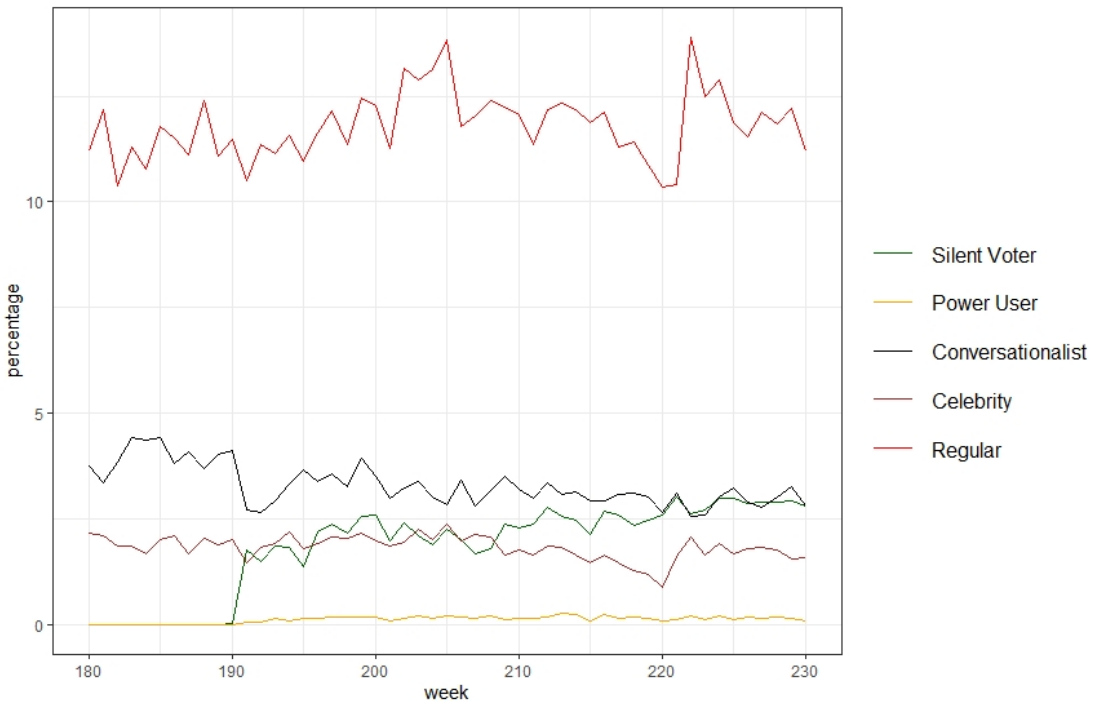


Figure 6.7: Distribution of roles, without Taciturns, after a new voting system was introduced (Jun. 2006 - Jun. 2007).

[1] https://about.derstandard.at/unternehmen/die-chronologie-des-standard-und-derstandard-at

This feature created the role of the Silent Voter, as well as the Power User. Figure 6.7 shows this birth around the 190-week mark. From then on Silent Voters have continuously gained shares of the total user base. It proved that a crucial way of engagement was missing for a lot of users, without which they would not have started to participate in the forum.

Long term shift

A long-term trend towards active roles is visible. At the beginning of the forum, Taciturns made up around 80% of the user roles. Other roles, except Regulars with around 10%, were negligible. Over time all roles gained shares at the expense of Taciturns.

Two possible explanations for the redistribution of shares from Taciturns to more active roles could be

1. Users evolve from a spectating role to a more involved one with time, and thus fall out of the Taciturn bracket and into any of the other roles.

2. Users are more open to participate in online social forums as time passes, due to it becoming normalized over time.

Based on the linear trend it is fair to assume that digitalisation might be among the main drivers of this effect. Connecting and engaging in social forums became more common with time, considering the start of the news forum happened at a time when using the internet for such tasks was not yet a mainstream activity. Thus over time, the percentage of actively engaging members increased compared to purely observing ones.

An important part falls to Silent Voters. This role was introduced after the voting system was introduced to the news forum. Before that users were only able to comment on each other's postings, but could not engage in other ways. Once the voting system was introduced this role quickly increased its shares of the overall distribution. It became the second most popular active role after Regulars. Being able to vote on postings appears to have converted an increasing number of otherwise mostly inactive users towards becoming active engaging members of the user base, highlighting the importance of providing different means of engagement for different personalities.

Another observation is that Power Users appear to be a relatively new phenomenon. As stated in Subsection 6.1.7, Power Users tend to vote a lot, as such they only started to emerge after the voting system was added to the news forum. Yet, while the number of Silent Voters began to rise quickly after this feature's introduction, it takes far longer for Power Users to gain shares. A possible explanation could be mobile phones and mobile internet connection. Power Users are defined by the volume with which they engage. More access to the news forum could therefore mean more opportunities to engage and participate.

One role that decreased in importance with time is the Conversationalist. This role is defined by longer discussions and almost no voting behavior, as stated in Subsection 6.1.6. As such, it appears that this particular style of communication plays a less important role nowadays than in the early days of the news forum. This behavior is more commonly associated with regular forums or boards and might either have gotten out of style in this news forum context or be less suitable for the fast-paced news cycle of the current era.

6.3 Implication of the COVID-19 pandemic

The corona pandemic provides a unique occasion, for a detailed analysis of a global event, which might have had an impact on online communication behavior in the recent past. As a result of the pandemic and through government-mandated lockdowns, a shift towards digitalization could be observed throughout society ranging from adopting work from home, digital payments, online shopping, and online communication, such as video meetings and team chats [FGL+21].

This fast-paced change in online culture combined with the unknown impact of the pandemic during the early days and the anxiety-inducing nature of a pandemic, raises the question if there was a measurable impact on the behavior during these times, as studies have shown increasing levels of stress and anxiety during this time according to Király et al. [KPS+20]. Special attention is given to the periods of government-mandated lockdowns, as these times were forcing many people to stay at home, providing the potential to increase their online activities.

In 2020 the Austrian government enforced three mandated lockdowns, which restricted citizens in their right to move freely aside from defined essential tasks. As a result, people had to remain in their homes for longer periods and where possible fall back to working from home instead of commuting to their office.

These lockdowns saw surges in access to online services from online shop software to remote video calls [FGL+21, AAS+21]. It therefore begs the question if these times, during which interactions were limited to online services, affected online communities and specifically with regards to this work, online behaviour in a noticeable manner.

Table 6.2 shows the starting dates of the three lockdowns in Austria for the year 2020. This work will especially focus on the 1st lockdown, as it appears a more isolated event compared to the other two. The 2nd lockdown was preceded by a less strict lockdown on November 3rd, 2020 which one day prior was preceded by a terror attack in Austria, both impactful events on the daily news blurring the lines of this analysis. The 3rd lockdown was during the Austrian Christmas vacation period, which does not represent an average period suitable for comparison.

6.3.1 Notable differences in roles

Figure 6.8 shows the role distribution between February 2020 and November 2021, including the beginning of the COVID-19 pandemic in 2020. The three major lockdowns

	1st lockdown	2nd lockdown	3rd lockdown
starting date	16.03.	17.11.	26.12.
calendar week	12	47	52

Table 6.2: Austrian government mandated lockdown dates for the year 2020.

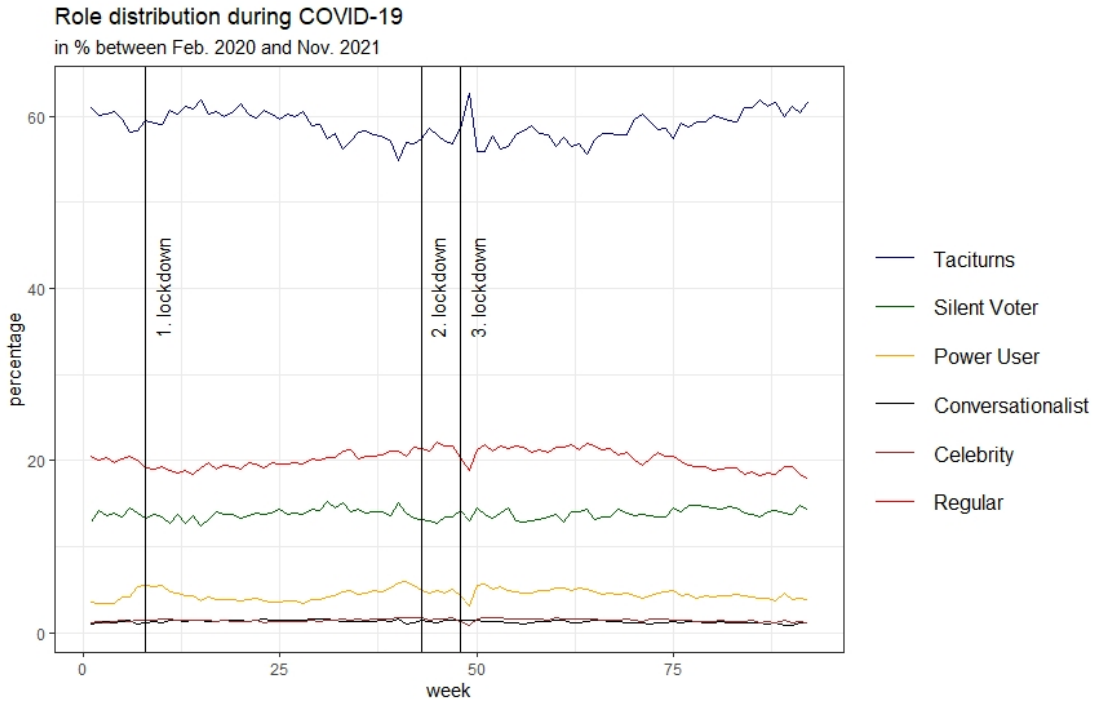


Figure 6.8: Distribution of roles during the COVID-19 pandemic (Feb. 2020 - Nov. 2021).

are marked with vertical lines.

To understand the shift in behavior it is important to remember that the lockdown does not coincide with the emergence of the coronavirus, which started to spread in the weeks and months before. March 2020 is about the time when the initial reports of observed virus infections in Austria began. Before this, news coverage about the pandemic was mostly focused on China and Italy, but it was already the main topic of the news cycle. As such, and due to the intimidating nature of the pandemic, the upward trend of the number of postings in the forum can be observed in the weeks before the first lockdown, as can be seen in Figure 6.9.

The effect on the distribution of roles can be seen in Figure 6.8, as well as in the data of Table 6.4. The latter contains the weekly changes in percentage for each role, positive values show increases compared to the previous week and negative values decrease. Leading up to the first lockdown, which begins in week 12, there is a strong increase detectable for Power Users and Celebrities, which share very similar traits.

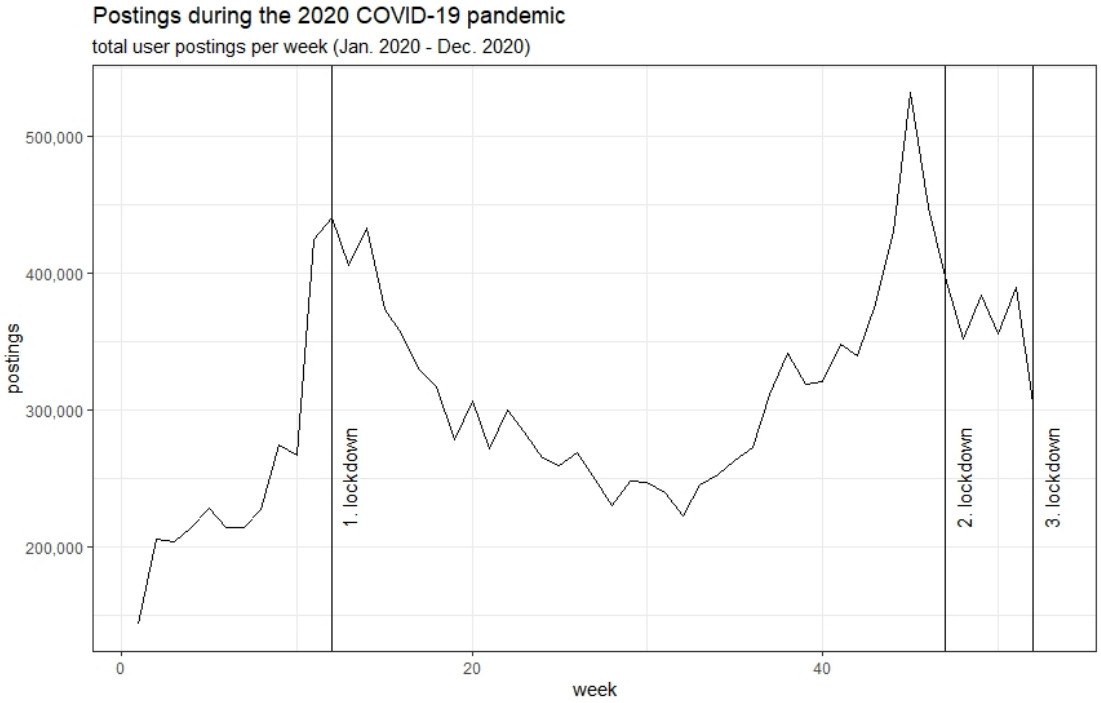


Figure 6.9: Number of user postings in 2020.

Feb. 2019 - Feb. 2020

	Taciturns	Silent voter	Power User	Conversationalist	Celebrity	Regular
mean	1.432427	2.992976	7.705423	8.334694	8.308633	2.060604
median	1.223845	2.618348	6.424245	7.660466	6.675946	1.470948

Feb. 2020 - Feb. 2021

	Taciturns	Silent voter	Power User	Conversationalist	Celebrity	Regular
mean	1.985150	2.943646	8.401767	9.200321	7.459001	1.787123
median	1.548543	2.762132	5.643923	5.690644	4.926617	1.417980

Table 6.3: Average changes per week in percent for each role.

Table 6.3 lists the average changes per week for each role. It includes mean and median values for the first year of the pandemic and the year before. Based on these values, we can determine that some roles appear more inclined to change than others. However, averages for both periods are very similar, which can be interpreted that there is no major underlying shift in communication behaviour detectable. Yet, looking at the individual weeks, there is some evidence that suggests specific events, like the announcement of a lockdown, do have a short-term impact on the discussion behavior and volume.

The two major observations regarding the roles of the news forum are increases for Power Users and Conversationalists. Especially Power Users appeared to have a drastic increase up to the first lockdown, beginning a few weeks prior, which then plateaued for the

duration of the lockdown. After that, it slowly returned to a level slightly above the average, before increasing again right up to the start of the second lockdown.

At the same time, between the first and second lockdowns, Taciturns were at an all-time low, meaning there were more engaged users than ever before. Regulars and Silent Voters both showed the same upward trend between these two events. Conversationalists also showed a noticeable increase right after the initial lockdown, despite being one of the smallest classes within the forum.

Without a doubt, the lockdowns, especially in the beginning, had a short-term effect on the posting behavior of the user base as portrayed by the shifts in role assignment. It is not possible to determine any long-term effects from this data though. A longer period after the end of the COVID-19 pandemic would have to be observed to make a statement about long-term changes.

week	Taciturns	Silent Voter	Power User	Convers.	Celebrity	Regular
6	-2.38725971	5.83210070	-3.30187728	29.0916830	6.2184824	-1.77409399
7	1.30050277	-3.61234739	0.99506164	-3.3705231	-3.5069695	1.21541577
8	-1.24004521	3.62185695	2.05499794	6.4911084	3.4323370	-1.74442607
9	-1.99758574	-1.72574483	21.23636228	2.7102459	19.0058766	1.60900247
10	-1.51112883	1.69462419	-0.96076322	5.6906439	-6.8235355	1.41798036
11	0.85418509	-5.02253622	33.68129572	-25.7821581	5.1893814	-0.75534772
12	2.18732115	-3.05430157	0.75377018	2.9974410	4.4626312	-1.86837184
13	-0.42285249	3.32903347	-2.28695048	29.8432305	-4.3108483	-2.99505249
14	0.29745339	-3.08660469	3.52880758	-10.3012775	16.4423664	1.37835658
15	3.14037917	-2.92253531 -	13.46968531	18.7749259	-4.8465843	-1.32125311
16	-2.27975549	7.68903047	-4.36938287	-0.3169651	-3.1156696	-1.09660126
17	3.09312797	-5.86122867	-5.11797184	-5.0822951	-1.5819854	0.89612808
18	-1.22258677	4.89073276	-1.55274158	10.1445096	-0.1708300	-2.34528128
19	2.44190998	-5.33771779 -	12.13916961	-2.0670286	-4.7085070	2.76734512
20	-3.79385726	5.42193964	8.71531518	-11.6560349	1.2171077	2.54290470
21	-0.47942583	6.19632521	-8.60865825	3.0972772	-4.9266169	-3.29510740
22	0.22587347	-3.11167463	5.17203688	4.1674738	5.8329699	1.28070064
23	0.75181307	0.30147990	-4.40453705	2.2908447	-4.5940673	-0.98822745
24	1.84931075	-2.76213183	-3.30347995	1.8072281	-2.8418969	-0.51174037
25	-2.08953065	1.69437008	5.64392311	-9.5667075	-8.4068279	3.05387976
26	-0.94686765	0.55927356	3.89708276	8.1365112	10.2613093	-0.30224530
27	1.00184334	1.03794075	-8.39399804	11.0688414	-14.0356268	-1.67368764
28	0.12888367	-0.17366243	-7.07326535	-4.1844139	1.5822649	1.26908056
29	-1.52722077	1.57798758	1.20769262	3.2850292	8.7768020	0.69808182
30	0.37789002	-0.79564076	8.12974642	-4.7585435	0.2054935	-0.73638468
31	-0.79667266	-0.09921417	-1.81508980	-1.4293911	3.1609069	1.90899949
32	1.69189706	-1.83105692	-8.57258314	-0.2801418	0.3577712	-0.23226647
33	-3.70623575	2.84482011	10.23548878	4.9498985	8.3749990	2.41506737
34	1.00023731	-2.02209257	3.01500591	8.9212180	-6.1806050	-0.72251851
35	-3.37766282	6.27952321	6.73783489	-10.3320910	7.8174108	-0.07867868
36	0.82225398	-3.16136973	1.27716697	-0.2974372	3.4960736	1.27722321
37	-2.56193492	1.47862567	12.74229605	-11.0768373	2.4334939	1.85220346
38	1.70496255	-3.72856894	4.87288224	-8.6490333	-4.7154956	0.53687841
39	1.25705826	1.05137320	-7.94272792	9.6573796	2.3814521	-2.36888932
40	1.99877315	-3.57866546	1.67593268	3.0142505	0.4448939	-0.78781452
41	-1.42745856	2.48646989	5.32622615	-4.3969068	5.4189917	-0.57455683
42	-0.73359838	-1.03879677	0.25130588	4.1206267	-5.1901089	2.33109616
43	-1.69146123	-1.28698815	7.06253215	-5.2657765	7.3945826	2.82948012
44	-4.17731453	4.97488311	12.46648287	17.4998662	-0.3594998	-0.67310074
45	3.49897563	-2.57615769	2.87083206	-39.6649982	8.8663223	-1.66793986
46	0.28739471	-4.22372192	-6.78474327	34.8022704	-5.0942811	3.70527412
47	1.74465549	-2.22683398	-10.40250914	15.1785738	0.7605219	0.17326649
48	1.54854278	1.67447908	-6.91788359	-9.0862139	-11.6753506	-1.44369295
49	-0.92950330	-2.44368305	4.19125567	-8.7570651	5.9024646	3.21463297
50	-2.23730182	4.02351081	-2.93770921	23.6395022	-1.0341704	-0.16395657
51	-0.22322020	-1.29352076	7.72313098	-3.0475867	9.5634582	-0.18946933
52	3.48577875	4.56906014	-16.51206145	3.5308276	-22.6655786	-5.38302138
1	7.39923783	-2.23624975	-27.96695901	10.7775395	-30.7415650	-6.06356577
2	-11.27713841	2.81545319	78.93042983	-13.3058530	78.0057949	9.09910105
3	-0.49620322	-2.37317504	3.01151316	-9.4018288	7.2005059	2.56310606
4	3.76900575	-2.36404422	-9.83004713	-3.6255946	-5.4275888	-1.55772252
5	-3.10078432	4.20741072	5.99084698	2.1193083	0.9410502	0.46066350
6	0.71906716	1.81065839	-7.23160363	-4.1060628	-3.2251011	-0.90665096

Table 6.4: Weekly changes in percent for each role (Feb. 2020 - Nov. 2021).

CHAPTER 7

Evaluation

7.1 Validation of the model

The classification model created by this work is quantitatively validated through statistical tests to confirm the predictive power of the model. The metrics used for evaluation are described in detail in Section 4.5.

7.2 Statistical tests

Based on the process outlined in Subsection 4.5.1 the accuracy will be tested by measuring `precision`, `recall`, and `F1 score` for the overall model, as well as each role individually.

The classification model assigns roles consisting of nominal categorical values, as there is no inherent order among them, so it is necessary to apply a multinomial logistic regression to validate the model [EH12, Böh92]. The results of the trained model and the confusion matrix of the predictions are listed in Table 7.1.

7.2.1 Multinomial logistic regression results

To conduct the validation a dataset was manually labelled with the roles defined in Section 6.1. A total of 2570 user entries consisting of weekly user data was assigned the six active user roles. This dataset served as the source of truth for the statistical validation, further described in Subsection 5.3.1.

First, the model was trained using a 75% subsample of the manually classified dataset, resulting in 1929 users being assigned roles by the model. The method used in the R function `train` was `multinom`, a penalized multinomial regression model[Kuh08, KWW+20]. The trained model was then used to predict the roles of the remaining 25%

of users of the manually classified dataset. Finally, the predicted roles were compared with the manually assigned ones to test the capabilities of the model. The individual class results are listed in Table 7.2.

At first, the accuracy of the model appears very high with 92.2%, with a 95% confidence interval between 89.85% and 94.16%. This means of all the predictions the model made, 92.2% were assigned the correct role. In very imbalanced datasets this metric can be deceiving, because as long as it predicts the majority classes correctly, accuracy will be high even if other roles were assigned incorrectly[MDDD20, GZC09, Pow11].

Therefore it is important to look beyond accuracy alone. The `no information rate` shows the largest class of the dataset, and accuracy should always be above this value. The `P-Value [Acc > NIR]` tests whether accuracy performs better than the `no information rate`. As the p-value for this test is far below 0.01 this test confirms the hypothesis.

Finally, it is necessary to determine the significance of accuracy by comparing it to the performance of random chance. The `kappa` measures how many labels would be assigned correctly were only the distribution of the resulting classes known, disregarding the model characteristics themselves. Its score is 0.8855 and therefore very high on a scale of -1 to 1, meaning the predicted values show high levels of agreement with the observed values of the manual classification [LK77].

7.2.2 Individual role results

The statistical analysis results for each role are listed in Table 7.2. As discussed, the main evaluation metric for evaluating the role assignment will be the `F1 score`, the harmonic mean of precision and recall.

Table 7.2 extracts the core metrics `precision`, `recall` and `F1 score` of all roles for easier comparison. The roles **Taciturn**, **Silent Voter** and **Power User** all exhibit F1 scores well beyond 0.90 and, except for Power User, beyond 0.95 meaning the model classifies a high percentage of the users correctly. It performs slightly worse for the second half of roles: **Conversationalist**, **Celebrity**, and **Regular**. However, both Conversationalists and Celebrities still show scores beyond 0.8 and Celebrity almost reached 0.9 in F1 score.

Conversationalists have high precision, meaning that of all the roles classified as Conversationalists, the majority are labeled as this role. But it lacks somewhat in recall, so not all instances of Conversationalists can be retrieved reliably. Celebrities are more balanced in precision and recall, diverging only slightly from one another.

The worst performance is shown by Regulars. With an F1 score of only about 0.6 it hints at a potentially poorly delimited role. An explanation for this weaker performance compared to the other roles could lie like the role itself. Regulars are defined by their all-round behavior, which paired with their low interaction volumes, does not emphasize specific behaviors, making it border to almost all roles in regards to how they behave.

Penalized Multinomial Regression
1929 samples 7 predictor 6 classes: '1', '2', '3', '4', '5', '6'
Pre-processing: scaled (7) Resampling: Bootstrapped (25 reps) Summary of sample sizes: 1929, 1929, 1929, 1929, 1929, 1929, ... Resampling results across tuning parameters:
decay Accuracy Kappa 0e+00 0.9182442 0.8805769 1e-04 0.9180235 0.8802155 1e-01 0.8526989 0.7783416
Accuracy was used to select the optimal model using the largest value. The final value used for the model was decay = 0.

Overall Statistics
Accuracy : 0.922 95% CI : (0.8985, 0.9416) No Information Rate : 0.4743 P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.8855
Mcnemar's Test P-Value : NA

Table 7.1: Multinomial logistic regression results for 1929 manually labeled users.

	Taciturn	Silent V.	Power U.	Conver.	Celebrity	Regular
Sensitivity	0.9934	0.9742	0.9492	0.7353	0.8667	0.5254
Specificity	0.9288	0.9918	0.9948	0.9967	0.9951	0.9759
Pos Pred Value	0.9264	0.9742	0.9492	0.9259	0.8966	0.6889
Neg Pred Value	0.9937	0.9918	0.9948	0.9853	0.9935	0.953
Precision	0.9264	0.9742	0.9492	0.9259	0.8966	0.6889
Recall	0.9934	0.9742	0.9492	0.7353	0.8667	0.5254
F1	0.9587	0.9742	0.9492	0.8197	0.8814	0.5962
Prevalence	0.4743	0.2418	0.092	0.053	0.0468	0.092
Detection Rate	0.4711	0.2356	0.0874	0.039	0.0406	0.0484
Detection Prevalence	0.5086	0.2418	0.092	0.0421	0.0452	0.0702
Balanced Accuracy	0.9611	0.983	0.972	0.866	0.9309	0.7507

Table 7.2: Multinomial logistic regression results by class.

At the same time, it is difficult to distinguish them from Taciturns, as they are both low-volume roles and the borders of this are blurry. They do however appear in most cluster solutions and are found throughout the literature, which still warrants their inclusion in the model, despite their weaker performance in statistical tests.

Taking a closer look at the `prevalence` of the individual roles, or the frequency of the appearance of each role in the labeled dataset, we can determine that the more common roles are better classified than the rarer ones. However, occurrences alone do not explain the differences in scores, as roles with similar frequencies can also differ greatly in comparison with each other, for example, Power Users and Regulars. As such, the dataset used to evaluate the model appears to be selected well for the task.

7.2.3 Micro and macro F1 scores

Attempting to evaluate the model as a whole gives us two options: the micro & macro average F1 score[SL09]. Both of these scores summarize the results of the individual clusters into one metric. The macro average is calculated as the mean of the F1 scores, as such it treats all classes equally. It is a simple way to evaluate the model quickly but does not consider the underlying distribution of the roles by including the actual amount of observations.

The micro average regards the actual observations and thus is impacted by the underlying distribution. Its result emphasizes the performance of larger roles compared to less frequent ones. However, in very imbalanced datasets, micro averages can deceive and overrate the performance of the model, as they hide the poor performance of smaller roles, as long as the majority of roles perform well enough[CJK04].

The classes of this classification model are not equally distributed, which is why the macro average F1 score seems a better-suited summary compared to the micro average. According to Yang et al. [YL99] it is still more informative to consider both measures,

Micro Average F1	0.9220
Macro Average F1	0.8632

Table 7.3: Micro and macro F1 scores for the classification model.

despite potential flaws within one score regarding the selected subset. Table 7.3 displays both values for completion.

7.3 Evaluation summary

According to the conducted tests, the classification model explains the role assignments of the users in sufficient quality. It does perform better for larger roles (eg. Taciturns) and such with strong behaviors (eg. Silent Voter) compared to more general roles (eg. Regular). Given the scores of overall model performance, it explains the provided datasets in reasonable quality and significantly outperforms random probability-based role assignments.

CHAPTER 8

Conclusion

8.1 Summary

In this work, a classification model for users of an online news forum, with commenting and voting capabilities, was proposed. After an initial exploratory and manual analysis of salient behavioural patterns, variables were defined to build a robust model using statistical analysis.

The hierarchical clustering resulted, based on suggestions of literature, in several roles showing similar behaviours to the ones observed during the exploratory phase. By comparing the results of both analyses a merged model resulted in the definition of six roles of user behaviour, with an additional role for inactive users.

This model was then applied to the long-term forum data provided by an Austrian newspaper. Based on this application of the model, we discovered that role distribution follows observations of previous research when it comes to the distribution of active and inactive users. We could also confirm several roles found by existing user behaviour research of different contexts in Subsection 6.1.1.

Despite being able to validate some of the observed distributions, the long-term analysis in Subsection 6.2.3 provides a clear picture of a change in these behaviour-based distributions, with more users engaging and engagement frequencies increasing throughout the observed twenty-year period.

Contrary to initial assumptions about potential negative influences and sentiments affecting user behaviour negatively, observed interactions show a trend towards general positivity, especially regarding the provided voting data. And while changes in behaviour within users do not appear to have a high frequency, showcased by increased jumps between roles detailed in Subsection 6.2.1, the gradual overall shift in behaviour of the user base as a whole was visualized by Figure 6.6 in Subsection 6.2.3.

8.1.1 Research questions

RQ1: To what degree, based on standard performance measures, can users of an online news forum be classified based on known and latent characteristics, as well as interaction data?

The model developed in this thesis shows that a general-purpose model for the use case of an online news forum is possible, even with a small set of variables. To preserve the universal application of the model, it was intentionally kept small with 7 classes, one of them inferred. While a more fine-grained division of some roles is possible, it would reduce the general applicability or require additional variables and dimensions to keep its expressiveness.

As shown in Subsection 7.2.1, the model explains observed behaviours in the provided dataset with sufficient quality, suggesting that it is possible to classify news forum users with the selected set of parameters. In the statistical tests conducted, the model achieves a macro F1 score of 0.8632 and a micro F1 score of 0.9220. Except *Regulars*, all roles tend to be well explained by the model. *Regulars* are adversely affected by the circumstance that their behaviour can show signs of most roles, as it could serve as the fallback role during low interaction periods of users with other inherent roles. Their limited interactions make it hard to discern itself enough, from the more active roles.

RQ2: Furthermore, does the classification of users change over time?

User behaviour appears to have shifted over time. People active in news forums tend to interact more now, than they did over the last 20 years, based on the data we examined. While the behaviours that characterize the roles of the classification model remained similar, the distribution of the assigned roles changed significantly over the years. The only major change to the model itself occurred when the ability to vote on other users' postings was introduced, spawning a new roles labelled *Silent Voter* and *Power User*.

The long-term development of role distribution suggests a long-continued trend towards more activity of forum users. Inactive roles such as the *Lurker* have shown a continuous decline over the observed period and the second most inactive role, the *Regular*, has remained steady for the past several years. In contrast, active roles independent of their behavioural characteristics have gained shares over the years, indicating an ongoing trend towards more participation. One explanation could be found in the continuous digitalisation and the observation that participating in online social communities is already a normalized behaviour for Internet users.

Additionally, impactful short-term events, such as the COVID-19 pandemic, have shown that these events can have an immediate impact on role distribution, suggesting a temporary shift in behaviour for some users. But, these effects are time-sensitive and do not appear to have a long-lasting effect, instead, distributions seem to return to their original long-term trend relatively fast after the event ends.

8.2 Contributions

This thesis contributes a generally applicable classification model for users of an online news forum with interaction means of comments and votes, by only using a small subset of available variables to assign users one of seven roles. These variables, described in detail in Subsection 5.1.2, should in general be available to operators of comparable forum setups with commenting and voting capabilities.

As such, it should enable news forum operators to apply this model in retrospective to their data, to gain insights into distributions and development of roles for their user base. It allows them to gain a quick impression of the composition of their current users. In addition, they can use the model to gain insight into the long-term development from their beginnings to the present, allowing them to discover ongoing trends in distribution changes but also impactful single events that had lasting effects on user composition. On attempts to encourage some types of behaviours, it could serve as a measurement of success, by observing the changes of the distributions.

Research can use the model as a starting point for further investigation into specific behaviours shown by individual roles, or enhance the model with further dimensions such as textual information or user relationships. Analysing online user behaviour and user interaction is a contemporary and relevant problem in data and computer science.

8.3 Limitations

This work only regarded the quantitative information provided by the data. It only looked at the frequencies with which users post, vote etc. without regarding the context of their postings. It is therefore impossible to say with certainty what votes express. For instance, downvotes could range from slight disagreement to disdain. Including contextual information could provide another layer of information not available in this work.

Content analysis like sentiment analysis could add more information to postings, especially in the context of replies. Without looking at the words it is impossible to determine whether users are discussing something, harassing each other or if they are having an argument.

Another layer of information could be provided by looking at the relationships of users through network analysis. Patterns between regularly interacting users could give more insight into the relations of individual forum participants and could potentially open up new roles. Taking these personal relationships between users into account could provide deeper insight into the behaviours for specific roles.

Analysing and defining the behaviours is a very open-ended task, as the granularity with which the analysis is conducted will directly influence the number of roles and detectable patterns. Therefore it is necessary to define a level, which suits the scope of the analysis. Many of the defined roles could, most likely, by isolating them, be broken into further

sub-categories. For this work, it was important to strike a balance, between having informative results and roles, while providing a very generalized top view of the total user base, instead of diving too deep into specific behaviours.

Similar to the possibility of taking the contents of postings into account, it is possible to divide the data into news categories and analyse these individually, to detect patterns and role distributions related to certain topics.

8.4 Future work

Several analyses have been left for future work due to self-imposed and time restrictions, such as including content-based or topics-based analysis. As it is not possible to test all of these ideas, without diluting the scope of this thesis or breaking the self-selected boundaries of what data limitations were to be used, future work needs to address these problems.

As the data provided for this thesis is quite extensive, there are more dimensions one could use for further analysis. Building on the model this thesis proposed, the following two ideas could be investigated further:

- **Topic-based distributions and topical role changes of users.** It might prove worthwhile to analyse distributions of various topics and sub-topics within the forum, similar to [CHD10]. Additionally, this concept could be expanded to analyse the roles users take on different topics. As we have seen in Subsection 6.2.1 users do, if limited, switch roles and while this work primarily focused on time as the underlying variable, it might be reasonable to assume topic and subject-matter knowledge could influence these roles.
- **User lifecycle.** Given the overall trend in Subsection 6.2.3 and the observations of users having often more than one inherent role of Subsection 6.2.1, a deeper analysis of a potential social participants lifecycle might provide insights into how users are shaped and their behaviours adapt in such social community settings, under the assumption that users start as Lurkers (see Subsection 6.1.2) and then progress into more engaged roles with time (see Section 6.1).

Using the classification model and roles of Section 6.1 could be the stepping stone to further analysis of the dataset with different dimensions. Due to the extensive dataset dating back far into the past, it is well suited for long-term analysis, which could also be done for deeper investigations of singular roles and changes within them over time.

Finally, an argument can be made to include experts in social studies within the analysis to provide the necessary interpretation of the observed behaviours and to assess the underlying reasons for potential behavioural shifts over time or during certain periods of significance.

List of Figures

List of Tables

Glossary

downvote A negative evaluation of a user's posting. 28

posting A user-generated comment below a news article. Postings can be nested, all nested postings are referred to as a reply. 27, 79

reply A posting posted below another user's posting (or nested posting). Replies allow for nested conversations between users. 79

UGC User generated content, describes all types of text written and posted by users on a platform.. 5

upvote A positive evaluation of a user's posting. 28

Bibliography

[AA15] Tarique Anwar and Muhammad Abulaish. Ranking radically influential web forum users. *IEEE Transactions on Information Forensics and Security*, 10(6):1289–1298, 2015.

[AAS+21] Ziyad R Alashhab, Mohammed Anbar, Manmeet Mahinderjit Singh, Yu-Beng Leau, Zaher Ali Al-Sai, and Sami Abu Alhayja'a. Impact of coronavirus pandemic crisis on technologies and cloud computing applications. *Journal of Electronic Science and Technology*, 19(1):100059, 2021.

[AMS19] Alessia Antelmi, Delfina Malandrino, and Vittorio Scarano. Characterizing the behavioral evolution of twitter users and the truth behind the 90-9-1 rule. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1035–1038, New York, NY, USA, 2019. Association for Computing Machinery.

[BA78] Roger K Blashfield and Mark S Aldenderfer. The literature on cluster analysis. *Multivariate behavioral research*, 13(3):271–295, 1978.

[Bak01] Paul Baker. Moral panic and alternative identity construction in usenet. *Journal of Computer-Mediated Communication*, 7(1):JCMC711, 2001.

[BH11] Petter Bae Brandtzaeg and Jan Heim. A typology of social networking sites users. *International Journal of Web Based Communities*, 7(1):28–51, 2011.

[BHK11] Petter Bae Brandtzæg, Jan Heim, and Amela Karahasanović. Understanding the new digital divide—a typology of internet users in europe. *International journal of human-computer studies*, 69(3):123–138, 2011.

[Bin12] Amy Binns. Don't feed the trolls! managing troublemakers in magazines' online communities. *Journalism Practice*, 6(4):547–562, 2012.

[Böh92] Dankmar Böhning. Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 44(1):197–200, 1992.

[Bra10] Petter Bae Brandtzæg. Towards a unified media-user typology (mut): A meta-analysis and review of the research literature on media-user typologies. *Computers in Human Behavior*, 26(5):940–956, 2010.

[BRCA09] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, pages 49–62. ACM, 2009.

[BTP14] Erin E Buckels, Paul D Trapnell, and Delroy L Paulhus. Trolls just want to have fun. *Personality and individual Differences*, 67:97–102, 2014.

[CBDNML17] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1217–1230. ACM, 2017.

[CC03] Jyh-Shen Chiou and Cathy Cheng. Should a company have message boards on its web sites? *Journal of Interactive Marketing*, 17(3):50–61, 2003.

[CDNML14] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[CDNML15] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anti-social behavior in online discussion communities. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[CHD10] Jeffrey Chan, Conor Hayes, and Elizabeth M Daly. Decomposing discussion forums and boards using user roles. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[CJK04] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

[Die85] Andreas Diekmann. Volunteer's dilemma. *Journal of conflict resolution*, 29(4):605–610, 1985.

[DN11] Nicholas Diakopoulos and Mor Naaman. Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 133–142. ACM, 2011.

[DSR14] Lalindra De Silva and Ellen Riloff. User type classification of tweets with implications for event recognition. In *Proceedings of the Joint Workshop*

on Social Dynamics and Personal Attributes in Social Media, pages 98–108, 2014.

[EH12] Abdalla M El-Habil. An application on multinomial logistic regression model. *Pakistan journal of statistics and operation research*, pages 271–291, 2012.

[EMG07] JM Ortega Egea, M Recio Menéndez, and MV Román González. Diffusion and usage patterns of internet services in the european union. *Information Research*, 12(2):12–2, 2007.

[F+03] George Forman et al. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, 3(Mar):1289–1305, 2003.

[Faw06] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[FGL+21] Anja Feldmann, Oliver Gasser, Franziska Lichtblau, Enric Pujol, Ingmar Poese, Christoph Dietzel, Daniel Wagner, Matthias Wichtlhuber, Juan Tapiador, Narseo Vallina-Rodriguez, et al. A year in lockdown: how the waves of covid-19 impact internet traffic. *Communications of the ACM*, 64(7):101–108, 2021.

[FS02] Katherine Faust and John Skvoretz. Comparing networks across space and time, size and species. *Sociological methodology*, 32(1):267–299, 2002.

[Gar18] Kiran Garimella. *Polarization on Social Media*. Doctoral thesis, School of Science, 2018.

[GCBC20] Mattia Gasparini, Robert Clarisó, Marco Brambilla, and Jordi Cabot. Participation inequality and the 90-9-1 principle in open source. In *Proceedings of the 16th International Symposium on Open Collaboration*, OpenSym '20, New York, NY, USA, 2020. Association for Computing Machinery.

[GD04] Scott A Golder and Judith Donath. Social roles in electronic communities. *Internet Research*, 5(1):19–22, 2004.

[GLZ15] Wei Gong, Ee-Peng Lim, and Feida Zhu. Characterizing silent users in social media communities. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[GM11] Scott A Golder and Michael W Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[GMS+20] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

[Gor00] PA Gore. Cluster analysis. in 'handbook of applied multivariate statistics and mathematical modeling.'(eds hea tinsley, sd brown) pp. 297–321, 2000.

[GWLS09] Eric Gleave, Howard T Welser, Thomas M Lento, and Marc A Smith. A conceptual and operational definition of 'social role' in online community. In *2009 42nd Hawaii International Conference on System Sciences*, pages 1–11. IEEE, 2009.

[GZC09] Qiong Gu, Li Zhu, and Zhihua Cai. Evaluation measures of the classification performance of imbalanced data sets. In *Computational Intelligence and Intelligent Systems: 4th International Symposium, ISICA 2009, Huangshi, China, October 23-25, 2009. Proceedings 4*, pages 461–471. Springer, 2009.

[Hag99] John Hagel. Net gain: Expanding markets through virtual communities. *Journal of interactive marketing*, 13(1):55–65, 1999.

[HCH16] Tobias Hecking, Irene-Angelica Chounta, and H Ulrich Hoppe. Investigating social and semantic user roles in mooc discussion forums. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 198–207, 2016.

[HJSSB02] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. Searching for safety online: Managing" trolling" in a feminist forum. *The information society*, 18(5):371–384, 2002.

[Hor07] John Horrigan. *A typology of information and communication technology users*. Pew internet & American life project, 2007.

[KEE93] Dacher Keltner, Phoebe C Ellsworth, and Kari Edwards. Beyond simple pessimism: effects of sadness and anger on social perception. *Journal of personality and social psychology*, 64(5):740, 1993.

[Koz99] Robert V Kozinets. E-tribalized marketing?: The strategic implications of virtual communities of consumption. *European Management Journal*, 17(3):252–264, 1999.

[KPS+20] Orsolya Király, Marc N Potenza, Dan J Stein, Daniel L King, David C Hodgins, John B Saunders, Mark D Griffiths, Biljana Gjoneska, Joël Billieux, Matthias Brand, et al. Preventing problematic internet use during the covid-19 pandemic: Consensus guidance. *Comprehensive psychiatry*, 100:152180, 2020.

[KR09] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[Kuh08] Max Kuhn. Building predictive models in r using the caret package. *Journal of statistical software*, 28:1–26, 2008.

[KWW+20] Max Kuhn, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, R Core Team, et al. Package 'caret'. *The R Journal*, 223(7), 2020.

[LB07] Charlene Li and Josh Bernoff. Social technographics. *Mapping Participation In Activities Forms The Foundation Of A Social Strategy*, 2007.

[Ler05] Jürgen Lerner. Role assignments. In *Network analysis*, pages 216–252. Springer, 2005.

[LK77] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[LS68] Michael E Lesk and Gerard Salton. Relevance assessments and retrieval system evaluation. *Information storage and retrieval*, 4(4):343–359, 1968.

[MDDD20] Sankha Subhra Mullick, Shounak Datta, Sourish Gunesh Dhekane, and Swagatam Das. Appropriateness of performance indices for imbalanced data classification: An analysis. *Pattern Recognition*, 102:107197, 2020.

[MJ09] Kent Marett and Kshiti D Joshi. The decision to share information and rumors: Examining the role of motivation in an online discussion forum. *Communications of the Association for Information Systems*, 24(1):4, 2009.

[Nie06] Jakob Nielsen. Participation inequality: Encouraging more users to contribute. *Online: https://www.nngroup.com/articles/participation-inequality (1.6. 2019)*, 2006.

[NP00] Blair Nonnecke and Jenny Preece. Lurker demographics: Counting the silent. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 73–80, 2000.

[OD08] David L Olson and Dursun Delen. *Advanced data mining techniques*. Springer Science & Business Media, 2008.

[Pow11] David Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[PRM+10] Kristen Purcell, Lee Rainie, Amy Mitchell, Tom Rosenstiel, and Kenny Olmstead. Understanding the participatory news consumer. *Pew Internet and American Life Project*, 1:19–21, 2010.

[PT11] Maria Teresa Pazienza and Alexandra Gabriela Tudorache. Interdisciplinary contributions to flame modeling. In *AI* IA 2011: Artificial Intelligence Around Man and Beyond: XIIth International Conference of the Italian Association for Artificial Intelligence, Palermo, Italy, September 15-17, 2011. Proceedings 12*, pages 213–224. Springer, 2011.

[Rog62] Everett M. Rogers. *Diffusion of innovations*. Free Press, New York, 4th edition, 1962.

[Ski65] Burrhus Frederic Skinner. *Science and human behavior*. Simon and Schuster, 1965.

[SL09] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[SV04] Chuan-Fong Shih and Alladi Venkatesh. Beyond adoption: Development and application of a use-diffusion model. *Journal of marketing*, 68(1):59–72, 2004.

[TSFW05] Tammara Combs Turner, Marc A Smith, Danyel Fisher, and Howard T Welser. Picturing usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication*, 10(4):JCMC1048, 2005.

[VGL+23] Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. Large language models enable few-shot clustering. *arXiv preprint arXiv:2307.00524*, 2023.

[VR79] C.J. Van Rijsbergen. *Information Retrieval*. Butterworths, 1979.

[VS04] Fernanda B Viégas and Marc Smith. Newsgroup crowds and authorlines: Visualizing the activity of individuals in conversational cyberspaces. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, pages 10–pp. IEEE, 2004.

[WCK+11] Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. Finding social roles in wikipedia. In *Proceedings of the 2011 IConference*, iConference '11, page 122–129, New York, NY, USA, 2011. Association for Computing Machinery.

[WGFS07] Howard T Welser, Eric Gleave, Danyel Fisher, and Marc Smith. Visualizing the signatures of social roles in online discussion groups. *Journal of social structure*, 8(2):1–32, 2007.

[WTHC03] Steve Whittaker, Loen Terveen, Will Hill, and Lynn Cherny. The dynamics of mass interaction. In *From Usenet to CoWebs*, pages 79–91. Springer, 2003.

[YL99] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49, 1999.

[ZLW+21] Zhijie Zhao, Yang Liu, Jiaying Wang, Biao Wang, and Yiqi Guo. Association rules analysis between brand post characteristics and consumer engagement on social media. *Engineering Economics*, 32(4):387–403, 2021.

[ZWS23] Yuwei Zhang, Zihan Wang, and Jingbo Shang. ClusterLLM: Large Language Models as a Guide for Text Clustering. *arXiv e-prints*, page arXiv:2305.14871, May 2023.

[ZZH21] Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. A comparative study of using pre-trained language models for toxic comment classification. In *Companion Proceedings of the Web Conference 2021*, pages 500–507, 2021.