Leopold Magurano, BA

Detecting the influence of opinion leaders in limited social networks

Effects of user-on-user interactions in online newspaper comment sections

MASTER'S THESIS

to achieve the university degree of

Master of Science

Master's degree programme:

Computational Social Systems

submitted to

Graz University of Technology

Supervisors

Univ.-Prof. M.Sc. Ph.D., Fariba Karimi

Institute of Interactive Systems and Data Science, TU-Graz

Univ.-Prof. Mag. Dr.rer.soc.oec. Markus Hadler

Institute of Sociology, Universität Graz

Graz, November 2024

Table of Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
CSV	Comma Separated Values
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DST	Der Standard
GCN	Graph Comment Network
GUN	Graph User Network
NYT	New York Times
NLP	Natural Language Processing
OL	Opinion Leader
ORS	Opinion Rank Score

Abstract

This thesis examines the role of influential users (opinion leaders) in shaping the discussions within online newspaper comment forums. The work is divided into two parts. Initially, a pipeline is developed for identifying influential users within comment forums. Subsequently, the impact of these identified users on their peers is analysed using this OL pipeline. Building on the opinion leader detection algorithm proposed by Cheng et al., this work recreates their algorithm using modern NLP techniques, as well as extending it to enable the identification of opinion leaders across an entire platform. The analysis is based on two datasets from The New York Times and Der Standard. Users of these platforms were categorized as either opinion leaders or normal users.

The results indicate that opinion leaders across both platforms are highly active participants and tend to attract other users. Articles with OL contributions see significantly more unique visitors, higher average comment counts, and increased engagement time. While no significant difference in content toxicity was detectable between the two user groups, it is evident that the behaviour of this influential group has some impact on other users.

1 Introduction

1.1 Thesis Outline

The influence of peers, both offline and online, has long been studied in sociology and related fields (Katz and Lazarsfeld, 1955, Brian et al., 2017). With the proliferation of online social media and the shift of public discourse to digital platforms, there is increasing attention being paid to the emergence and influence of opinion bubbles and user behaviour online. The majority of this attention goes to the big social media platforms like Twitter, Facebook and Youtube (Cha et al., 2010), all of which operate with algorithms aimed to attain higher retention of users' time and attention.

While it is important to study these ‘big’ platforms, as they have become an integral part of our lives, this master's thesis aims to focus on rather smaller social media platforms with less direct influence on users. Although less prominent, comment sections of news media websites also function as forums, where users can interact with each other. The goal of this thesis is to understand how influential users peer behaviour within these smaller forums, specifically by identifying and analysing the influence of these users on their peers across entire platforms.

With the shift of news delivery from print to digital, online social media platforms have also begun to play a role in news distribution. Since news media is largely funded by advertisements, print media faced a challenge in retaining readers on their own newspaper websites. To counteract this loss of users, they had to offer more than just a digital version of print papers, leading to the introduction of additional features on news websites. (Santana, 2014) One example is the online comment section under articles, which encourages user engagement by offering a platform for readers to comment and discuss. By enabling readers to contribute to public discussions, their role shifts from passive participants to active members of the discourse. Readers can now both observe and actively engage with their peers. This type of social network represents a modern form of town square discussion that has received little attention from academia.

On these platforms, contrary to the common structure of popular social-medias user interaction and popularity are not influenced by metrics such as follower counts, friend counts, or karma scores. The interaction between users on these platforms is not altered

through content delivery systems. Consequently, news media comment sections seem to offer themselves as a good test environment to examine interactions between users.

1.2 Theoretical Foundation

The aim of this chapter is conceptualizing the theoretical background of this thesis. A short overview of the historical emergence of the concept is presented followed by an overview of the modern concepts of opinion leaderships and the efforts of identification.

The concept of central figures or opinion leaders, who exert influence over others in their surroundings, has always been somewhat vague and is widely referenced across various academic disciplines since its inception (Van der Merwe & Van Heerden, 2009; Valente & Pumpuang, 2007; Dubois & Gaffney, 2014). The term "opinion leaders" was first coined by Lazarsfeld et al. (1944) in their effort to understand the dynamics of voter decision-making during the 1940 U.S. presidential election. They proposed the two-step flow of communication model, which postulates that most people form their opinions primarily through interactions with certain influential individuals, termed "opinion leaders," who themselves derive their information from mass media sources. This model directly challenged the prevailing beliefs of the time, shaped by the emergence of mass media such as newspapers, radio, and television, which assumed that the general population obtained information directly from these sources.

While groundbreaking, the concept of the two-flow stage of communication was also met with widespread criticism for its simplicity, failing to incorporate more complex communication behaviours (Soffer, 2021). Nonetheless it is generally acknowledged that certain individuals influence others through interpersonal communication. (Odefey, 2011) The importance of opinion leaders has also been questioned in recent times (Bennett & Manheim, 2006) With the decline of traditional media formats like print media and newscasts, in which consumers had to actively seek and invest effort to inform themselves, a space was initially created for intermediaries—opinion leaders—to relay and interpret news information. However, with the rise of highly personalized media, where individuals receive tailored news directly through algorithms, the intermediary role of opinion leaders is being replaced. This shift toward what Bennett and Manheim (2006) describe as the "one-step flow" of communication reflects a transformation in how individuals engage with information. Instead of relying on peer networks or opinion leaders to mediate content,

people are now directly targeted with content that fits their personal interests and preferences (Bennett and Manheim, 2006).

Nevertheless, the challenge of identifying opinion leaders through algorithmic means, both online and offline, has engaged the attention of scientists across a range of disciplines, including healthcare, politics, media, and marketing. The identification of influential figures within the public sphere is considered a crucial step for the dissemination of a particular message. (Jungnickel, 2018)

For the purposes of this master's thesis, I will follow Jungnickel's (2018) definition of opinion leader. The author distils the concept through a literature analysis, defining opinion leaders by three core elements: firstly, they influence the opinions and behaviours of others; secondly, they can influence others through both mediated and non-mediated interpersonal communication and thirdly, opinion leaders are not professional communicators, meaning they are not employed to capture public attention or to inform the public.

The modern concept of opinion leaders can also be broadly applied to the phenomenon of power users on online platforms. Power users refer to individuals who disproportionately contribute to activity on online platforms, while also often assuming leadership roles within online communities through their high levels of engagement and participation (Bright et al., 2020). These users can be found on nearly any user-generated content platform, adhering to the 1% rule. This rule describes the highly unequal distribution of content creation and sharing online Which typically, follows the 90-9-1 rule, where 90% of users on online social networks do not participate in discussions and are merely observers often referred to as "lurkers", 9% participate occasionally, and 1% account for most contributions on the platforms where they are active (Nielsen, 2014). As Bright et al. (2020) discovered, power users have the ability to steer attention toward topics that align with their own interests. Therefore, the majority of users can be seen as passive observers in the online space, while the 1% lead the discussion.

Approaches to identify opinion leaders vary widely and are mostly dependent on the population being studied. Weimann and colleagues (2007) collected six methods to identify opinion leaders, which can be grouped into two types of approaches: survey-based approaches and observational techniques. Survey-based approaches have historically been preferred due to their lower costs. These methods involve querying individuals through

interviews or questionnaires to either position themselves and their perceived influence within their networks or to nominate individuals they see as opinion leaders. This approach relies heavily on participants' self-assessments and perceptions rather than direct observation of their interactions. Observational techniques, on the other hand, gather data by tracking social interactions and communication patterns within a community. By creating a "map" of social connections, these methods aim to identify key users, particularly within closed networks. While observational techniques often require significant effort and cost when conducted offline, they can also be adapted to digital platforms, where social connections and interactions are more readily trackable.

The state of the art, for identifying opinion leaders algorithmically in last decades has evolved beyond the use of superficial user metrics or simple centrality metrics such as retweets, likes, and comments to assess influence within online networks. These approaches now employ advanced network analysis techniques (Jungnickel, 2018). In a review of recent publications on identifying opinion leaders algorithmically in online forums, M. Kang et al. (2023) analysed the applied methods and identified three overarching approaches in the current state of opinion leader detection: social network analysis (SNA), rule-based scoring, and unsupervised data mining techniques. The most common approach is social network analysis, which identifies influential users based on centrality metrics such as degree, betweenness, or closeness centrality, to locate users in influential positions within a network. Rule based scoring approaches on the other hand, rank users on a platform according to self-established scoring rules. Typically combining engagement metrics with self-crafted proxies. The third technique relies on machine learning algorithms, which include the clustering of user to identify users with behavioural outliers or employ topic modelling algorithm to segment users on a contextual level.

On a theoretical level, Jungnickel identifies four criteria for assessing opinion leaders: Contacts, which measure the quality and influence of social relationships; Activity, which considers online communication frequency and content quality; Feedback, which gauges acknowledgment through views, likes, and comments and Citation/Imitation, which assesses the spread of a message. This matches the practical approach proposed by Song et al. (2011), a novel method to identify opinion leaders in news comment sections. Their methodology surpasses conventional link analysis by also including semantic similarity as well as the sentiment orientation of comments to build a relationship network, weighing the influence of

single comments on others down the line. While they applied their strategy to dynamically identify the most influential comments in singular threads, these core concepts can be adapted to expand the model beyond the original intent to identify influential users across an entire platform.

1.3 Research Questions

Research Question 1: In constrained social network scenarios, like comment sections on news websites, is it possible to identify opinion leaders?

Using the algorithm described ins section 2.2, I will categorize the user base of a news platform dataset into two groups: normal users and opinion leaders, assigning a label to each user. This classification enables me to perform statistical tests to compare the identified opinion leaders with normal users. By analysing these distinct groups, I aim to assess the characteristics and behaviours that distinguish opinion leaders from regular users, ensuring the robustness and accuracy of the algorithm in identifying influential individuals.

Research Question 2: Are opinion leaders merely prolific posters or do they actively shape the platform's discourse?

This research question aims to explore if OL are merely identified based on their high volume of activity on the platforms, or if they also differentiate themselves though other characteristics such as increased levels of user engagement. For this purpose, I use the likes and replies as a measure of interaction between users. Additionally, I want to examine whether high level of user engagement, regardless of user type, is a result of sheer prolific activity or if it reflects the influential nature of a user's contributions.

> **Hypothesis 1:** The engagement levels of comments by opinion leaders significantly differ from those of regular users, indicating their influential role.
>
> **Hypothesis 2:** As the volume of comments posted by any user increases, the engagement levels of comments by the same poster increase as well.

Research Question 3: What impact does the influence of opinion leaders have on shaping the commenting behaviour of their peers in these specific online environments?

This research question aims to answer if the presence of opinion leaders changes the behaviour of other users. The question is grounded in the idea that users are more inclined to engage in discussions when they observe or respond to opinion leaders. Uncovering potential patterns in the commenting behaviour may lead to a better understanding of the phenomena of opinion leaders itself.

Research Question 4: Does the tone of the conversation differ when users engage in discussions with opinion leaders compared to normal users?

This research question aims to uncover the possible differences in the quality of the discussion when an opinion leader is present in the conversation. Assuming that opinion leaders generate a higher level of engagement, a possible explanation could be that content of their comments encourages users to continue participating in the conversation. This would be reflected by a low Toxicity value assigned to by the perplexity API (see Section 2.3.2)

> **Hypothesis 3:** The toxicity levels of the conversations between opinion leaders and other users are lower in comparison to discussions between regular users**.**

Chapter 2

2 Methods

2.1 Datasets

For the purpose of testing the developed algorithm for identifying opinion leaders, two suitable datasets from two fairly respected newspaper websites for their country of origin, namely *Der Standard* (DST) and *The New York Times* (NYT), have been selected. DST is an Austrian newspaper which reaches around 12% of the population on a weekly basis with an overall public opinion brand trust of 58% (Newman et.al. 2024). NYT is based in the United States of America and reaches around 13% of the population on a weekly basis with an overall public opinion brand trust of 50%. (Newman et.al. 2024) Both datasets are publicly available and include the comments as well as comment properties of all users and on all articles over a defined period of time.

It is important to note that, while both datasets used in this thesis offer a comprehensive overview of past commenting activities of users on the respective platforms, the commenting activity itself represents only a fraction of the user behaviour. Due to the snapshot nature of both datasets and the reluctance of the platforms to share extensive data about their users, information such as the first activity of a user on the webpage, the number of views of each comment, and the article reading behaviour of users is not available.

2.1.1 The New York Times Dataset

The New York Times dataset was collected by Dornel (2021) over the course of one year (01.01.2020 – 31.12.2020) through the now depreciated NYT comment application programming interface (API). The dataset was subsequently published in the form of two comma-separated values (CSV) files on Kaggle (Dornel 2021). Each file contains a table: the comments table and the articles table. The articles table contains all 16.787 articles published within the observed time frame. Each article entry is described by 11 features. The key columns in the article table include the newspaper section in which the article was posted, the headline of the article, the comment count, and a unique identifier of each article. This unique identifier links the articles table to the comments table.

The comments table contains 4.986.460 comments posted by 300.000 users and is comprised by 23 features. In the comments table the most important features for the OL detection pipeline are the user identifier, which allows tracking of the users over the entire timespan; the comment body; and the number of recommendations (likes) a comment has received.

Commenting on articles and accessing the comment section on the NYT-website is only available for subscribed users. This relatively high barrier of entry to the comment forum, set by a standard monthly subscription cost of 25$ (The New York Times, 2024), raises expectations for a more engaged and civil community compared to a "free" platform. Additionally, not every article has comments enabled. According to Etim (2013), a community manager for the New York Times, comments are enabled according to the news value of the story, the projected reader interest in the story, if a similar issue recently had comments enabled as well as whether moderation is possible in a timely fashion.

2.1.2 Der Standard Dataset

Der Standard (DST) dataset draws many parallels to the NYT dataset, as it also encompasses all comments posted on the platform over the span of a year (01.06.2015 - 31.05.2016). This dataset involves significantly fewer users compared to the NYT dataset, with 31.000 accounts totalling 1.011.773 comments under 12.087 articles. This dataset was collected and published by Schabus et al. (2017) in cooperation with DST.

In contrast to the NYT dataset, the barrier to entry for posting on DST is lower, as no paid subscription is required to post comments on articles. Additionally, all published articles have the comment section enabled, which is reflected by the high volume of comments compared to the lower user base. Although its metrics are not included in the dataset, DST additionally allows users to follow each other, which enables users to track the postings of others. This has three functions: it first tracks all users one has labelled as followed in a list, through which one can access their profiles and see all comments posted by this user in the last 30 days; secondly, it highlights the followed accounts with a blue icon; lastly, beside each username on postings, the number of co-posters ("Mitposter:innen"), meaning the number of followers a particular user has, is displayed.

The structure of the DST dataset mirrors that of the NYT, comprising also two distinct tables: one for articles and another for comments. It offers the same functionalities for tracking

comments and viewing likes. A notable difference is that the DST dataset also tracks the possibility for users to dislike comments, adding an additional layer of interaction data.

2.1.3 Algorithmic Gatekeepers

NYT as well as DST follow a “slow moderation” philosophy, where each comment is initially screened algorithmically for inappropriate content. The algorithm either approves the comment directly or flags it for manual review by human moderators. This hybrid approach allows companies to host a discussion platform while also providing the legally mandated supervision and moderation of their platform. (Der Standard, 2013; The New York Times, 2017)

NYT employs a system called “Moderator”, developed in partnership with Jigsaw (an Alphabet Inc subsidiary). “Moderator” is a machine learning algorithm trained on past NYT comments in a supervised fashion to identify toxic comments. This was achieved through human coders, which were given a set of past NYT comments. These comments were then labelled as either toxic or not toxic according to their subjective perception. (Perspective API, n.d.; The New York Times, 2017)

Similarly, DST utilizes an automatic moderation system known as “Foromat”, developed by the Austrian Research Institute for Artificial Intelligence in 2005. This system primarily conducts a check for empty posts, unknown links, foreign languages or dialects, advertisements, and offensive language as well as incorporating a “karma” system that affects the likelihood of a user's comments being automatically approved based on their posting history. (Der Standard, 2013; Austrian Research Institute for Artificial Intelligence, n.d.)

According to DST this system allows them to moderate the platforms more efficiently with around 75% of comments passing the automated checks. (Der Standard, 2013) While both newspapers emphasize transparency and accountability in their moderation processes, none of them have published a set of examples or a clear threshold of their systems.

One downside of this approach on forum moderation from the user perspective is the delayed nature of conversation. After a comment is written an uncertain amount of time passes by before the comment can be read by others, hindering the flow of communication. This problem is somewhat alleviated by the option on both platforms to get notified when the comment is published, as well as if another user replies to a posted comment.

This delay can be deduced from the NYT dataset, to display the extent of it for a better understanding of its possible effects on fluid communication. The distribution of time deltas between the time of writing and the time comments are greenlit by the automatic moderator of NYT (Figure 1) suggests that around 25% of all comments are immediately greenlit, being published within 2 seconds of being written. The next quartile may undergo further analysis, given that half of the comments are approved within about 26 minutes and 32 seconds. This might still be too fast for human oversight, especially considering that approximately 15,000 comments are posted on an average day. Consequently, the remaining comments might have to be manually checked by a human in the loop, with 75% of comments approved within about 2 hours. We can also observe that sometimes comments fall under the radar of human supervisors, as the longest time taken for a comment approval is nearly 340 days.

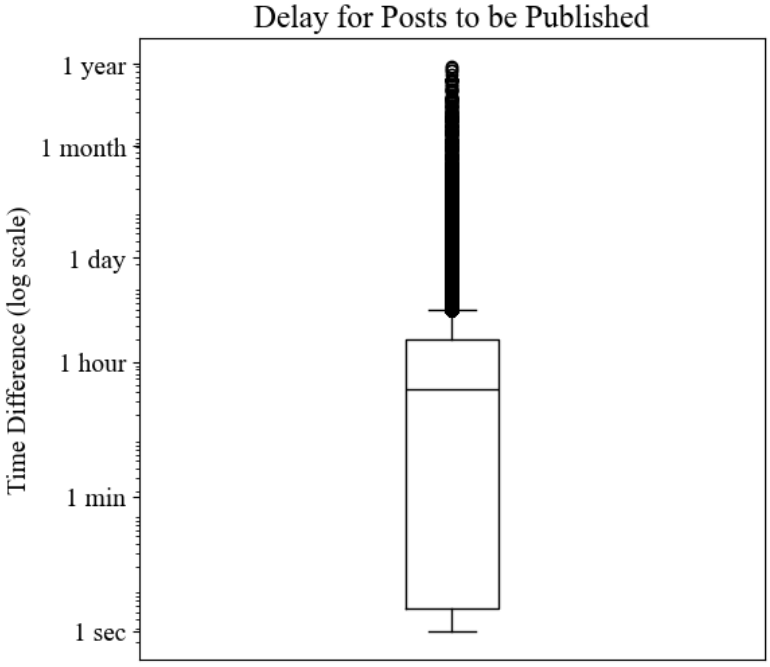

Figure 1 Moderation delay of comments on the NYT

The needed data for the analysis shown above is not included in the DST dataset, therefore the extent of their moderation delay cannot be measured.

2.1.4 Anatomy of a Comment

On NYT as well as on DST the comment section of each article is accessible via a dedicated icon, that appears on multiple locations throughout the article page. By clicking on this icon or scrolling to the end of the article, the users arrive to the comment section. Both comment sections of the websites are two-dimensional. By default, all comments are arranged in chronological order, with the most recent comment displayed first. The user is also able to reverse this or sort the comments in descending order according to the number of recommendations (likes) they have received. When sorting comments, only the metadata of the top-level comment is taken into account and not the user replies.

A comment may be submitted as a standalone contribution, referencing the article (Top-level comment) or it can be posted in response to a previous comment. In the latter case, the reply will be visible in an indented position beneath the original comment. The hierarchical structuring of comments ends here. A further reply to a comment already posted as a reply will not lead to a further indentation, leaving it to the reader to follow any discussion by paying attention to the reply handle at the beginning of the post.

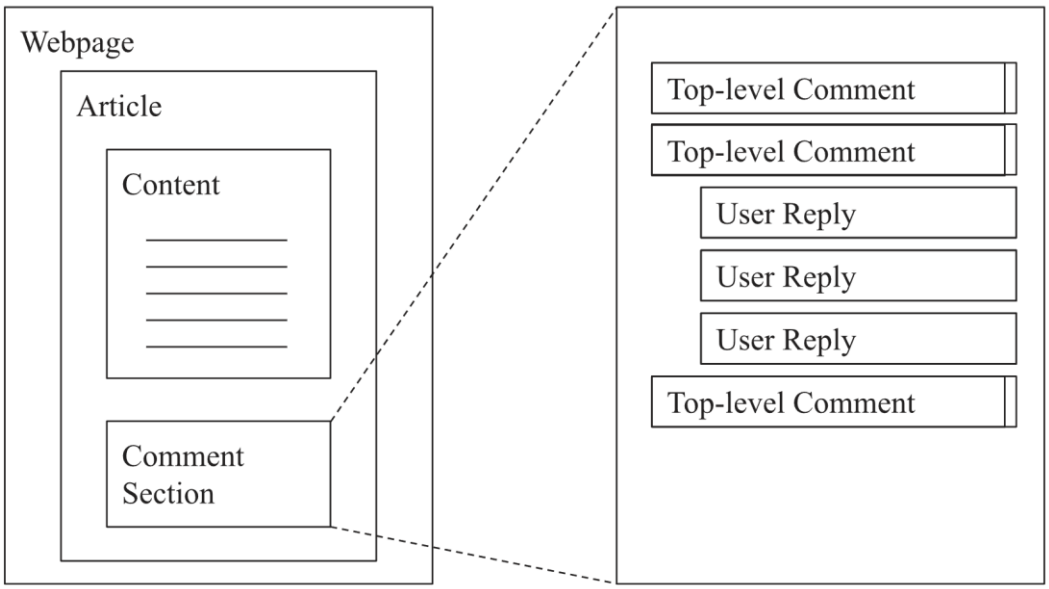

Figure 2 Anatomy of a Newspaper Comment Section

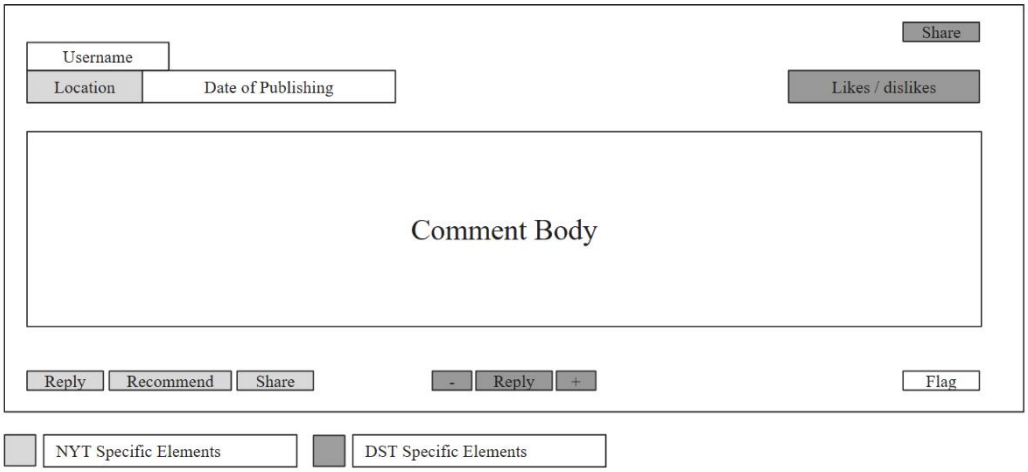

Figure 3 Anatomy of a Comment

A comment is composed of multiple elements, that largely overlap between the two platforms. In both cases, the upper left-hand corner of the display shows the username of the person who has posted the comment. The NYT also prominently displays the self-reported location of the user, while the DST places a greater emphasis on the likes and dislikes a comment has received. These are positioned on the same height as the username in the upper right-hand corner and visualized with the colors red and green, respectively. The New York Times, on the other hand, places the number of recommendations a comment has received, beside the recommendation button in the lower left corner, using the same color as the rest of the text.

The main body of the comment is centered on the page. The DST also allows users to optionally assign a title. To reply to a comment, both platforms position these options on the end of the comment frame. Both platforms also provide a button to quickly express their opinion about the comment. The NYT limits users to upvoting, whilst DST allows both downvoting and upvoting.

Another feature that both platforms have in common is a flag button in the right lower corner, which allows users to report a comment. Clicking this button opens a new dialogue-window in which a user can further explain why they deem the comment report-worthy.

2.1.5 Exploratory Quantitative Analysis of the Datasets

A comparison of the two datasets reveals a clear skew in the distribution of user activity levels on the respective platforms. A long-tailed distribution is visible in the number of comments made by users. As depicted in the Lorenz plot (Figure 4), a small proportion of users account for the majority of comments. Approximately 20% of the user population is responsible for 80% of the comments posted on both platforms. This indicates that most users are not actively participating in the discussions. The proportion of lurkers versus posters is unknown, given that the view count on comments is missing from the data. A comparison of the two platforms reveals that DST has a steeper tail, which can probably be explained by the fact that it is a free website to comment on, resulting in a higher throughput of users.

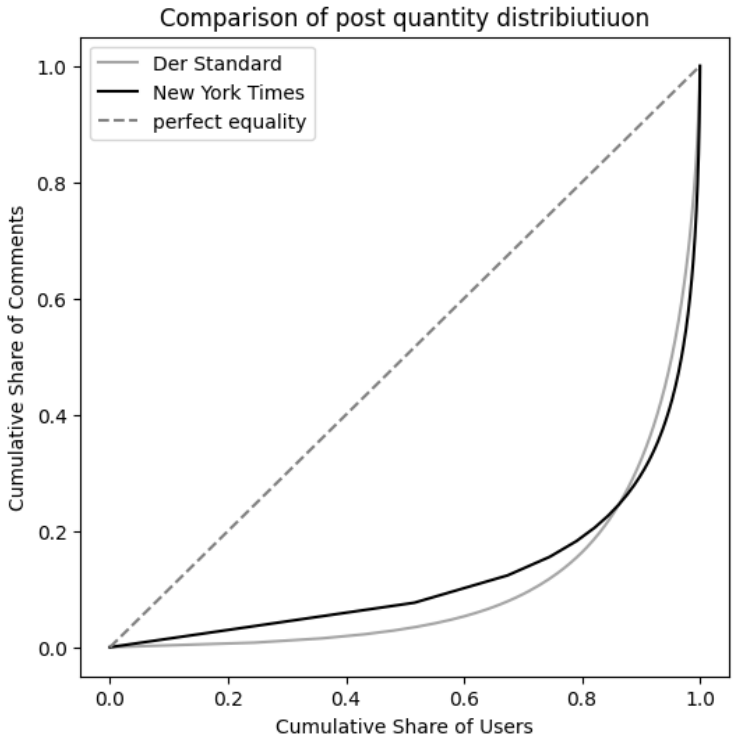

Figure 4 Comparison of Post Distributions

As mentioned before, the NYT dataset contains a self-reported optional location datapoint, in which users can self-report their location. Unsurprisingly, the majority of the users (77 %) report to be located in the United States of America (USA), followed by three major English-speaking nations, namely Canada, the United Kingdom and Australia.

User Location	Percentage Users	Percentage comments
USA	77.05	82.7
Canada	3.32	4
United Kingdom	1.81	1
Australia	1.06	1.2
Italy	0.87	0.5
France	0.85	0.6
Germany	0.77	0.6
India	0.75	0.26
Spain	0.65	0.9
Philippines	0.46	0.15

Table 1 NYT user locations

The DST dataset doesn't contain any location information and user statistics are also not published separately. However, given that the articles are published in German with the majority of the articles relating to Austria, one can assume that most of the readers would be located in Austria. Following the trend in NTY, because of the article language, it is expected that a certain percentage of users are located in Germany or Switzerland.

This assumption is supported by a comparison of the distribution of the time at which comments are posted on each newspaper. (Figure 5) The "posting times " on DST follow a more natural activity pattern, with comments becoming more frequent in the morning hours starting at around six o'clock, peaking during the lunch hour, and staying at a high until the late afternoon with a gradual drop until midnight. The lowest activity is seen in the night hours, reflecting a presumed wake and sleep pattern. This suggests that the readers of the DST are most likely to be located within the same time zone. In contrast, the NYT demonstrates a shifted peak at approximately 3 p.m. according to New York City local time (GMT -4). This difference in user activity might be explainable by the 6 different time zones in the USA, with the East Coast being up to 6 hours behind New York, inevitably leading to a warped distribution. Additionally, a less abrupt reduction in postings during nocturnal hours

is present when compared to the DST. The higher nightly activity may be attributed to the fact that the NYTs international readers are accessing the platform from around the globe.

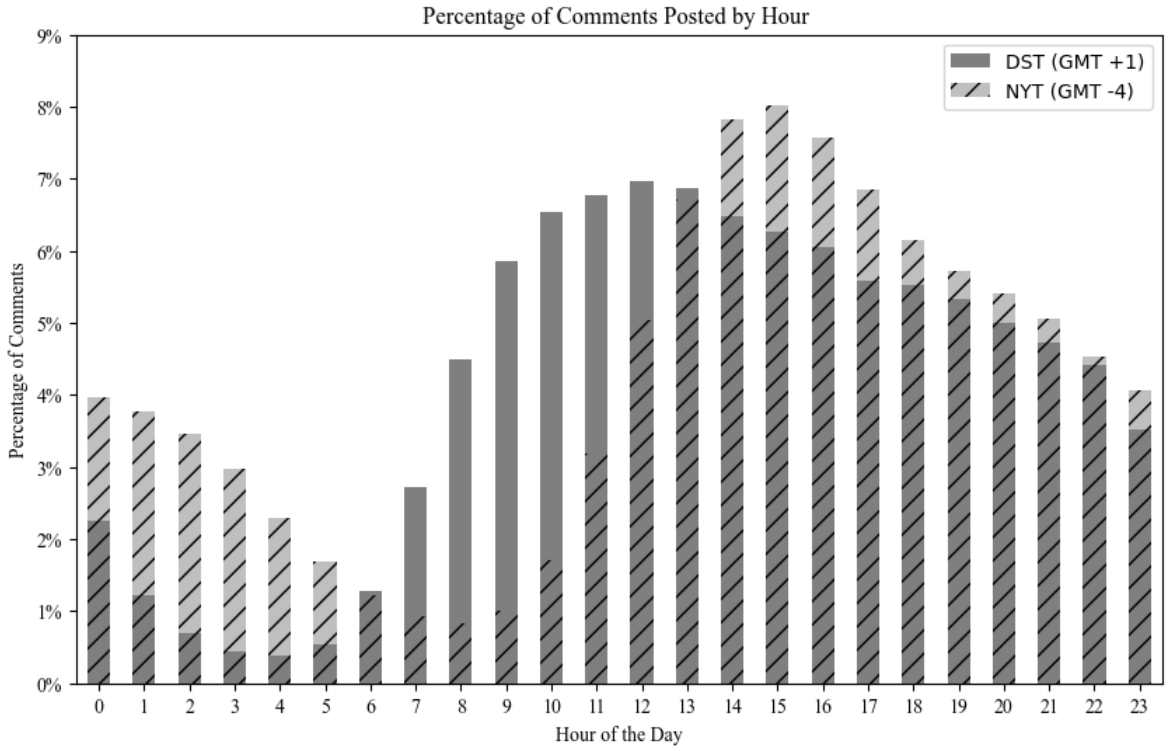

Figure 5 Hourly Post Activity

Looking at the posting behavior on a more zoomed-out timescale, we can observe a very prominent weekend slump in commenting activity. This is contrary to the assumption that on weekends, where people supposedly have more 'free' time, they spend more time online. This may indicate that commenting and discussing in online forums is predominantly an office-hour activity. This cyclical pattern is more developed in the NYT dataset, with activity on Saturday and Sunday dropping to half of the volume compared to weekdays. The DST on the other hand only experiences a 20% drop. Interestingly, zooming out further over the entire span of the year this recurrent pattern in the NYT dataset can be observed throughout the entire year, as seen in Figure 7. In contrast, the weekly patterns in the DST dataset are less prominent, suggesting that the commenting activity is mainly influenced by news events.

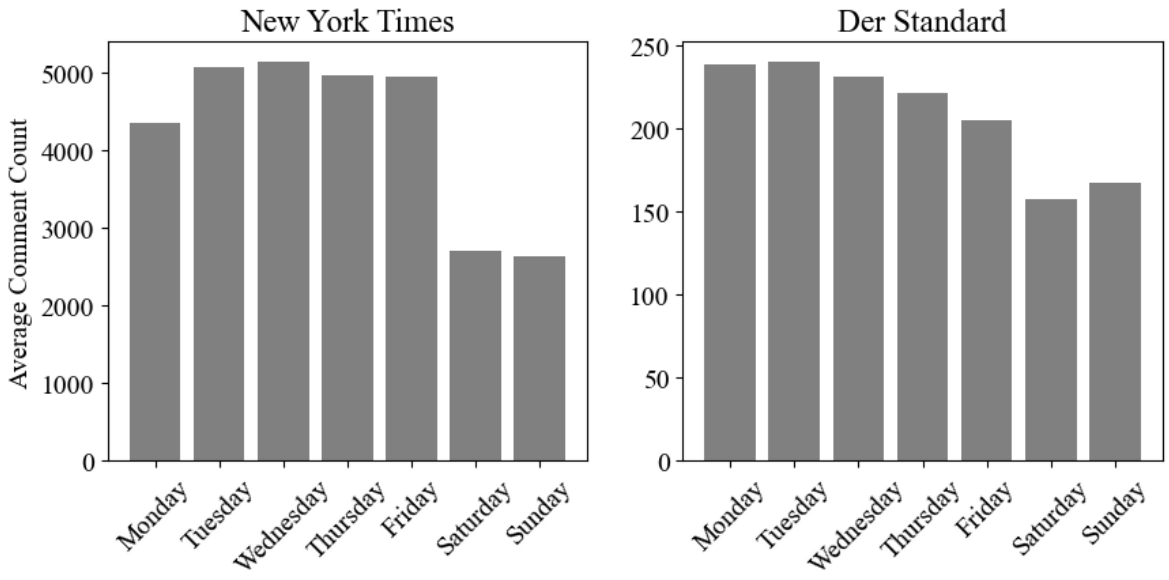

Figure 6 Weekly post activity

However, the peaks in activity for both newspapers can be explained by news events, especially of political nature. Specifically, the New York Times often exceeds 20,000 comments per day during such peaks. The most commented article of peak days makes up the difference between the average comment value and the peak comment value. For more detail on the type of articles please refer to the table in Appendix 3 which lists the headlines of the most commented articles of the peaks.

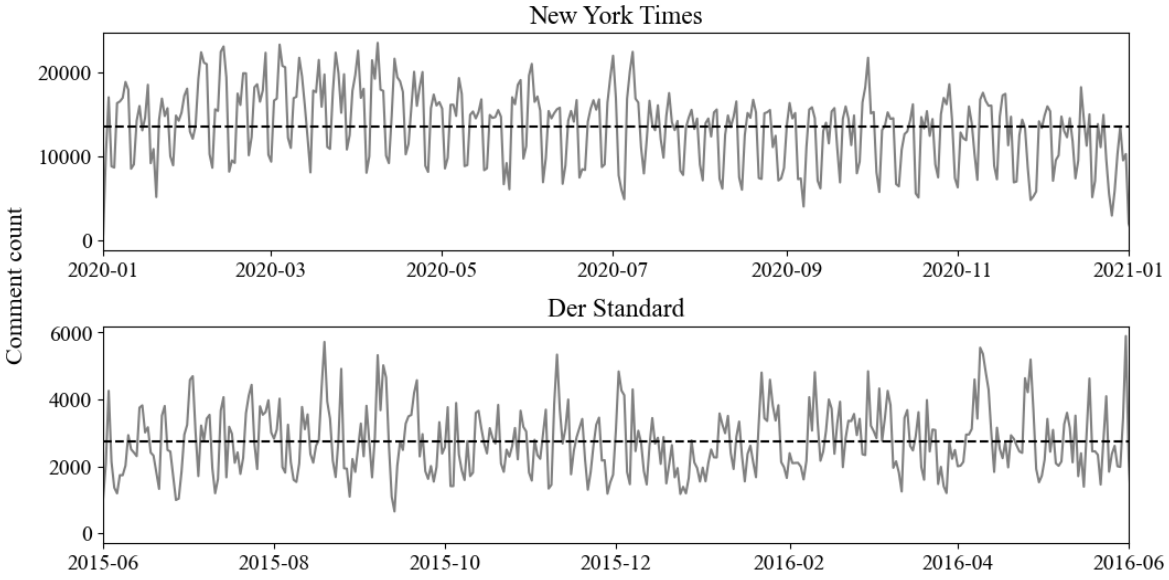

Figure 7 Total Post Frequency Distribution

2.2 Algorithmic Opinion Leader Detection

The approach for detecting users of interest follows the conceptual proposal of Song et al. (2011). This method was selected from the numerous methods published in the literature as it goes beyond simple social media metrics, like likes, reposts or network centrality to detect opinion leaders and puts an emphasis on the dimension of time and content similarity (Kang et al., 2023). The reference paper focuses on algorithmic detection of opinion leaders dynamically in a single thread or an article over time, which is not fully applicable in the context of a whole platform, but its core functionalities will be adapted to be used in a platform wide opinion leader detection.

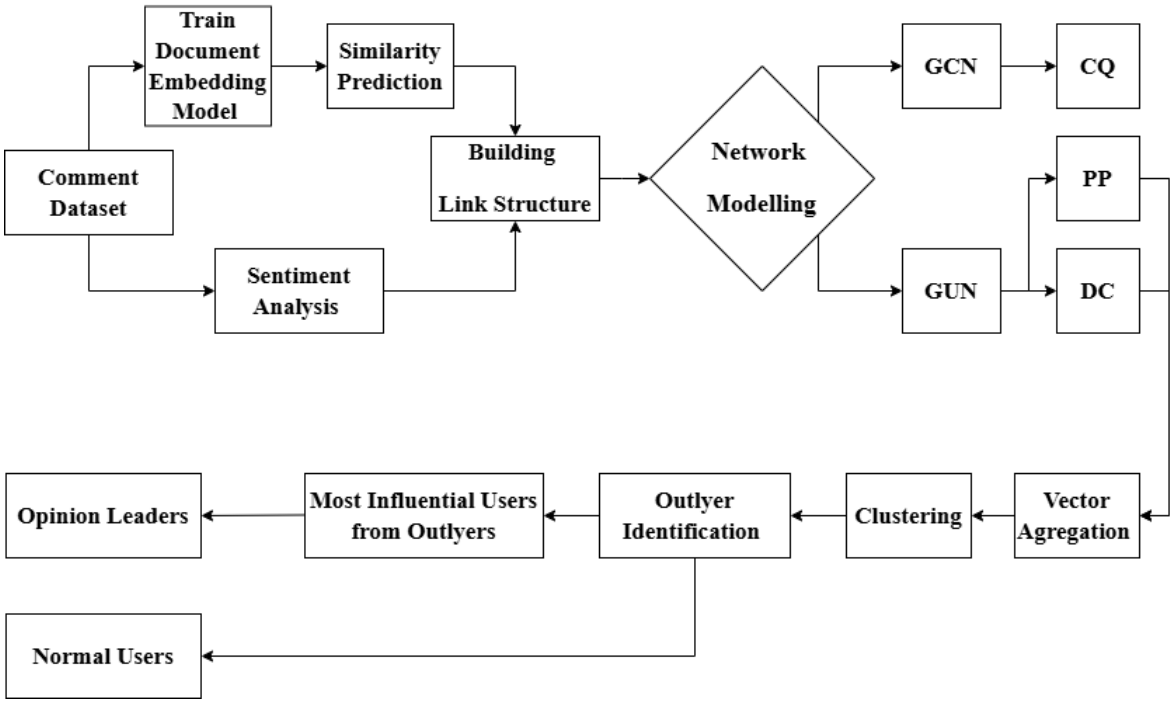

Figure 8 Pipeline Flowchart

2.2.1 Preprocessing and Standardization of the Datasets

The pre-processing of the datasets consists of standardising the structure of the input tables for the opinion leader detection algorithm. Although the two datasets are similar in structure, the column names and data types have to be reformatted to be consistent.

Column Label	Description
commentID	unique ID of comment
userID	unique ID of the user who made the comment
parentID	unique ID of the parent comment (NA if comment is not a reply)
articleID	unique ID of the article where the comment was posted
createDate	date and time when the comment was created
approveDate	date and time when the comment was visible to others
commentBody	text content of the comment

Table 2 Standardised input table columns

As illustrated in Table 2, the standardised input matrix for the OL pipeline is comprised by 10 parameters. Each entry corresponds to a comment posted on the designated platform. The comment ID and the user ID serve as unique identifiers. The user ID remains constant over time for each user, making it possible to track a user's posting history, while a single user can have multiple comment IDs attributed to him. A comment ID can only be associated with a single user. The article ID, which is present in each entry, represents the article under which the comment has been posted. This allows comments and the comment authors to be grouped and linked to the article's metadata. In addition, each comment entry includes the comment body, which contains the content of the comment, as well as the date and time the comment was created and approved by the forum. Since the DST dataset doesn't differentiate between the time of posting and the time of approval, both variables are set equal. The parent ID is the user ID of the comment that is being replied to, in case of it being a top-level comment the value is “none”, making it possible to track comment threads where users are replying to each other. This is the closest form of conversation that can take place on these platforms, as private messaging is not facilitated.

2.2.1 Relationship Identification and Weight Calculation

The first step in the pipeline is to identify the types of links between comments for further use in network construction. Here Song's (2011) definition, as stated below, is followed, which distinguishes between two types of links.

> ***Definition 1 - link type:*** *For C_i and C_j ($1 \leq i,j \leq n$), suppose C_i be published earlier than C_j. If C_j is a follower or reply of C_i, C_j is regarded as having an explicit link to C_i. If they don't have the relationship, but C_i has semantic similarity with C_j (same or different), C_j is regarded as having an implicit link to C_i.*

As stated in definition 1, explicit links in a comment section occur between comments, when a comment is a direct reply to another comment. A direct reply assumes that the comment being replied to has been read, and that the reader feels the need to add their own opinion to the previous statement.

Implicit links, on the other hand, are links between comments that do not directly refer to each other. If a comment is posted under an article without any references to other comments, but it is semantically similar to a comment previously posted under the same article, the connection between these two comments is labelled as an implicit link. This is because it can be assumed that the earlier comment might have been read by the author of the later comment, thereby influencing its creation.

> ***Definition 2 - link polarity:*** *If C_j has the same sentiment orientation with C_i, the link (explicit or implicit) is called as "positive link", otherwise as "negative link".*

In addition to the two link types (implicit and explicit), a weight component is also applied to each link, indicating whether the connection is positive or negative. A positive link signifies that both comments exhibit a similar sentiment orientation. A negative link indicates that the comments display opposing sentiment orientations. The weight itself is not indicative of an absolute opposition or agreement of user opinions as it is possible to find both positive and negative links between comments of the same users even under the same comment thread. However, the weight can be used as a proxy for the opinions of users regarding a specific article, as the discourse on news platforms is designed by policy to exclusively revolve around the content of the news article and demand civil discussions. (Der Standard, 2024; The New York Times, n.d.)

In practice, the explicit linkage required by Definition 1 is already present in the datasets, as a reply-comment references the comment to which it responds. To identify implicit links, both the comment sequence and content must be considered to calculate the similarity metric. To compare comments, I chose the Distributed Representations of Sentences and Documents (Doc2Vec) algorithm, first introduced by Le and Mikolov in 2014. Doc2Vec is an unsupervised algorithm designed to translate paragraphs or entire documents into a machine-understandable vector. It builds on the principles of the Word2Vec framework, which translates natural language into machine-readable formats. However, instead of averaging word vectors from a text body, Doc2Vec infers a new "paragraph vector" from the previously trained word vectors.

In my pipeline, I used the Doc2Vec implementation from Gensim (Řehůřek & Sojka, 2011), an open-source library for unsupervised topic modelling and natural language processing. By training the Doc2Vec model on the entire corpus of comments in the dataset, the algorithm learns word representations, also called word embeddings, that encode semantic relationships between words. With this trained model, I generated a paragraph vector for each comment. This vectorized representation allowed me to compare comments using a similarity function to calculate the distance between vectors, providing a measure of semantic similarity.

The similarity between comments is computed using cosine similarity between their vectors. Cosine similarity measures the cosine of the angle between two paragraph vectors. The angle indicates how similar the comments are in terms of semantic content. The values range between -1 and 1, where a value of -1 means the comments are opposing in meaning, and a value of 1 suggests they share semantic similarity. For this application, I chose a threshold value of 0.5, labelling comment pairs with a cosine similarity higher than 0.5 as implicitly connected. The threshold of 0.5 represents a balance between, filtering out every comment that doesn’t reflect the exact same arguments but isn’t too liberal that all comments are labelled as implicit links resulting in a median implicit link ratio of 2:2.

The task of comparing comment vectors, although conceptually simple, grows in complexity the more comments an article has. Having a quadratic time complexity O(n2), with n being the number of comments posted under an article, it is crucial to reduce calculation time by selecting only the necessary comments. Given the assumption that a commenter can only read comments published before their own, the similarity between comments only needs to be calculated for those posted prior to the comment in question.

Following the link type identification, the link polarity needs to be established. This can be achieved by using a sentiment analysis classifier. Sentiment classification is a natural language processing (NLP) application that tries to quantify the emotional orientation of a piece of text. Sentiment classification algorithms can be categorized into following two methods:

Dictionary-based approaches, explained in a simplified way, rely on a dictionary, which contains words with predefined sentiment scores. When analysing a text, the dictionary-based methods count the occurrences of words to determine the overall sentiment of the text. With some clever additions to this method, Algorithms like the Valence Aware Dictionary and Sentiment Reasoner (VADER) (Hutto, C. J., & Gilbert, E. E. 2014) can factor in negations, punctuation, acronyms and degree modifiers. However, they still fail to capture sarcasm or irony, since they work with fixed sentiment values per word.

Transformer-based approaches represent a more recent method that work with pretrained models, which were trained on a corpus of labelled texts. At the time of inference, transformer-based models attempt to predict the sentiment values according to learned patterns previously seen in the training data. Through the immense size of their training data, they are capable to predict more subtle and complex sentiment patterns. While they’re still not perfect, these models at least have a chance to capture sarcasm, irony, and other language complexities, leading to more accurate sentiment analysis results.

Despite being almost ten years old, VADER still holds up well against newer transformer-based models like BERT (Borrelli, F. M., & Challiol, C. 2023). In terms of processing speed, dictionary-based algorithms demonstrate superior performance compared to transformer-based models, even when the latter are executed in CUDA-accelerated environments. Nevertheless, I opted for a transformer-based model over a dictionary approach, in order to make the pipeline language agnostic. Implementing a dictionary-based approach would have required a sentiment dictionary for each processed language. I circumvented this by employing a multilingual language model, which also facilitates a more meaningful comparison of outcomes.

For the sentiment analysis algorithm, I chose the pretrained model 'twitter-XLM-roBERTa-base for Sentiment Analysis' from the Cardiff NLP group at Cardiff University (Barbieri et al., 2021). This model was fine-tuned with 198M tweets in over thirty languages. While

multilingual models underperform compared to monolingual models due to the linguistic diversity across multiple languages (Rust et al., 2020), the slight reduction in precision is an acceptable compromise for the simplicity of dealing with a single model, given that the objective of the sentiment classification in this use case is only to determine the polarity of the comments.

By employing the compound score proposed by VADER and establishing a separation threshold of zero, the comment orientations are determined. The calculation of the compound score is achieved by assigning each sentiment output of the classifier a weight (positive: +1, neutral: 0, and negative: -1) and computing a weighted sum of the sentiment scores. This sum is then scaled using the hyperbolic tangent function to limit the score between -1 and +1, providing a normalized polarity score. (Erwe & Wang, 2024) The recommended interpretation of the compound score is as follows: a compound score above 0.05 indicates a positive sentiment analogue, a compound score below -0.05 indicates negative sentiment, and anything between is considered neutral. If two related comments are classified with the same sentiment orientation, a positive link weight is assigned, alternatively, in the case of opposing orientations, a negative link type is applied.

In summary this step resulted in the identification of the relationship (implicit or explicit) between comments in an article as well as the weight of these links (positive or negative) with the value of the weight being the cosine similarity between the two comments. A simplified example can be seen in Figure 9.

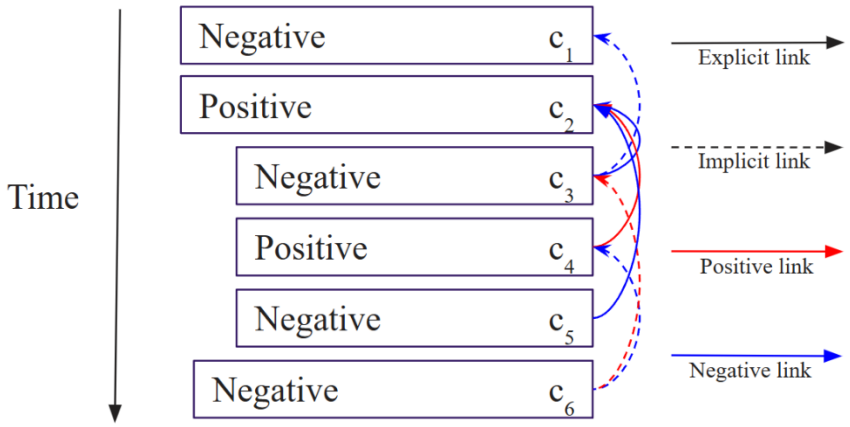

Figure 9 Comment Relationships

In this example, comment c_1 is the first comment made by a user in response to an article, with c_2 being the second comment, and so on. To simplify the visualization, the order in which comments are displayed in the figure is also the order in which they were posted. The indentation of comments signifies that they are part of a reply thread. c_1 doesn't have any outgoing links, as there are no preceding comments in this article that could influence it. However, c_1 has one incoming implicit connection from c_3, meaning that c_3 has similar content to c_1 (cosine similarity > 0.5).

c_2 has no explicit or implicit outgoing connections either, meaning it is neither a direct reply to c_1 nor sufficiently similar in content to c_1 to warrant an implicit connection. However, as c_2 has three replies, it has three explicit incoming connections from c_3, c_4, and c_5.

Comments are not limited to only one outgoing connection; for example, both c_3 and c_6 have two outgoing connections. In the case of c_3, this is because the comment is a direct reply to c_2 and also has high content similarity to c_3. c_6, on the other hand, has two implicit outgoing connections to c_3 and c_4. It is important to note that a comment can theoretically have an unlimited number of implicit links.

2.2.2 Network Modelling

With the defined link types and calculated sentiment orientations, it is now possible to build the two networks according to the definition of Song et al. (2011).

The modeling of the network is being performed with the Python package network X (Hagberg et al., 2008). The package provides the necessary data structures to represent the tabular data in graph format, as well as the specialized network analysis algorithms

Two distinct graph networks, a graph comment network (GCN) and a graph user network (GUN) are created according to definition 3 of Song et al. (2011). Each network is constructed on per-article basis, meaning that all comments under an article, and all users who comment under that article, are included in the respective networks.

Definition 3: *Let E be the set of the links of definition 1 and definition 2 (as edge), and V be the comment set, i.e. C={C1, C2, …, Cn} (as vertex), then GCN (V, E) be a comment network (as graph). Moreover, if the vertex in GCN (V, E) is defined with the user who published comments instead of comment, GCN (V, E) will express the relationship among users. We named it as GUN(V, E), i.e. user network. Different from GCN (V, E), the edges in GUN(V, E) are only explicit links. GCN(V, E) and GUN(V, E) are directed graphs (...) the edge in GCN(V, E) or GUN(V, E) has a weight wt, and wt can be calculated according to similarity of two vertexes linking the edge.*

The graph GCN is a directed network composed of vertices (comments) and edges (links between comments). Each comment occurs once in a GCN and can be connected with multiple other comments, either directly or indirectly. This implies that a single comment could potentially have n–1 edges. The edge between two vertices indicates the influence one comment might have had in the creation of the comment. The edge itself has a weight component, that is composed of the semantic similarity between the two comments.

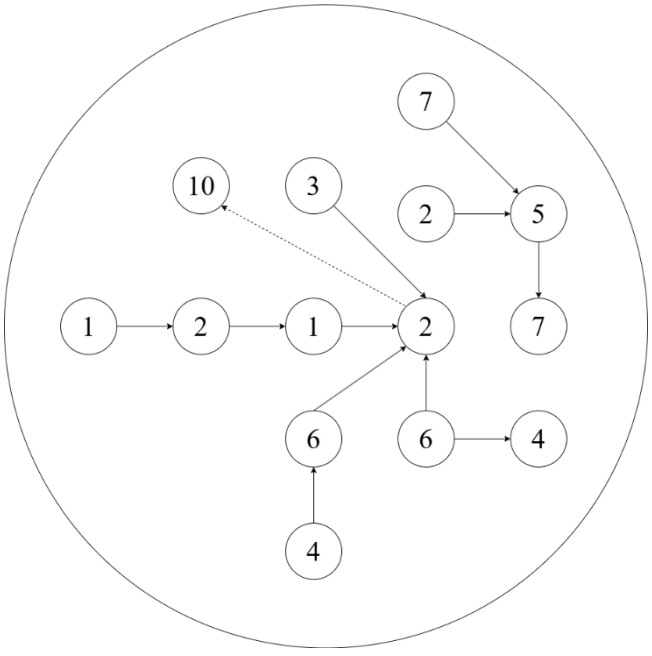

Figure 10 Graph Comment Network Example

Figure 10 is an example representation of a GCN. Each node in the graph represents a comment made under the same article. The node labels correspond to the user ID, and the directional edges represent the influence a comment exerts on another. The GCN does not

differentiate between implicit or explicit connections. Depicted in this example we can see ten users, some of whom have posted multiple comments. A comment without any outgoing connections must be a top-level comment, that is, a comment that does not reply to a comment, nor is it similar to a preceding comment.

The objective of GCN generation is to calculate the Opinion Rank Score (ORS) for each comment within an article. The ORS concept, introduced by Song et al. (2011), quantifies the potential influence of comments by incorporating both browsing patterns and the tendency of users to engage with more recent comments. This recency bias—a well-known phenomenon in social media engagement—suggests that user interaction with posts typically peaks within 24 hours of publication (Vassio et al., 2022). Although there is no literature regarding this topic in the context of comment sections of online news platforms, evidence of this behaviour is also found in the two datasets at hand, as shown in Figure 11. On average, between 80% and 90% of comments are posted within the first ten hours after an article is published.

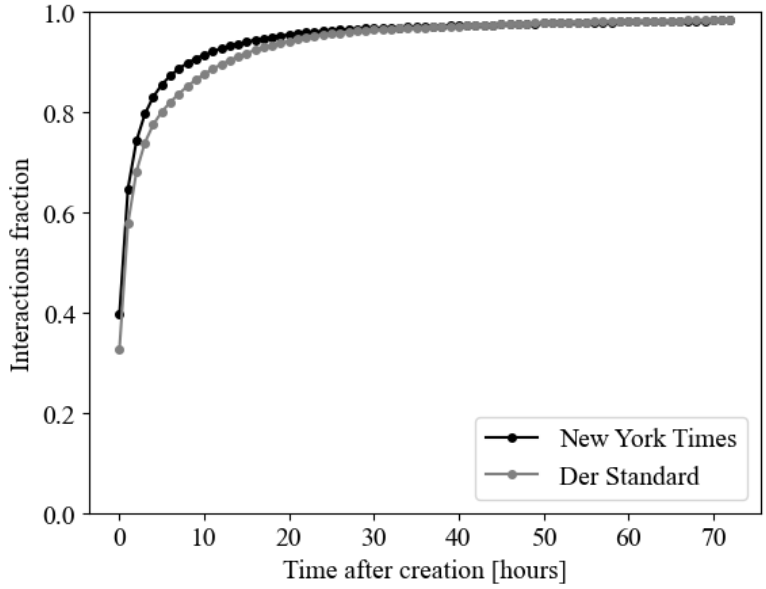

Figure 11 Average Evolution of a Post in terms of Interactions

The ORS is derived from the sum of all PageRank values assigned to each comment from the same user. Originally introduced by Page (1999) and used by Google to quantify the importance of webpages, the PageRank algorithm assigns a ranking to each page based on the number and characteristics of links in a directed network. The PageRank algorithm in this implementation estimates the relative importance of nodes in the network by simulating how a user randomly surfing might browse the comment section. The algorithm models this

behavior by moving between comments as it follows the edges (links) between them. By randomly moving around the network and keeping track of how often the "random user" passes a comment node, the PageRank algorithm assigns each comment in the network a probability score, reflecting the likelihood of that comment being influential. The random walker is not moving truly randomly through the network, as it is also influenced by the edge weights (f), the "random user" is more likely to follow edges with a higher weight.

Here, f represents the probability of a comment being influenced by the linked comment and is calculated individually for each edge, based on its time difference between the two comments being published. If a comment influencing another is farther in the past from the time of posting, its probability of influence is lower. At the time of constructing the GCN for the PageRank calculation the weight f is calculated according to the formula below.

$$f_{(t_1,t_2,D)} = D^{\frac{t_2 - t_1 \vee \times K}{3600 \times 24}}$$

t_1 ..... Time of posting influenced comment
t_2 ..... Time of posting influencing comment
D ..... Dampening coefficient (0.85)
K ..... Control parameter (2)

Since the network of commenting threads under an article is not fully connected and the user browsing behavior doesn't always follow an entire comment thread as users might lose interest, the random walker has a probability of 1-D of jumping to another randomly selected comment in the network, with D being the damping parameter set to 0.85 - the default value of PageRank.

The algorithm runs iteratively updating the probability scores of each comment until one of two criteria is met. Either the scores converge, meaning that the probability change is less than a predefined threshold or the maximum number of iterations is reached. In this implementation the threshold is set to 1e-06 and the max iteration is set at 2000.

The PageRank implementation of the NetworkX package is used to perform the personalized calculations. Finally, once the PageRank algorithm has converged, all the PageRank values of the comments belonging to a user are summed and stored as the ORS.

The GUN on the other hand consists of the set of users that commented under a specific article. Each user is represented once as a vertex, the connections between the users are represented as arrows. Each user that had an explicit connection or interaction with another user is captured and visualized in the graph, meaning all the connections are explicit links. It is a much sparser graphic compared to the GCN, as the user-comment-ratio is 1:n. The GUN is utilized to calculate the Degree Centrality (DC) and Proximity Prestige (PP) of the user in the network as described in the succeeding chapter.

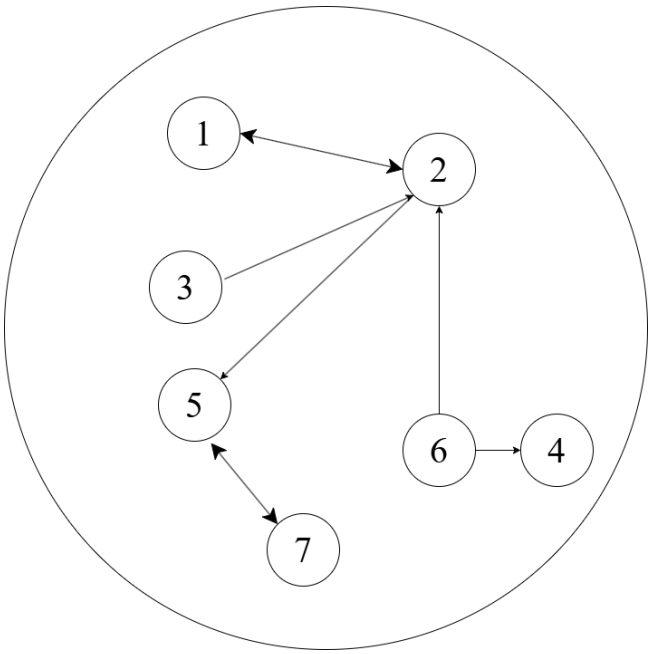

Figure 12 Graph User Network Example

2.2.3 Network Parameters

2.2.3.1 Degree Centrality

The degree centrality (DC) of a user measures the interconnectedness of this user with other users under a specific article. Given that the GUN is directional, the DC is solely concerned with the out-degree connections of a user. The DC-values lie between 0 and 1. A score of 0 indicates that the user did not reply to any other user's comments. A higher score indicates that the user has a bigger number of outgoing connections to other users. A score of 1 would mean, that a user has replied at least once to all the other users under that article, which makes this score virtually impossible except in article networks with very few users.

$$DC(i) = \begin{cases} \frac{OD_i}{n-1} & if\ OD_i > 0 \\ 0 & if\ OD_i = 0 \end{cases}$$

n …….. total users in the network
OD_i…… out degree of user i

2.2.3.2 Proximity Prestige

The proximity prestige (PP) is a measure of a user's prominence, that considers both direct connections and connections over multiple nodes to all users that are reachable ($|I_i|$). (Liu, 2011) In the case of indirect connections, the shortest path to other users ($d(j,i)$) is assessed. The PP-values lie similarly to DC-values between 0 and 1. A PP-value of 0 indicates that a user has no connections to any other users. Conversely, increasing values towards 1 signify that the user is connected directly or indirectly through other users to a higher number of total users in the network. The value is also higher, when the path to reach other users is shorter.

$$P(i) = \frac{|I_i|}{n-1} \Bigg/ \left(\sum_{j \in I_i} d\,(j,i) / |I_i| \right)$$

n …….. total users in the network
$|I_i|$……. users that can reach i
$d(j,i)$… shortest path length from j to i

2.2.3.3 Comment Quality

Comment quality represents the overall influence of a user's comments in a given article, relative to other comments. CQ is made up of four components: The sum of all PageRank scores belonging to the comments a user has posted under the article, divided by the number of comments published by the user, times the average length of all comments made by the user, divided by the maximum length of all comments made by that user.

$$CQ(i) = \frac{\sum_{j \in UC_i} \text{Score}(j)}{|UC_i|} \times \frac{CL(i)}{CML}$$

$\text{Score}(j)$ Page Rank score

UC_i ……Number of comments posted

$CL(i)$ ….Mean word count of comments posted

CML …..Average length of comments made by User j

2.2.4 Vector Aggregation

As the end-goal of the pipeline is to identify user types on a platform-wide level, each of the three user metrics need to be aggregated to represent a single user's behaviour on the entire platform. The aggregation is performed as follows: for each article a user has commented on at least once, the previously described parameters are determined. The calculated values are assigned to each user. These values are averaged per parameter, resulting in three parameters (DC, PP, CQ) per user. Using the mean values instead of the sum seemed to be the most reasonable aggregation method to perform on the data, as summation would have led to higher scores for heavy posters. The result of the aggregation is a single vector per user (user vector) consisting of the three components: DC, PP, CQ.

2.2.5 Outlier Detection

The final stage of the process is to identify any outliers within the user pool. This is achieved by clustering the users based on the user vector with the assistance of an unsupervised labelling algorithm. While Song et al. (2011) propose to use the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester, 1996) to cluster the data, applying DBSCAN, while reasonable on a small set of users, turned out to be a challenging task on a large dataset with up to 400.000 users, as the entire matrix needs to fit in memory. Because of the algorithms memory complexity of $o(n^2)$, using this algorithm would have exceeded reasonable calculation costs. Therefore, I opted to use a more recent parallel DBSCAN implementation by Wang et al. (2020), which not only improves the calculation times drastically, but also lowers the memory complexity. This allowed me to stay within my hardware limitations to compute the clustering. This DBSCAN implementation parallelizes the workflow of the clustering by decomposing the 3D-space into smaller cells, enabling local clustering within each partition before merging the results globally.

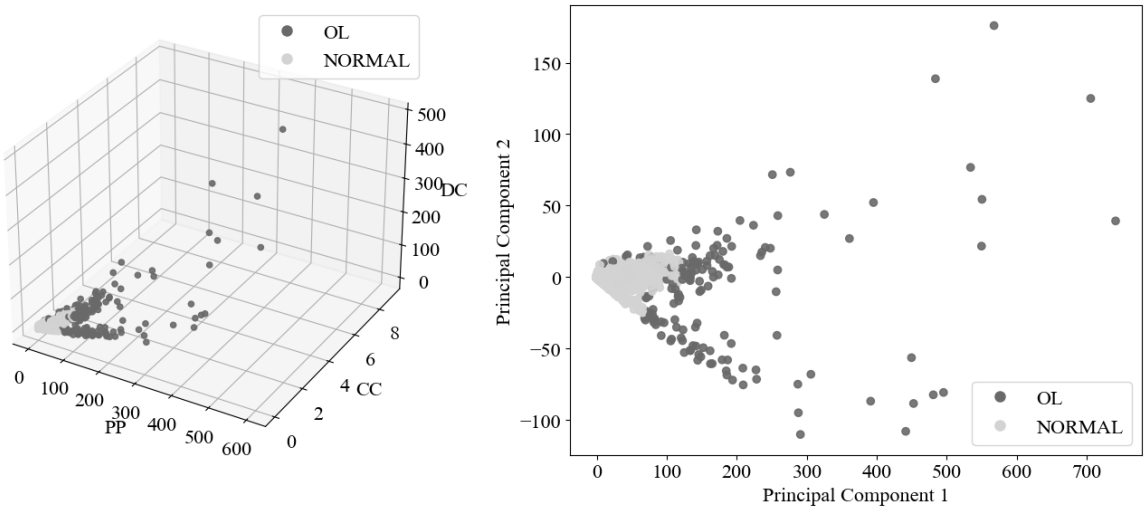

Figure 13 Spatial Plotting of NYT Users

Through plotting the users in the 3D-space, with the 3-axis being the DC-, PP-, CQ-values, the outlier detection performed by the DBSCAN can be visualized. As seen in Figure 13, the majority of users form a single, large, homogeneous cluster. This cluster found in each dataset can be conceptualized as a collection of "normal" users who do not exhibit any significant outliers in any of the three dimensions. The remaining outlying clusters contain anywhere between a single user up to a low double-digit number of users. All users that belong to any of these outlying clusters form the basis of the user ranking to identify potential opinion leaders. This final step of the pipeline ranks the outlying users according to the score value (S). Here, the same method proposed by Song et al. (2011) is employed, which is taking a weighted mean of the individual DC-, PP-, CQ-values of a user. The weights used are predetermined by Song et al. through their testing and have not been changed.

$$S(i) = ((10 * PP(i) + 0.3 * CQ(i) + 10 * DC(i)) / 3$$

The weights are set to favour the proximity prestige as well as the degree centrality in the score value, given the sparse nature of the comment networks. Calculated with the above formula, all outlying users now have a single score value attributed to them (user score), which are used to rank them. Figure 14 shows the distribution of score values for each outlying user. In the outlier ranking we can clearly see a jump in the score values at the

threshold line of the mean plus one standard deviation. This statistical line has also been chosen as threshold to classify the outliers as highly influential users or opinion leaders.

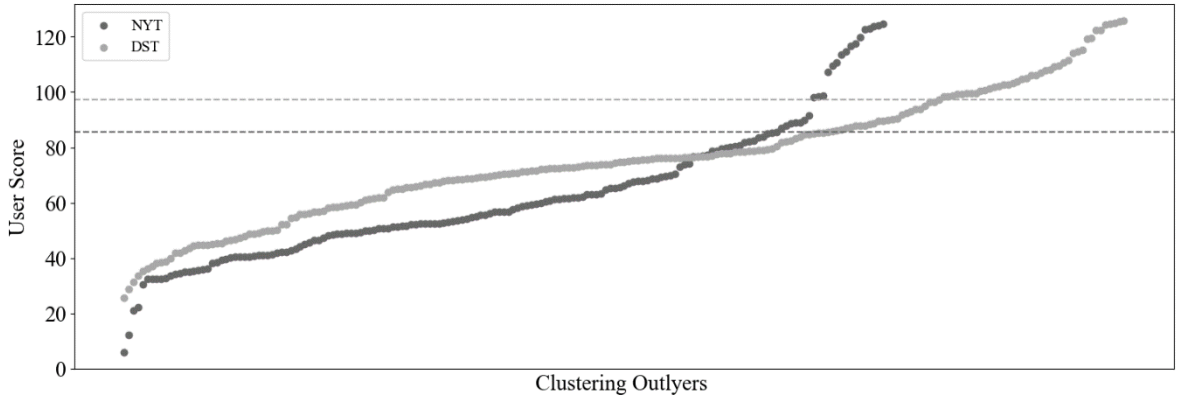

Figure 14 User Score Comparison

2.3 Parameter Analysis

A critical challenge in the development of the opinion leader identification pipeline poses the evaluation stage. The absence of a definitive ground truth prevents direct comparison of the algorithm's output with a benchmark, making traditional statistical evaluation methods inapplicable. To work around this limitation, two alternative methods have been used to evaluate the performance of the algorithm.

To test the robustness of the algorithm, the NYT dataset was split into two equal parts using the threshold date "2020-06-01". Firstly, the entire dataset was processed through the pipeline, identifying 18 opinion leaders (OL). Then, two halves of the dataset were evaluated separately. In the first half of the dataset, the algorithm identified 12 OL. These 12 OL were also among the 18 OL identified in the full dataset. In the second half of the dataset, the algorithm identified 13 OL, with 12 OL overlapping with the OL from the full dataset and 9 OL overlapping with the OL from the first half of the dataset. These two half-runs combined successfully identified 15 of the 18 OL from the full dataset, resulting in a recall of 83%.

Additionally, a qualitative feasibility check for the results was conducted by manually reviewing a small (n< 100) sample of the post histories of the 18 identified opinion leaders in the NYT dataset. The qualitative test primarily served to rule out potential mistakes, such as system messages or posts from NYT staff accounts being mistakenly classified as opinion leaders.

2.3.1 Engagement

User engagement with comments extends beyond what the dataset captured. As it is missing crucial measures like the view count of comments or abandonment figures. Abandonment describing in this context if a user leaves after viewing a comment. However, engagement can also be operationalised through two proxy variables, specifically upvotes and the number of replies, both of which are included in the dataset. The effort associated with upvoting a comment is lower, as a user only needs to click once to create a like, while typing out a reply takes more effort especially in a moderated forum where comments should contribute value to the conversation, increasing the effort it might takes. This assumption is backed by observed ratio between the sum of unique comments under an article and the sum of recommendations or likes. The slope fitted to the data suggest a ratio of approximately 1: 28. Demonstrating that recommendations are more frequently given than comments.

As in this case I am attempting to capture the comment engagement I have chosen to use the number of likes as a proxy for engagement, as this measure is not used in the opinion leader detection pipeline. In contrast, replies are a necessary element in constructing and identifying opinion leaders. Leaving out the comment count in the operationalization of Engagement helps to avoid getting a tautological result, as replies are already a component of engagement.

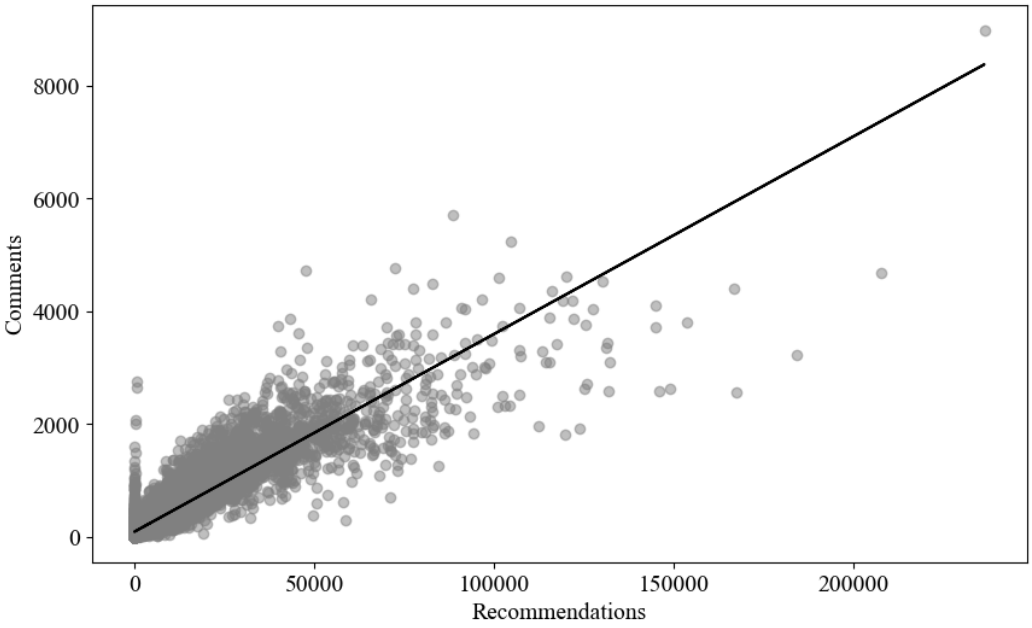

Figure 15 Relationship between Recommendations and Comments per Article

2.3.2 Toxicity

The evaluation of comment toxicity was conducted using the Perspective API, a machine learning model developed by Jigsaw a subsidiary of Alphabet. The following section shortly presents the API and its scoring system according to the developers themselves. It should be noted that although I have been granted access to the API the results and the performance of the API might not be comparable to the “Moderator” system employed in the NYT comment scoring.

In the operation of the Perspective API, a request with the text of a single comment is sent to the Google Api services, and a subsequent response containing a prediction about the potential influence of the comment on ongoing conversation is returned, by examining the comment ”across a spectrum of emotional concepts, termed attributes”. (Perspective API, n.d.)

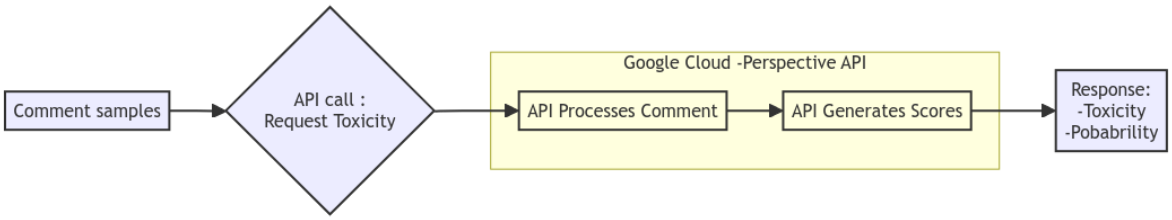

Figure 16 Perspective API-Call

According to jigsaw, the primary attribute, toxicity, is defined as "a rude, disrespectful, or unreasonable comment that is likely to compel one to exit a discussion." (Perspective API, n.d.) Within the response from the Perspective API, a score is provided which quantifies the probability that a reader would perceive the submitted comment as manifesting a particular attribute. This scoring reflects the evaluated emotional impact of the comment relative to the requested attributes. (Lees et.al 2022)

The default quota limit for the non-commercial Perspective Api is 60 requests per minute. Yet thanks to a last-minute increase in the quota to 36000 requests per minute, it was possible to process both datasets in their entirety. This rendered previous representative sampling efforts obsolete but enabled me to obtain a overview of the distribution of toxicity within the forums.

After having collected all the comment's toxicity scores an oddly similar toxicity distribution between both datasets became evident. The toxicity distributions are overlapping and follow a similar distribution raising concerns about the validity of the results. This unexplainable similarity between the two distinct datasets is hard to explain without further insights into the inner workings of the perspective API.

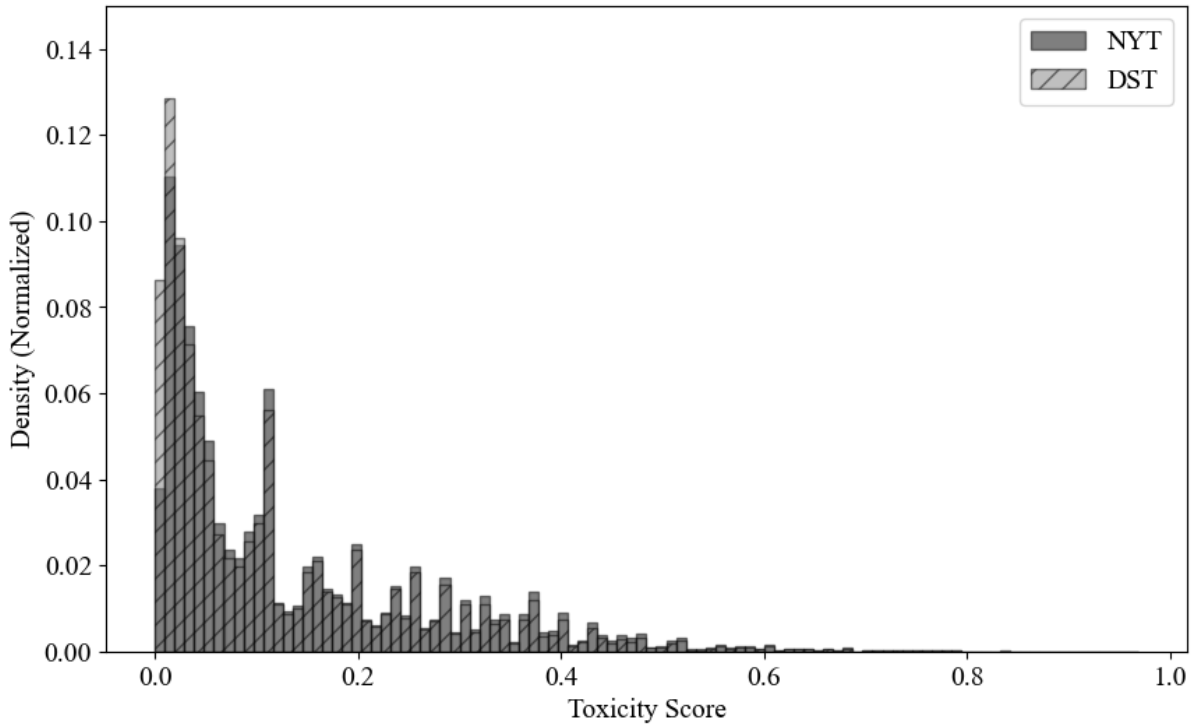

Figure 17 Distribution of comment toxicity scores

Observing the two toxicity distributions without comparison the results seem reasonable. In both datasets we can observe a strong left leaning distribution of scores, with 75% of comments lying in the lower fifth of the value range. This can most likely be attributed to the pro-active comment moderation being performed on both platforms.

Interestingly, the data at hand contradicts the findings of Nogara, et. al. (2023), which have found in their analysis that the Perspective API misreads German as more toxic compared to English. Comparing the two datasets the comments in the DST have a lower overall mean of 0.119 compared to the NYT mean of 0.134. However, due to the black box nature of the model, any changes between their time of testing and my analysis could account for a change in the model's behaviour.

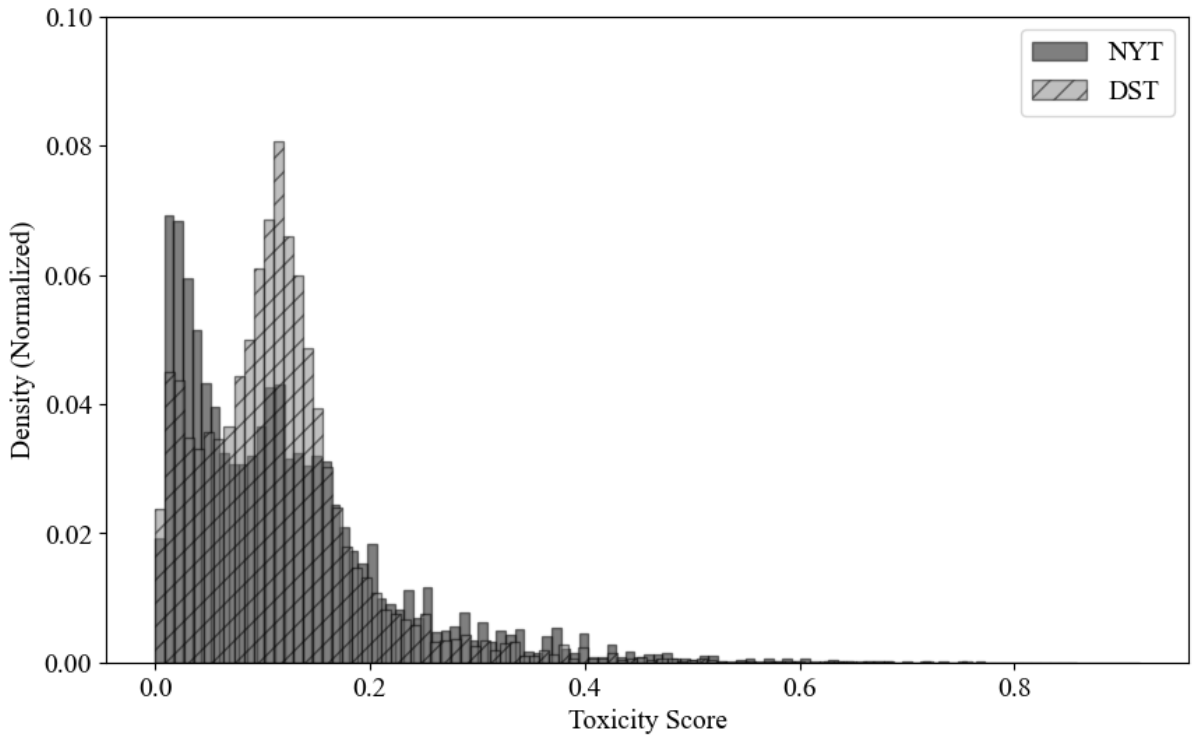

Figure 18 Mean user toxicity values

Looking at the mean toxicity scores per user, the distributions for both datasets show that the majority of the users are still within the lower fifth of the toxicity value range. Although now the users in the DST dataset have a shifted peak towards slightly higher toxicity values. The shift in the distribution indicates that a large proportion of users have a wide range of toxicity scores attributed to their comments. The majority of users in the datasets exhibit low mean toxicity scores. with only a small number of users having a high (> 0.4) average toxicity score. Notably, the majority of high-scoring users have a relatively low number of comments associated with their profiles.

As is the case with the majority of machine learning models, particularly those of a closed-source nature, the algorithms themselves are a black box, returning scores with little to no information regarding the reasoning behind them. Therefore, it cannot be stressed enough that all results should be interpreted with caution.

3 Results

RQ1

Research Question 1: In constrained social network scenarios, like comment sections on news websites, is it possible to identify opinion leaders?

The results of the opinion leader detection pipeline returned for each dataset consists of a labeled user list. placing each user in either of the following categories Normal, for users that are not deemed too special in any sense and Opinion leaders (OL) users that express a deviating characteristic in the measured metrics and exhibit a high influence potential. In this chapter the results of two runs with the two previously described datasets are analyzed, to get an overall understanding of the results and to try to answer my research questions.

User type	NYT	DST
Normal	402,795	31,167
Outlier	165	217
Opinion leader	18	29

Table 3 OL-pipeline results

In both datasets, the majority of the user base has been classified as 'Normal' users. Interestingly, despite maintaining consistent parameters across both runs, the German DST run resulted in a pipeline output with a proportionally higher number of 'Influential' users, by a factor of 20. With the outlier category being allocated in the middle ground I choose to count them towards normal users to streamline the analysis.

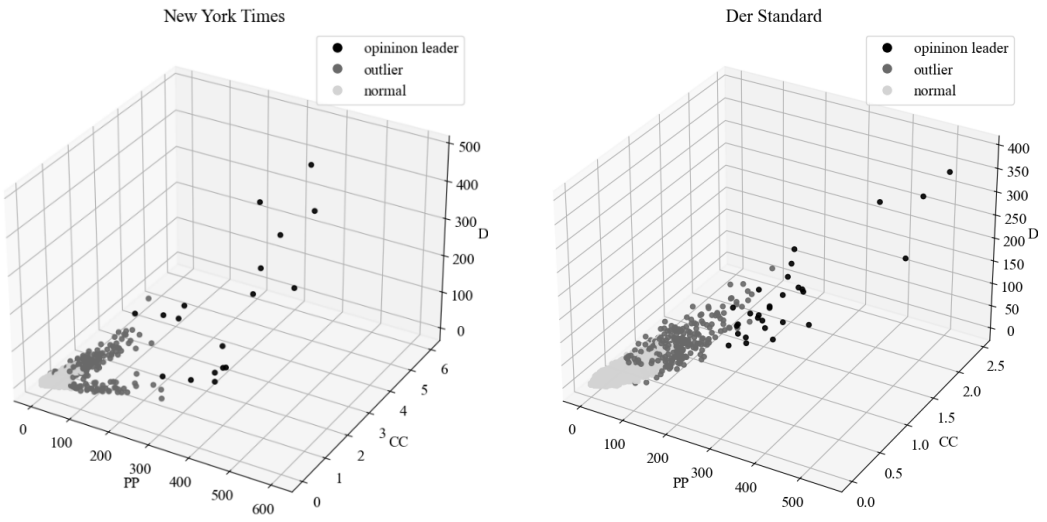

Figure 19 Comparison of pipeline results

RQ2

Research Question 2: Are opinion leaders merely prolific posters or do they actively shape the platform's discourse?

The users identified as being most influential in the respective datasets appear to have posted a markedly higher number of comments than other users. This indicates that, as anticipated, this category of users is a prominent contributor to the comments section. The number of comments per user in the Opinion leader set range from 205 to 7019 for the NYT and 415 to 2879 for DST, as illustrated in Figure 20. It is noteworthy that the algorithm did not select users based solely on their comment volume, given that there are 3765 other users in the NYT Dataset, and 182 other users in the DST Dataset, not categorized as opinion leaders, which also exhibited a high volume of comments. High volume in this context, means as many or more comments than the opinion leader with the fewest comments in that dataset.

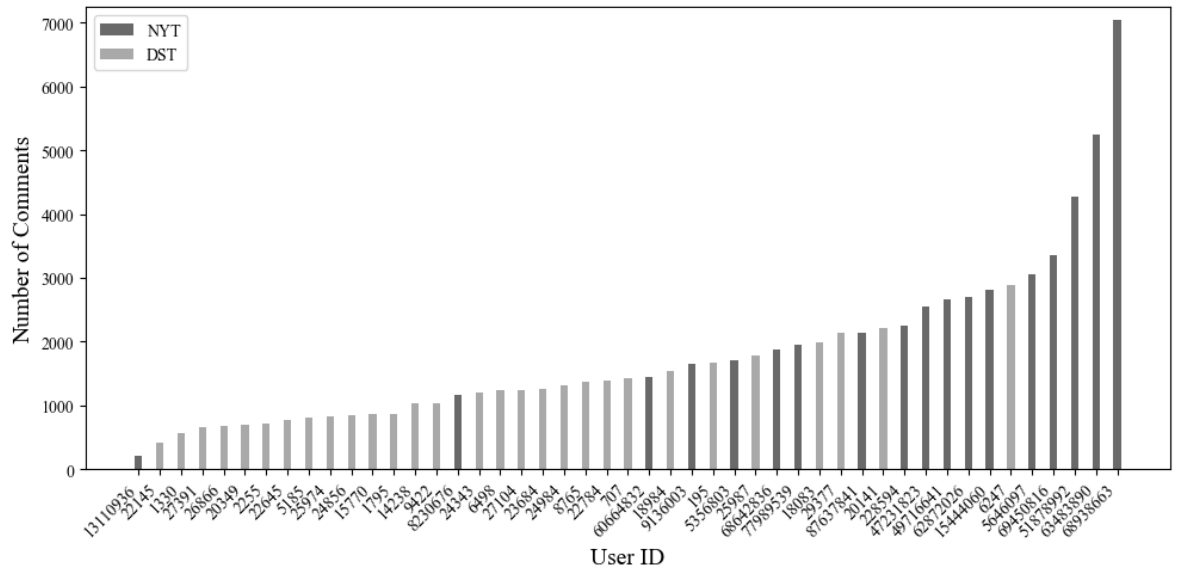

Figure 20 Commenting volumes

To test the differences numbers of comments posted between normal users and opinion leaders the two-sided Mann-Whitney U test was conducted with the alternative hypothesis that the distribution between the groups is not equal. As shown in Table 4, the test results for both datasets are highly significant, which allows to reject the null hypothesis and to conclude that there is a significant difference in comment distribution between normal users and opinion leaders.

Dataset	User label	median Comments	U	p
NYT	Normal	2	7234892	< .001
	OL	2393.5		
DST	Normal	5	903467.5	< .001
	OL	1196		

Table 4 Comparison of mean comments per user

Hypothesis 1 The engagement levels of comments by opinion leaders significantly differ from those of regular users, indicating their influential role.

Comparing the engagement levels with all the comments users have obtained over the observed timespan reveals that in the NYT dataset, opinion leaders receive, on average, more replies per comment (mean of 0.88) compared to regular users (mean of 0.41). However, the distribution of mean recommendations (likes) per user shows a small difference between opinion leaders (mean of 17.75) and regular users (mean of 19.39) in which normal users receive slightly more recommendations on average.

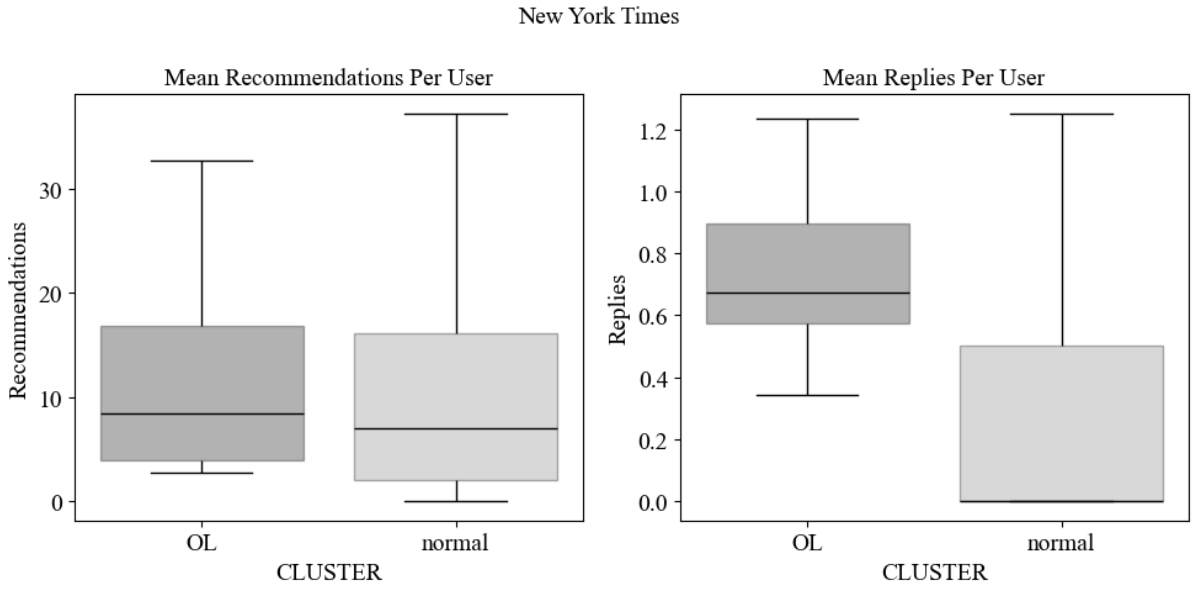

Figure 21 NYT user engagement

A two-sided Mann-Whitney U test was conducted for each comparison (see Table 4). The test results indicate that the distributions of mean recommendations per user do not significantly differ between the two groups. However, there is a significant difference in

mean replies per user between opinion leaders and regular users, with opinion leaders receiving a higher median reply count compared to regular users.

New York Times				
	Median OL	Median Normal	U	P
Mean Recommendations per User	8.4	7	4168732	0.264
Mean Replies Per User	0.67	0	5889524	<0.001

Table 5 Mann Whitney U test results NYT H1

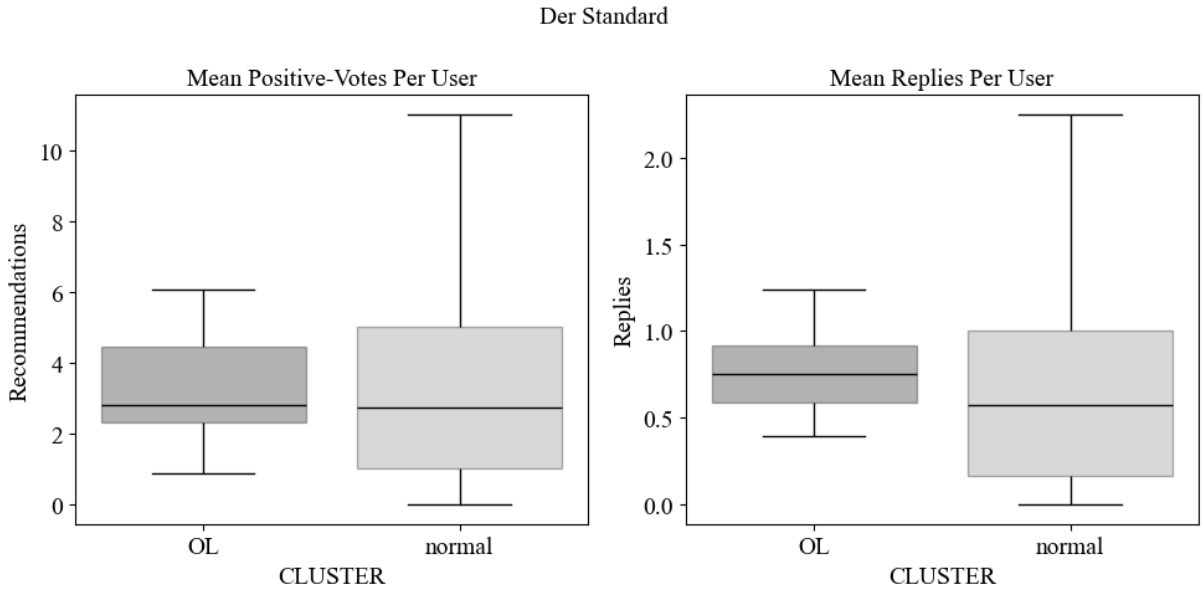

Figure 22 DST user engagement

A similar pattern is observed in the DST dataset (see Table 5). The distributions of mean recommendations per user do not significantly differ, with opinion leaders receiving slightly fewer upvotes on average (mean of 3.34) than regular users (mean of 3.85). Additionally, while the difference in mean replies per user between the groups is yet again significant, it is less pronounced in der Standard: opinion leaders receive an average of 0.76 replies per comment, compared to 0.66 for regular users.

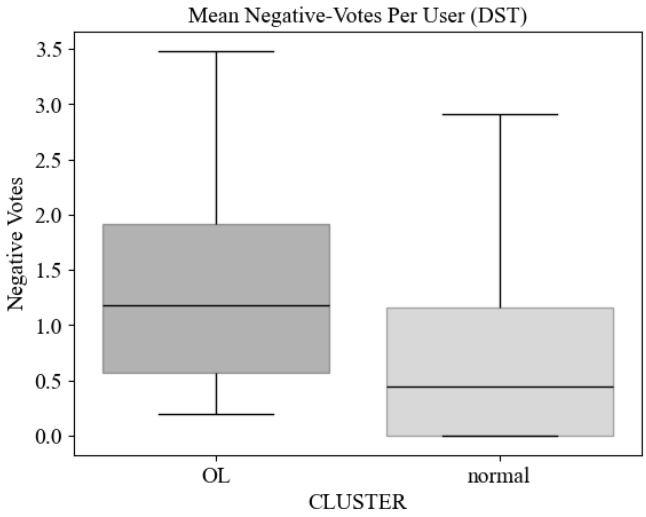

Figure 23 DST Negative votes

Something that can be only observed in the DST dataset, where negative downvotes are included, opinion leaders receive a higher average number of negative votes (mean of 1.38) compared to regular users (mean of 1.06), with a Mann-Whitney U test confirming that the distributions of negative votes significantly differ between the two groups (see Table 5).

Der Standard				
	Median OL	Median Normal	U	P
Mean Replies per User	0.75	0.57	570530.5	0.006
Mean Positive-Votes per User	2.79	2.75	499028.5	0.165
Mean Negative-Votes per User	1.18	0.44	641079	<0.001

Table 6 Mann Whitney U test results DST H1

To summarize, the analysis indicates that users labeled as opinion leaders receive a significantly higher volume of replies to their comments in both datasets. However, when examining the feedback received via the recommendation or "like" button, no significant difference was found between the two groups across both datasets. Only a significant

difference was observed in terms of negative feedback (limited to the DST dataset) with opinion leaders receiving significantly more negative feedback.

These results support the hypothesis that engagement levels, in terms of replies to comments by opinion leaders, are significantly different from those of regular users. However, when focusing on one-click feedback, the findings are less conclusive, as there is no significant difference in the distribution of positive feedback, with the exception of a higher volume of negative feedback observed in one dataset.

Hypothesis 2 As the volume of comments posted by any user increases, the engagement levels of comments by the same poster increase as well.

A straightforward linear correlation demonstrates that the total number of recommendations a user has received over the observed time span correlates positively with the number of comments posted by a user. This is true in both datasets with the DTS having a stronger correlation ($\rho(31411) = 0.872$, $p < .00$) than the NYT ($\rho(402367) = 0.645$, $p < .00$) This is not a surprising result, as the sum of recommendations or upvotes a user can receive can only increases and by posting more comments the opportunity of receiving recommendations increases, especially given that the lifespan of a comments in regards to interaction is fairly short as seen in figure 11.

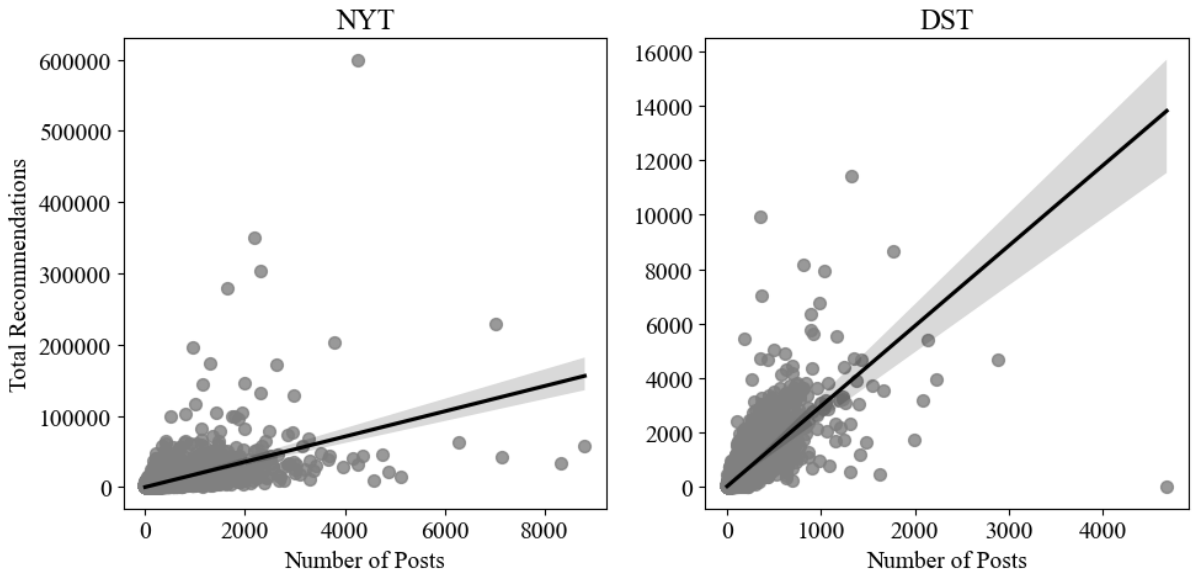

Figure 24 Recommendations to post distribution

To test this hypothesis on a per comment basis each comment made by a user is assigned a sequential number. Then, the Spearman correlation coefficient between the sequence of the comments and the number of recommendations each comment has received is calculated. To ensure the results were meaningful, the sample was restricted to users who had posted at least 10 comments. This results in a reduction of the user sample in the respective datasets: NYT to 61,252 and in DST to 12,265.

It should be noted that the sequential numbering only represents the order in which comments were made during the observed timeframe and is not reflective of the actual sequence of numbering of comments made by the user. It is therefore necessary to exercise caution in the interpretation of these results, which nevertheless can provide an indication of an underlying trend.

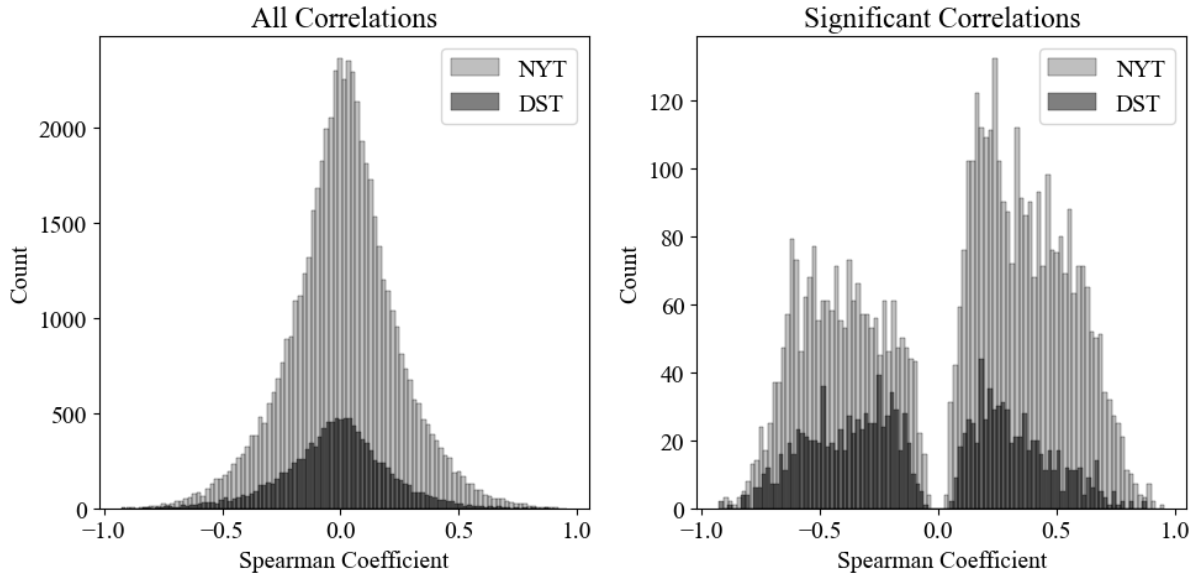

Figure 25 Histogram of correlations between comment sequence and recommendations / up-votes

The plotting of the calculated correlation coefficients for each user in a histogram reveals an approximate normal distribution, with the majority (90%) of the correlations not being statistically significant. The majority of users exhibit no correlation between the number of posts and the recommendations their comments receive. The statistically significant correlations are somewhat symmetrically distributed, with the majority of users exhibiting only weak correlations. The distributions are primarily skewed towards zero from both sides, then taper off towards -1 and 1.

While the negative correlations are somewhat counterintuitive given the inclusion of only positive upvotes or recommendations, users with negative correlation coefficients initially received a high number of likes on their comments, followed by a decline in recommendations with an increase in comments.

Overall, the evidence does not support the hypothesis that as the volume of comments posted by a user increases, so too does the engagement with their comments from other users.

RQ3

What impact does the influence of opinion leaders have on shaping the commenting behaviour of their peers in these specific online environments?

Dividing the Article base into two subsamples depending to whether opinion leaders have commented under them enables a comparison of user commenting behaviour with and without their presence. In detail I look at the engagement relations between users and the article as well as the time users have spent in an articles comment section. The split over both datasets behaves similarly with opinion leaders appearing in approximately 45% of articles as seen in table 7.

	Articles with OL	Articles without OL
NYT	7704	9083
DST	5420	6667

Table 7 Article distribution by presence of opinion leaders

For the comparison of the time spent on an article, the time difference (in minutes) between a user's first and last comment on the same article is considered. Due to the necessity for users to have posted at least two comments under the same article, this measure excludes users with a single comment from this analysis.

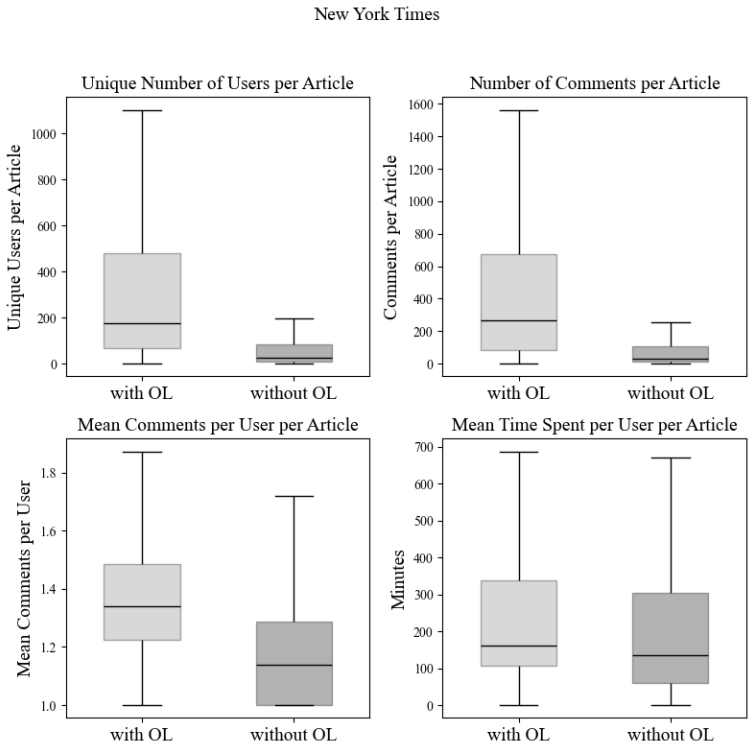

Figure 26 Comparison of articles with and without the appearance of OL NYT

The outliers (see Table 9) are excluded from the above visualization (Figure 26) to improve the readability of the boxplots. However, they are present in the dataset and the following tests.

Looking at the boxplot comparisons in Figure 26, it is noticeable that articles with opinion leaders have a higher number of unique users and comments per article. In detail, the median and distribution of unique users per article are considerably higher for articles with opinion leaders than for those without. Parallel to that, the median number of comments per article is also higher for articles with opinion leaders. It is noteworthy that the mean number of comments per user per article also increases for articles with opinion leaders, which may indicate a correlation between the presence of opinion leaders and increased commenting activity. Furthermore, the mean time spent per user per article also exhibits a distinction, which is most likely attributable to a correlation between the number of comments a user writes, and the time spent under that article.

New York Times				
	With OL		Without OL	
	Outlier range	Number of Outliers	Outlier range	Number of Outliers
Unique Number of Users per Article	1300 - 5638	386	357 - 4383	453
Number of Comments per Article	1837 - 8786	385	493 - 4718	454
Mean Comments per User per Article	1.88 - 4	386	1.54 – 2.66	455
Mean Time Spent per User per Article	1661 - 262000	373	1868 - 145000	326

Table 8 Outlier comparison of Fig. 26

For each comparison, a one-sided Mann-Whitney U test was conducted with the alternative hypothesis that the distribution of articles in which OL appear is higher and not equal than that of articles without OL. The results of this test can be found in table 9 and 11.

New York Times				
	Median With OL	Median Without OL	U	P
Unique Number of Users per Article	178	28	55565012.5	<0.001
Number of Comments per Article	268	33	56072377	<0.001
Mean Comments per User per Article	1.34	1.14	53352061.5	<0.001
Mean time spent per User per Article	161.62	134.65	28170665.5	<0.001

Table 9 Mann Whitney U test results NYT RQ3

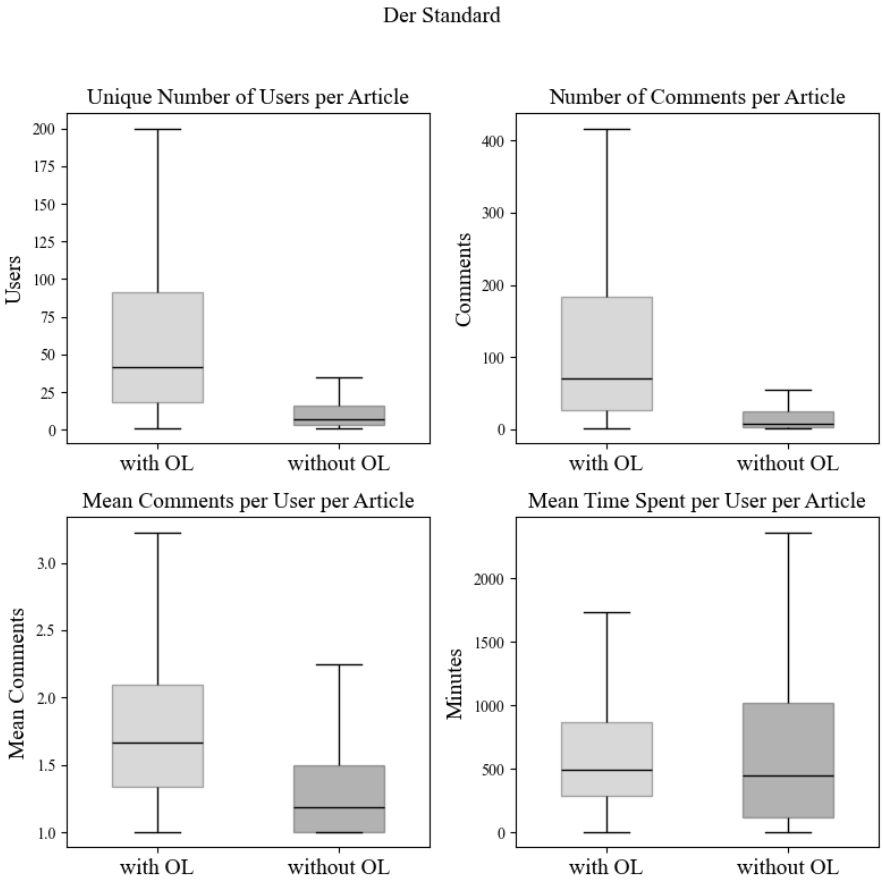

Figure 27 Comparison of articles with and without the appearance of OL DST

In the DST dataset, the distributions behave similarly to those in the NYT. The distributions for the number of users and the number of comments are both significantly higher for articles featuring OL. Furthermore, the difference in the distributions of mean comments per user per article is even more pronounced in the DST dataset, with a mean of 1.8 comments per user for posts with OL, in comparison to a mean of 1.4 comments per user for posts without OL.

One noteworthy distinction is that the mean time spent per article is higher for users in the DST dataset compared to the NYT dataset, irrespective of whether opinion leaders are present or not. The mean time spent per article by DST users is approximately five hours longer. It is important to note that the majority of users will most likely not spend the entirety of this time in the comments section, as this observation measures the difference between the first and last post. Some users probably revisit the comment section after a period of time has elapsed to follow up on their initial comments. Suggesting that DST users tend to engage in follow-up conversations spanning a longer period of time compared to NYT users.

Der Standard				
	With OL		Without OL	
	Outlier range	Number of Outliers	Outlier range	Number of Outliers
Unique Number of Users per Article	262 - 1371	268	48 - 244	332
Number of Comments per Article	602 - 3656	271	88 - 1386	329
Mean Comments per User per Article	3 - 9.6	271	2.4 – 2.5	334
Mean Time Spent per User per Article	2404 - 313919	253	3471 - 836137	200

Table 10 Outlier comparison of Fig.27

Der Standard				
	Median With OL	Median Without OL	U	P
Unique Number of Users per Article	42	7	30762887.5	<0.001
Number of Comments per Article	71	8	30641621.5	<0.001
Mean Comments per User per Article	1.7	1.2	27083401	<0.001
OP-Replies	492.6	448.8	10925711	<0.001

Table 11 Mann Whitney U test results DST RQ3

A comparison of the two samples of articles in both datasets (see Figure 25 and 26) reveals that the distribution of the number of users posting comments under articles as well as the distribution of the number of comments per article are significantly higher ($p < .001$) for articles featuring opinion leaders. It is not possible to conclude that the presence of OL is the cause of this difference in distributions, given that the content of the articles was not taken into account in the analysis. An alternative explanation for this trend may be found in the articles themselves rather than in the presence of opinion leaders. This could be the subject of further analysis, although it is beyond the scope of my thesis.

Regardless of the user or comment volume an article has attracted, it is visible that the mean number of comments per user remain significantly higher in articles featuring opinion leaders. This indicates that there is a higher rate of interaction among users when opinion leaders are present in the conversation.

RQ4

Does the tone of the conversation differ when users engage in discussions with opinion leaders compared to normal users?

Hypothesis 3: The toxicity levels of the conversations between opinion leaders and other users are lower in comparison to discussions between regular users.

The overall difference in toxicity scores between normal users and those labelled as opinion leaders differs significantly according to a one-sided Mann-Whitney U test. Indicating that the distribution of toxicity scores for comments posted by opinion leaders (OL) is stochastically less compared to the distribution of normal users. While statistically significant in both cases der Standard only exhibits very minor difference as is visible in Figure 28, with the majority of the distributions overlapping and the biggest differences being visible in the NYT dataset toward the tail of the distribution.

Dataset	NYT	DST
Mann-Whitney U	13892786470	13892786470
p	<0.01	<0.01

Table 12 Mann Whitney U test results DST H3

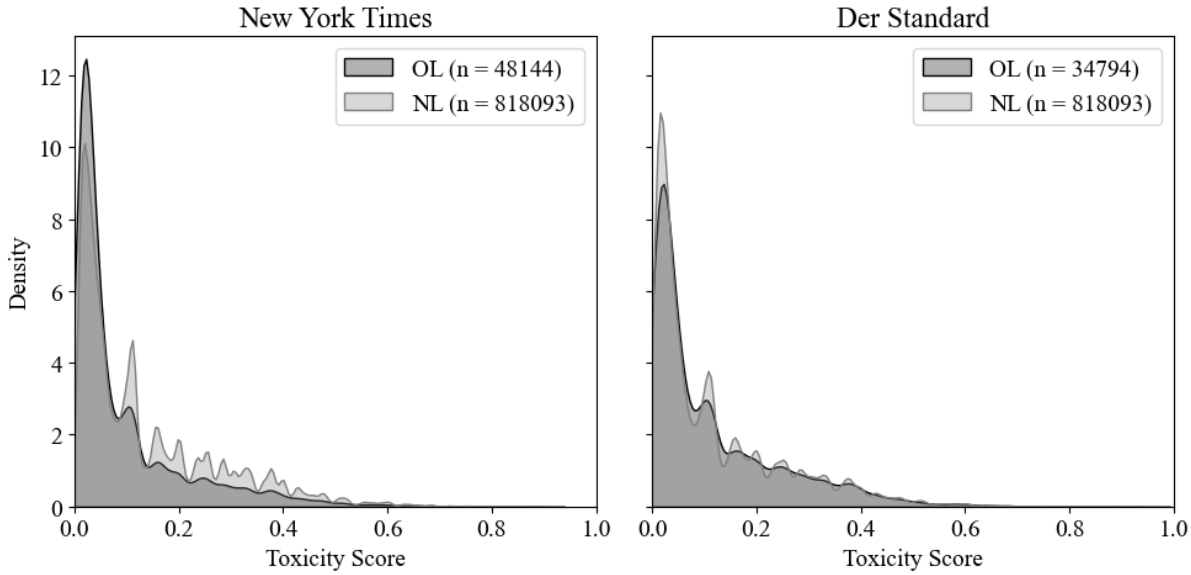

Figure 28 Comment toxicity value comparison by user type

Figures 28 and 29 present the distributions of toxicity scores as smoothed approximations using kernel density estimation (KDE). The density curves illustrate the relative shape of the distributions within each group, with normalization applied within each plot to allow for meaningful comparisons between the groups.

A clearer view of the differing communication behaviours between the two user groups emerges when we analyse the average toxicity across all posts made by each user. This results in a much narrower distribution of toxicity values for OL in both datasets. This can be explained firstly by the much smaller set of users in the OL sample but might also be indicative of a more balanced language used. The OL group is drawn toward a central tendency of 0.1, while the Normal users, in comparison, are more varied with two peaks in toxicity levels, suggesting that there might be two subsets of users: those whose comments are less toxic and some whose comments are more toxic.

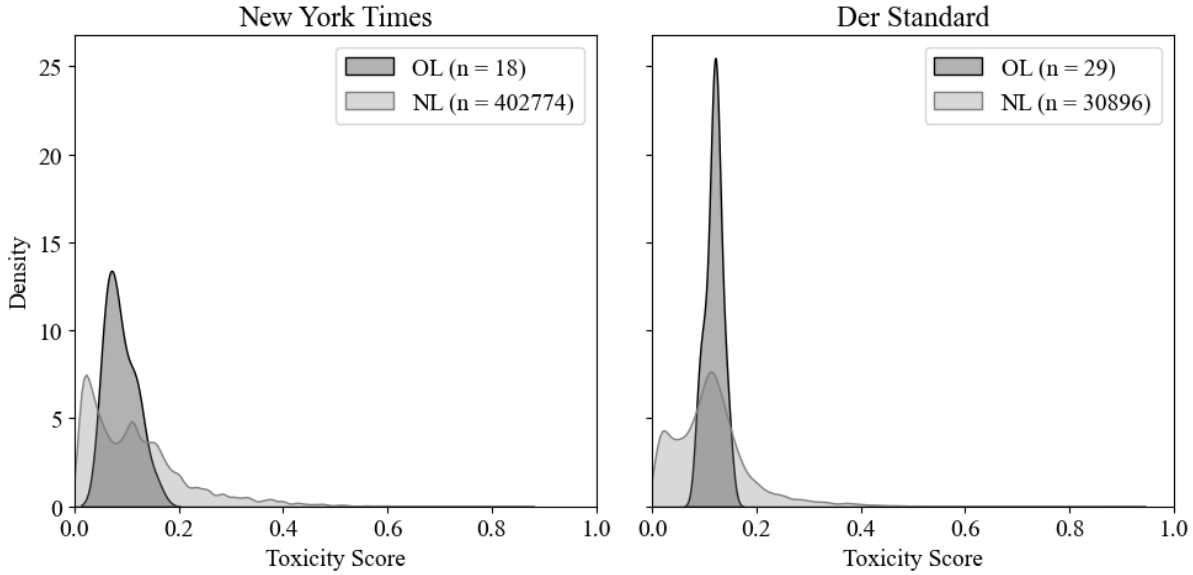

Figure 29 Mean user toxicity comparison by user type

To analyse how the interaction between users differs when they engage in discussions with opinion leaders compared to normal users, I took a sample from all top-level comments made by users (Original Posters OP) from each of the two groups that received at least ten replies in the NYT dataset and at least 5 replies in the DST dataset, indicating that they initiated a discussion of some sort. From these top-level comments, I then retrieved all corresponding replies. This enabled me to compare how users respond to comment threads initiated by OL

(Non-OP Replies), as well as how the OL respond to replies (OP Replies), all of which I can compare to a sample of top-level comments made by normal users.

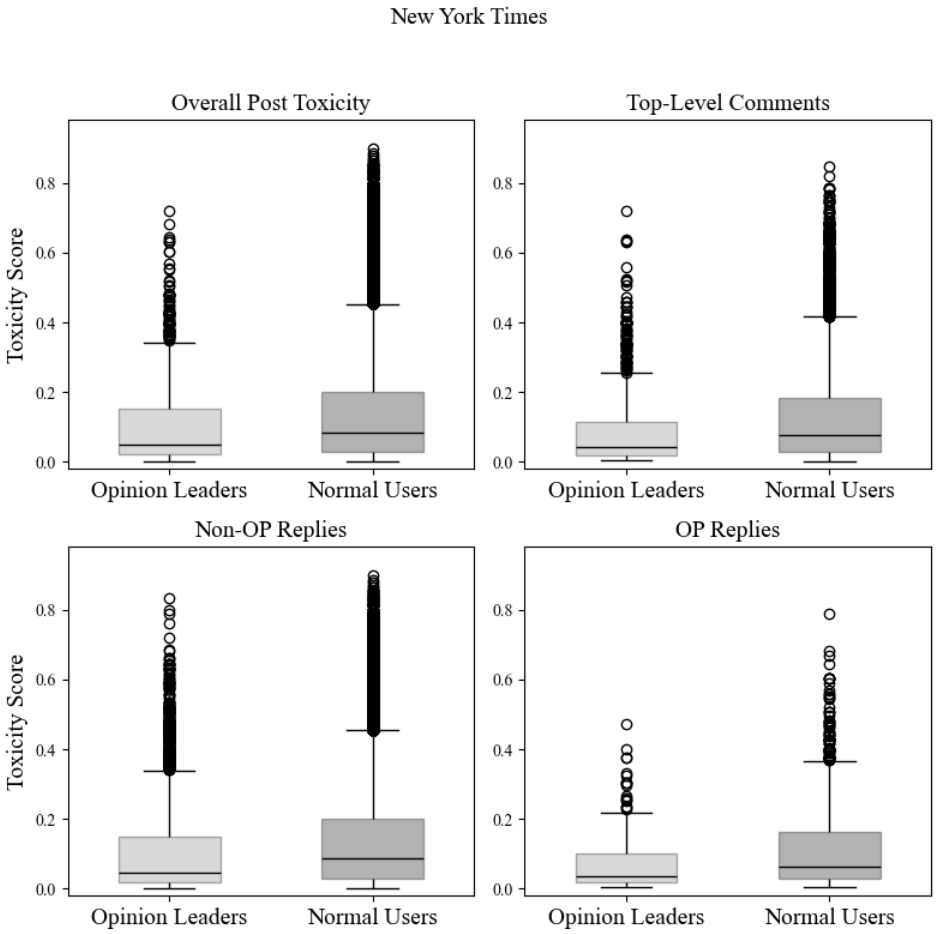

Figure 30 NYT conversation toxicity

As indicated in Figure 30 and supported by the consistently significant Kolmogorov-Smirnov test for all pairs in the NYT dataset, the post-toxicity distribution thread initiated by opinion leaders differs significantly from those started by normal users. Similarly, all the one-sided Mann-Whitney U tests indicate, as seen in table 13, that toxicity scores are significantly lower in opinion leader-initiated threads. While these differences are significant, they are rather small.

New York Times				
	Median OL	Median Normal	U	P
Overall, Post Toxicity	0.052	0.086	190032093	<0.001
Top-Level Comments	0.043	0.076	8948672	<0.001
Non-OP-Replies	0.047	0.086	877785482	<0.001
OP-Replies	0.037	0.064	101471	<0.001

Table 13 NYT OP-Comparison, Medians and Mann-Whitney U Test Results

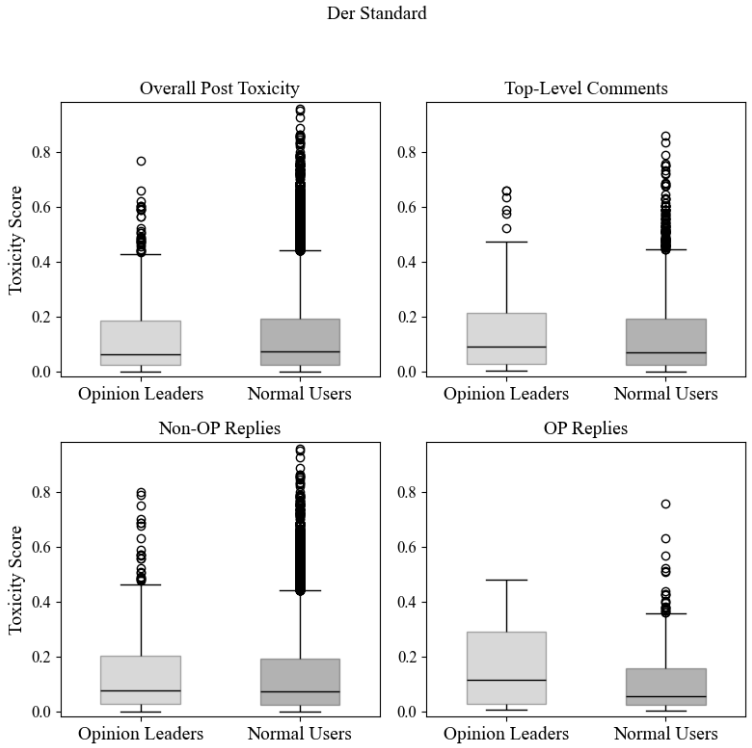

Figure 31 DST conversation toxicity

In the DST the results, Figure 31, the findings show no significant difference between the distributions in toxicity.

Der Standard				
	Median OL	Median Normal	U	P
Overall, Post Toxicity	0.062	0.070	19076324	0.113
Top-Level Comments	0.091	0.068	331912	0.905
Non-OP-Replies	0.076	0.071	13602892	0.606
OP-Replies	0.011	0.056	1594	0.618

Table 14 DST OP-Comparison, Medians and Mann-Whitney U Test Results

To summarize the combined results of the two datasets are conflicting making it not possible to conclusively support the hypothesis that conversations between opinion leaders and other users do not significantly differ from discussions between regular users.

4 Discussion

4.1 Summary of the Thesis

The aim of this thesis was to algorithmically identify opinion leaders within the comment sections of news websites. The experimental pipeline developed is capable of categorising users as normal users or OL. However, due to the limited tools available to evaluate the pipeline's results, further analysis focused on a comparative examination of the behaviour between the two groups, instead of testing the pipeline’s accuracy.

The analysis demonstrated that opinion leaders are, in general, prolific commenters, posting at a significantly higher rate than their peers. However, the mere fact of high commenting volume does not automatically result in the categorisation of a user as an OL, as proven by the presence of numerous other high-volume posters, who are not labelled as such. The relationship between posting volume and received engagement was also examined. No evidence has been found to suggest that prolific posting results in greater levels of received engagement.

Opinion leaders appear to impact overall engagement in comment sections. Articles where opinion leaders contribute, tend to attract more unique users and comments, with users spending more time on these articles. This suggests that opinion leaders may boost community activity. However, no causal relationship can be established, as it remains unclear whether the increased engagement is driven by the presence of opinion leaders or by other factors, such as the article content, timing, or external influences. Additionally, opinion leaders attracted a greater number of replies to their comments compared to normal users, supporting the idea that they stimulate interaction. While these identified users do not receive significantly more positive feedback (in the form of recommendations or likes) compared to normal users, opinion leaders did receive a greater number of negative feedback messages within the DST dataset. This can be interpreted that due to their more prominent status within the forum, other users tend to be more critical to their contributions, or it could be due to the nature of the content they disseminate.

Further the analysis of toxicity within interactions yielded mixed results. In the NYT dataset, threads initiated by opinion leaders exhibited lower toxicity than those started by regular

users, indicating that opinion leaders may contribute to more neutral discourse. In contrast, the DST dataset showed no significant difference in toxicity between user groups.

4.2 Limitations

A clear limitation of this thesis was, that it was designed around second-hand datasets collected half a decade ago, which drastically limited the amount and quality of information available. The age of the datasets limited the extent of user tracking on the platforms studied and a more comprehensive dataset, including post-reader tracking and user browsing patterns, would have been necessary in constructing a more comprehensive representation of user behaviour.

Secondly, although I used modern NLP techniques such as sentiment orientation and toxicity detection, these algorithms used offer limited insight when it comes to gain understanding of participants standpoints within discussion. A dedicated stance detection algorithm (a task that has yet to be solved by NLP algorithms) to detect agreement and disagreement would be needed. Without stance detection, the current toolset lacks the ability to accurately detect whether users are responding to each other in a supportive or disapproving manner.

Additionally relying on privately distributed third party machine learning models to classify comments proved limiting. Besides the vulnerability of access at the mercy of the API owner, the non-transparent algorithms make it difficult to interpret the returned results and weaken confidence in the results received. In consideration of the rapid advancements made in the field of NLP over the past decade, it is likely that these issues will be addressed in the near future.

4.2 Discussion and Outlook

The experimental algorithm developed to detect influential individuals demonstrated limited success, it is able to classify users yet without being able to quantitatively benchmark the results, it is challenging to empirically assess the effectiveness of the algorithm, which in turn complicates any claims of success.

The large sample size of users and comments, while initially a motivating factor for the project, turned out to be its Achilles' heel. This volume of users and the fractured nature of the comment forums make it difficult to identify small deviations in user influence behaviour. A potential avenue for further research could be to segment the user base into smaller, more focused groups of interest, potentially identified through topic modelling on the comments and the news articles content. By clustering users around specific topics, this approach could yield a more detailed analysis of user groups within these subgroups. Such segmentation might reveal differences in influence patterns which are currently overshadowed by the platform wide outliers.

This thesis demonstrates that opinion leaders in online news comment sections exhibit behaviours measurably distinct from their peers. While OL can serve as catalysts for discussion, there is also potential for them to misuse their influence by spreading misinformation or undermining the credibility of news content. Consequently, it would be of interest to more closely analyse the content produced by these users to further understand their impact on public discourse.

Literature:

Austrian Research Institute for Artificial Intelligence. (n.d.). Foromat: The formal and pragmatic foundations of formatting [Project]. Retrieved September 7, 2024, from https://www.ofai.at/projects/foromat

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2021). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. arXiv preprint arXiv:2104.12250. https://doi.org/10.48550/arXiv.2104.12250

Bennett, W. L., & Manheim, J. B. (2006). The one-step flow of communication. The ANNALS of the American Academy of Political and Social Science, 608(1), 213-232. https://doi.org/10.1177/0002716206292266

Bright, J., Bermudez, S., Pilet, J. B., & Soubiran, T. (2020). Power users in online democracy: their origins and impact. Information, Communication & Society, 23(13), 1838-1853. https://doi.org/10.1080/1369118x.2019.1621920

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. (2010, May). Measuring user influence in Twitter: The million follower fallacy. In Proceedings of the international AAAI conference on web and social media (Vol. 4, No. 1, pp. 10-17). https://doi.org/10.1609/icwsm.v4i1.14033

DER STANDARD. (2013). Rätselhaftes Wesen Foromat. Retrieved September 7, 2024, from https://www.derstandard.at/story/1363709332729/raetselhaftes-wesen-foromat

DER STANDARD. (2024). Allgemeine Geschäftsbedingungen: Forum. Retrieved September 7, 2024, from https://about.derstandard.at/agb/#Forum

Dubois, E., & Gaffney, D. (2014). The multiple facets of influence: Identifying political influentials and opinion leaders on Twitter. American Behavioral Scientist, 58(10), 1260-1277. https://doi.org/10.1177/0002764214527088

Erwe, L., & Wang, X. (2024). Comparison of VADER and pre-trained RoBERTa: A sentiment analysis application (Undergraduate thesis, Lund University).

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD '96) (Vol. 96, No. 34, pp. 226–231). AAAI Press.

Etim, B. (2013). Re: How does the NYT determine which articles have comments? [Comment on a Quora post]. Quora. Retrieved August 24, 2024, from https://www.quora.com/How-does-the-NYT-determine-which-articles-have-comments

Jigsaw. (n.d.). About the API attributes and languages. Perspective API. Retrieved June 30, 2024, from https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US

Jungnickel, K. (2018). New methods of measuring opinion leadership: a systematic, interdisciplinary literature analysis. International Journal of Communication, 12, 23.

Kang, M., Liang, T., Sun, B., & Mao, H. Y. (2023). Detection of opinion leaders: Static vs. dynamic evaluation in online learning communities. Heliyon, 9(4), e14567. https://doi.org/10.1016/j.heliyon.2023.e14844

Lazarsfeld, P., Berelson, B., & Gaudet, H. (1948). The People's Choice. New York: Columbia University Press.

le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In International Conference on Machine Learning (pp. 1188–1196). PMLR. https://doi.org/10.48550/arXiv.1405.4053

Lees, A., Tran, V. Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., & Vasserman, L. (2022, August). A new generation of Perspective API: Efficient multilingual character-level transformers. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 3197–3207). https://doi.org/10.1145/3534678.3539147

Liu, B. (2011). Web data mining: Exploring hyperlinks, contents, and usage data (2nd ed.). Springer. https://doi.org/10.1007/978-3-642-19460-3

Newman, N., Fletcher, R., Robertson, C. T., Ross Arguedas, A., & Nielsen, R. K. (2024). Reuters Institute Digital News Report 2024. Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-vy6n-4v57

Nielsen, J. (2014). Participation Inequality: Lurkers vs Contributors in Internet Communities, (URL: http://www.nngroup.com/articles/participation-inequality/).

Nogara, G., Pierri, F., Cresci, S., Luceri, L., Törnberg, P., & Giordano, S. (2023). Toxic bias: Perspective API misreads German as more toxic. https://doi.org/10.48550/arXiv.2312.12651

Odefey, M. A. (2011). Interpersonal communication and opinion leadership in the context of the 2009 German federal election: How the Internet raises the bar for most, but lowers it for some; and how ideas seem to flow from the Internet to the general public via opinion leaders (Doctoral dissertation, Otto-Friedrich-Universität Bamberg).

Page, L. (1999). The PageRank citation ranking: Bringing order to the web (Technical Report).

Perspective API. (n.d.). About the API: Attributes and languages. Retrieved September 7, 2024, from https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US

Perspective API. (n.d.). Perspective API. Retrieved September 7, 2024, from https://perspectiveapi.com/

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., & Gurevych, I. (2020). How good is your tokenizer? On the monolingual performance of multilingual language models. arXiv preprint arXiv:2012.15613. https://doi.org/10.48550/arXiv.2012.15613

Santana, A. D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. Journalism Practice, 8(1), 18-33. https://doi.org/10.1080/17512786.2013.813194

Soffer, O. (2021). Algorithmic personalization and the two-step flow of communication. Communication Theory, 31(3), 297-315. https://doi.org/10.1093/ct/qtz008

Song, K., Wang, D., Feng, S., & Yu, G. (2011, September). Detecting opinion leader dynamically in Chinese news comments. In International conference on web-age information management (pp. 197-209). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28635-3_19

The New York Times. (2017, June 13). Have a comment? Leave a comment. Retrieved September 7, 2024, from https://www.nytimes.com/2017/06/13/insider/have-a-comment-leave-a-comment.html

The New York Times. (2024). Subscription options. The New York Times. Retrieved August 24, 2024, from https://www.nytimes.com/subscription

The New York Times. (n.d.). The comments section. Retrieved September 7, 2024, from https://help.nytimes.com/hc/en-us/articles/115014792387-The-Comments-Section

Valente, T. W., & Pumpuang, P. (2007). Identifying opinion leaders to promote behavior change. Health Education & Behavior, 34(6), 881–896. https://doi.org/10.1177/1090198106297855

Van der Merwe, R., & Van Heerden, G. (2009). Finding and utilizing opinion leaders: Social networks and the power of relationships. South African Journal of Business Management, 40(3), 65-76. https://doi.org/10.4102/sajbm.v40i3.545

Vassio, L., Garetto, M., Leonardi, E., & others. (2022). Mining and modelling temporal dynamics of followers' engagement on online social networks. Social Network Analysis and Mining, 12(96). https://doi.org/10.1007/s13278-022-00928-2

Wang, Y., Gu, Y., & Shun, J. (2020). Theoretically-efficient and practical parallel DBSCAN. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (pp. 2555–2571). Association for Computing Machinery. https://doi.org/10.1145/3318464.3380582

Weimann, G., Tustin, D. H., van Vuuren, D., & Joubert, J. P. R. (2007). Looking for opinion leaders: Traditional vs. modern measures in traditional societies. International Journal of Public Opinion Research, 19(2), 173–190. https://doi.org/10.1093/ijpor/edm005

Datasets:

Dornel, B. (2021). *New York Times Articles & Comments (2020), Version 1* [Data set]. Kaggle. Retrieved February 29, 2024, from https://www.kaggle.com/datasets/benjaminawd/new-york-times-articles-comments-2020

Schabus, D., Skowron, M., & Trapp, M. (2017, August). One million posts: A data set of German online discussions [Data set]. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1241-1244).

Appendix A: Peak commenting days in the NYT dataset

Publication Date	Headline	Number of comments
05.02.2020	In Private, Republicans Admit They Acquitted Trump Out of Fear	3807
06.02.2020	Trump Hails Acquittal and Lashes Out at His 'Evil' and 'Corrupt' Opponents	2616
07.02.2020	Trump Fires Impeachment Witnesses Gordon Sondland and Alexander Vindman in Post-Acquittal Purge	4105
12.02.2020	Paging Michael Bloomberg	4220
13.02.2020	Barr Says Attacks From Trump Make Work 'Impossible'	4363
28.02.2020	No, Not Sanders, Not Ever	4610
04.03.2020	Joe Biden Just Performed a Miracle	2565
05.03.2020	Elizabeth Warren, Once a Front-Runner, Drops Out of Presidential Race	3719
06.03.2020	Trump's Calamitous Coronavirus Response	2048
11.03.2020	Behind Bernie Sanders's Decision to Stay in the Race	2968
18.03.2020	Give Every American $2,000, Immediately	1735
24.03.2020	Trump Expresses Outrage at Having to 'Close the Country' to Slow Virus	2767
01.04.2020	Coronavirus Spreads Amid Supply Shortages, Stay-at-Home Orders and Sobering Economics	1763
06.04.2020	Has Anyone Found Trump's Soul? Anyone?	3861
08.04.2020	Bernie Sanders Drops Out of 2020 Democratic Race for President	4209
14.04.2020	Coronavirus Updates: Trump Halts U.S. Funding of World Health Organization	3250
21.04.2020	California Announces Early Coronavirus Deaths; Trump Narrows Immigration Ban	1920
24.04.2020	Trump Muses About Light as Remedy, but Also Disinfectant, Which Is Dangerous	2797

02.06.2020	Peaceful Protesters Defy Curfews as Violence Ebbs	2685
01.07.2020	Is Trump Toast?	3138
07.07.2020	Mary Trump's Book Accuses the President of Embracing 'Cheating as a Way of Life'	3306
08.07.2020	Biden Should Not Debate Trump Unless …	3793
30.09.2020	With Cross Talk, Lies and Mockery, Trump Tramples Decorum in Debate With Biden	4044

Appendix B: Python Packages and Modules Used

da Costa-Luis, C., Larroque, S. K., Altendorf, K., Mary, H., richardsheridan, Korobov, M., Yorav-Raphael, N., Ivanov, I., Bargull, M., Rodrigues, N., Chen, G., Dektyarev, M., mjstevens777, Pagel, M. D., Zugnoni, M., JC, CrazyPython, Newey, C., Lee, A., … McCracken, J. (2024). tqdm: A fast, Extensible Progress Bar for Python and CLI (v4.66.5) [Software]. Zenodo. https://doi.org/10.5281/zenodo.13207611

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX [Software]. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th Python in Science Conference* (pp. 11–15). Pasadena, CA, USA.

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy [Software]. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hutto, C. J., & Gilbert, E. E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of social media Text [Software]. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). scikit-learn: Machine Learning in Python (Version 0.24.2) [Software]. scikit-learn. https://doi.org/10.48550/arXiv.1201.0490

Reitz, K. (2020). Requests: HTTP for Humans (Version 2.25.1) [Software]. Requests. https://requests.readthedocs.io/en/master/

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP* [Software].*Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

The Matplotlib Development Team. (2024). Matplotlib: Visualization with Python (v3.9.2) [Software]. Zenodo. https://doi.org/10.5281/zenodo.13308876

The pandas development team. (2024). pandas-dev/pandas: Pandas (v2.2.2) [Software]. Zenodo. https://doi.org/10.5281/zenodo.10957263

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python [Software]. Nature Methods, 17(3), 261-272.

Guhr, O., Schumann, A.-K., Bahrmann, F., & Böhme, H. J. (2020, May). Training a broad-coverage German sentiment classification model for dialog systems. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 1620–1625). European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.202.pdf

Waskom, M. L. (2021). seaborn: statistical data visualization [Software]. Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021

Wang, Y., Gu, Y., & Shun, J. (2020). Theoretically-efficient and practical parallel DBSCAN. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (pp. 2555–2571). Association for Computing Machinery. https://doi.org/10.1145/3318464.3380582

As well as the following built-in Python modules:

Python Collections (collections)

Python itertools (itertools)

Python statistics (statistics)

Python math (math)

Requests (requests)

Python time (time)

Python json (json)