PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

UNDERSTANDING INCIVILITY AND ENGAGEMENT IN NEWS FORUMS THROUGH NLP AND GRAPH THEORY

FLORENCIA BARRIOS MONTERO

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Advisor:
JUAN LORENZO REUTTER DE LA MAZA

Santiago de Chile, January 2021

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

UNDERSTANDING INCIVILITY AND ENGAGEMENT IN NEWS FORUMS THROUGH NLP AND GRAPH THEORY

FLORENCIA BARRIOS MONTERO

Members of the Committee:

JUAN LORENZO REUTTER DE LA MAZA

FRANCO WILFREDO PEDRESCHI PLASENCIA

MAGDALENA CAROLINA SALDAÑA VILLA

DENIS ALEJANDRO PARRA SANTANDER

DocuSigned by: Juan Reutter 8BE1FF0C178C403...
DocuSigned by: Franco Pedreschi 7AE8EB3303994E3...
DocuSigned by: Magdalena Saldaña 85B0F8ECD5534ED...
DocuSigned by: 0F3E21932A2E493...

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Santiago de Chile, January 2021

Gratefully to my family, friends and
to my advisor Juan

ACKNOWLEDGEMENTS

I want to start this section by thanking my advisor Juan Reutter for supporting me throughout this journey. Thank you for believing in me and for encouraging me when everything felt uphill. Finally, thank you for not letting me settle for little and encouraging me to take a more challenging but more satisfying road. I would also want to acknowledge Magdalena Saldaña. Thank you for transmitting Juan and I all your knowledge in the area (because we entered an unknown territory) and being a great interdisciplinary team.

Special thanks to my parents Juan Pablo and Ana María, which encouraged me to challenge myself and work hard for my goals and dreams from a very young age. Also, for allowing me to study in a different city and for trusting me when I decided to go for a Computer Science major. To my brothers Juan Pablo, Benjamín, and Maximiliano, thanks for putting up with my mood swings and being there for me. To my grandmother, María Angélica, thank you for receiving me when I got to Santiago and making me feel at home.

It would be unfair not to mention my friends in this section, as this journey would not have been the same without you. Thank you for your constant worry and cheering me up every time I felt it was hard to finish this investigation. Special mention to the *Parruka Seminars*, which made us stay active working on our theses in the middle of the global COVID pandemic. To my dearest *La Tribu*, our friendship is irreplaceable.

Finally, I would like to thank all the DCC staff, professors, and friends for being a second family. I was fortunate to get to know all of you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Social networks are under scrutiny because of the perceived levels of incivility among users. Within these networks, news forums and comment blogs are particularly interesting. They gather discussion about a very mixed set of content, including political and entertainment and other quotidian material. This mix raises a series of questions: is the type of news related to incivility and engagement levels in the news comment blogs? Are these results the same for every culture? And lastly, is there a correlation between incivility and engagement?

To answer these questions, we have downloaded over a year of all posted news and comments of a well-known Chilean news portal and used this data to train a classifier that can automatically tag civil and uncivil comments with reasonable confidence. With the classifier implemented, we automatically labeled 4.5 million comments and later used them to measure incivility and engagement across all news topics. The engagement was measured in a novel way, as we introduced network theory.

Results showed that incivility varies among different news topics and different cultures. For example, sports present high incivility in the Global North, but not in Chile. About engagement, topics are also a determinant factor, but the subjects with higher engagement are not necessarily the same as the ones that present higher incivility. Finally, the levels of incivility are not directly proportional to the levels of engagement. We present a novel framework based on quadrants of behavior to understand the relation between these two phenomena.

Keywords: incivility, engagement, social networks, content analysis, network analysis, facebook comments, bert, machine learning, local news, online news, user comments.

RESUMEN

Las redes sociales están bajo escrutinio debido a los niveles percibidos de incivlidad entre los usuarios. Dentro de estas, los foros de noticias son particularmente interesantes, ya que reúnen debates sobre un conjunto muy variado de contenido, que incluye tanto material político, de entretenimiento y cotidiano. Esta mezcla levanta una serie de preguntas: ¿está el tipo de noticias relacionado con los niveles de incivilidad y participación en los blogs de comentarios de noticias? ¿Son estos resultados los mismos para todas las culturas? Y, por último, ¿existe una correlación entre la falta de civilidad y participación?

Para responder a estas preguntas, hemos descargado más de un año de todas las noticias y comentarios publicados de un conocido portal de noticias chileno, y usamos estos datos para entrenar un clasificador que puede etiquetar automáticamente los comentarios civiles e inciviles con razonable confianza. Con el clasificador implementado, etiquetamos automáticamente 4.5 millones de comentarios que luego los usamos para medir la incivilidad y la participación en todos los tópicos de noticias. La participación se midió de una manera novedosa, ya que se introdujo teoría de redes.

Los resultados mostraron que la incivilidad varía entre diferentes temas de noticias y diferentes culturas. Por ejemplo, el tema deportes presenta alta incivilidad en Estados Unidos pero no en Chile. En cuanto a la participación, los tópicos también son un factor determinante pero los temas con mayor participación no son necesariamente los mismos que presentan mayor incivilidad. Por último, los niveles de incivilidad no son directamente proporcional a los niveles de participación. Presentamos un marco novedoso basado en cuadrantes de comportamiento para comprender la relación entre estos dos fenómenos.

Palabras Claves: incivilidad, participación, redes sociales, análisis de contenido, análisis de redes, comentarios de facebook, bert, aprendizaje de máquina, noticias locales, noticias online, comentarios de usuarios.

1. INTRODUCTION

Over the last decade, news consumption has suffered drastic changes, as digital platforms have come to compete with traditional media such as the newspaper and radio stations (Ksiazek, Peer, & Lessard, 2016). With this, the experience of online news-consumption has mutated as Facebook pages, or other forms of social networks, and their user comments are now an inseparable part of the process (Su et al., 2018).

But the promise of the Web and social networks as a space to productively exchange ideas has been confronted with recent concerns about the political polarization exposed in these discussions (Rains, Kenski, Coe, & Harwood, 2017). Special attention has been put on commenting features on online media due to the prevalence and the impact it can cause on user's attitudes and behaviors (Su et al., 2018).

Several investigations have focused on studying the high level of incivility and aggressiveness that manifests a considerable number of individuals when exchanging opinions online (Coe, Kenski, & Rains, 2014; Rains et al., 2017; Saldaña & Rosenberg, 2020; Su et al., 2018; Van Duyn, Peacock, & Stroud, 2019; Chen, Fadnis, & Whipple, 2020). This problem is not only for readers who must face large volumes of aggressive content but also for the media outlets. In conversations with *Bío-Bío*, a national broadcast station, they perceived that incivility affects the credibility of their content and drives away the real interested users. For all this, it has been put into doubt whether genuine deliberation can be reached on the Internet (Rosenberg, 2018).

However, incivility is not the only concern of news outlets delivering their content through social media. The engagement, in the sense of how much people interact with each other, has also been a source of study among the academic community. Indeed, online activity consists of an equal share of interactions among pears and not just single-shot, unidirectional interactions (Shugars & Beauchamp, 2019), and thus it is important to understand why and how are these interactions occurring.

The intrinsic nature of social media involves engagement and discussion (Lysak, Cremedas, & Wolf, 2012), and these features seem–at times–to engage and correlate with each other. For example, the theory states that negative mood (being exposed to unrelated negative prior stimulus) and negative discussion context (exposure to prior trolling behavior) increases the engagement and the probability of a user trolling back (Cheng, Bernstein, Danescu-Niculescu-Mizil, & Leskovec, 2017). However, Shugars and Beauchamp (2019) suggests that there also exists a more productive mode of engagement, where participants exchange content in a genuine attempt to persuade or inform. This may be referred to as a true deliberative argument.

Our efforts on understanding engagement and why people come back to a repeated argument fall in the line of Shugars and Beauchamp (2019), as we are interested not only in the engagement of a user in a conversation like on Coe et al. (2014) but also to responses and re-engagements depending on the post characteristics. Moreover, previous studies have emphasized the importance of exploring the complexities of engagement beyond general measures (Ksiazek et al., 2016).

In this investigation, we focus on analyzing incivility and engagement on a large scale. We expect to answer how much incivility is present in online news discussions and how involved users are in them. Previous research has presented divergent results. While Coe et al. (2014) found that one in every five comments is uncivil, others found that this number increased to almost one in every three for political-themed discussions (Saldaña & Rosenberg, 2020).

The size of the data we work with puts manual classification completely out of reach, so we resort instead to state-of-the-art natural language processing (NLP) tools (Devlin, Chang, Lee, & Toutanova, 2019). However, deploying an NLP classifier that can detect such types of comments is not an easy task. First, we need to start from a pre-trained model, as annotating a dataset large enough to train a classifier from scratch is a costly and time-consuming undertaking (Davidson, Sun, & Wojcieszak, 2020). We rely on neural networks for our work as simpler methods like support vector machines (SVM) or

logistics regressions have been outperformed in this task before (Sadeque et al., 2019; Maity, Chakraborty, Goyal, & Mukherjee, 2018, Davidson et al., 2020).

To understand the factors that condition incivility and engagement, a study of topics will be carried out with a similar method to the proposed by Coe et al. (2014). We expect to answer if our findings follow the same results of previous literature, or due to the local context and cultural differences, the results vary in the topics that raise the most incivility and engagement. Remarkably, we introduced network theory to understand engagement, work that had not been done until now.

Finally, our last analysis will center on the correlation between incivility and engagement. Could it be that users tend to be more civil when there is more engagement, as proposed by Coe et al. (2014)? We expect no linear relationship between these two phenomena and that their relation responds to a more complex correspondence.

As a footnote, we need to bring out that, while altering the context of a discussion (like hiding uncivil comments) may increase civility perception (Cheng et al., 2017), it is not in our interest to provide a tool to produce censorship in discussions.

2. RELATED WORK

To begin, we review the literature on incivility and engagement in online discussions. The review will be mainly centered on topic analysis, as it represents the core of our investigation. Since we are also building an automatic classifier for detecting incivility, we also review the main studies in this area.

Incivility has a wide range of definitions. Coe et al. (2014) define it as features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topics. Others define incivility as a manner of offensive interaction that can manifest as aggressive commenting, incensed discussion, hate speech, and harassment (Antoci, Delfino, Paglieri, Panebianco, & Sabatini, 2016). Maity et al. (2018) remark that incivility is an act intended to hurt, embarrass, or humiliate another person mentally. Finally, Hwang, Kim, and Huh (2014) define incivility in political discussions as an extreme form of polarized discourse, where participants use disrespectful statements towards the opposite political party or its members. Although the definitions vary, there are core concepts transversal to all definitions: incivility is pronounced as an act intended to harm and disrespect others.

This phenomenon has been widely studied in different sources of discussion. Some studies focus on comments posted on newspapers websites (Coe et al., 2014; Saldaña & Rosenberg, 2020; Su et al., 2018; Chen, 2017), and others focus on Twitter (Shugars & Beauchamp, 2019; Maity et al., 2018), essentially because it has turned into a destination where people abuse and act in an uncivil mode (Maity et al., 2018). Posts on Reddit have also been studied, as the platform's anonymous nature tends to encourage long and complex discussions (Davidson et al., 2020).

Engagement can be understood as a phenomenon that describes assorted user attention and involvement with media (Ksiazek et al., 2016). Shugars and Beauchamp (2019) add that it can be understood as the decision to go back to a repeated argument and comment again.

2.1. Incivility, news sources, and topic analysis

Coe et al. (2014) were among the first to explore how incivility varies in public discussions depending on the article's topic. This was done by analyzing the incivility present in comments on articles of a newspaper website (Arizona Daily Star). They presented a dataset labeled by experts, with 300 articles and 6000 comments. Each message was given zero to five incivility forms: name-calling, aspersion, lying, vulgarity, and pejorative for speech. Among other results presented, they found that serious topics ("harder news") like economy, politics, taxes, and crime gathered greater incivility (one uncivil per four posted). On the contrary, topics like health and lifestyle gathered lower levels of incivility. Exceptions were sports and immigration. About sports, it was the topic that presented the higher levels of incivility reported (29.8%) despite not being in the category of "harder news". About immigration, the rate was significantly low (15.3%) even though it is considered a topic that triggers reactions in political matters.

Saldaña and Rosenberg (2020) analyzed incivility during election periods in Chile. Among their study questions, they analyzed to what extent users are uncivil when commenting on election-related articles. Results showed that 29% of the comments were uncivil and that when mentioned, female politicians received 13.16% more uncivil comments ($p < 0.001$) than their male counterparts. Finally, they found that incivility in online comments sections of Chile is 45% more than studies made in the United States (Coe et al., 2014).

Su et al. (2018) compared patterns of incivility in Facebook comments across a wide range of news outlets. They categorized them into four types: national, local, conservative partisan, and liberal partisan. They collected more than 200 million comments and labeled them through a content-analysis software (Crimson Hexagon ForSight). Among their findings, they present that there is a significant relationship between forms of incivility and the news outlet category. Liberal news outlets were the most civil, and local-news outlets generated the most uncivil environment.

The association between gender and incivility has also been studied. Van Duyn et al. (2019) found that women are less likely to comment online in political discussions than men. Also, they discovered that there are differences in perceptions of incivility. Women rated the comments in the study less uncivil and, therefore, more civil than men.

Finally, Chen et al. (2020) studied incivility in comments about race-related shootings in the United States. They conducted the study with 1840 comments posted on three different sites of news organizations. Among their research questions was whether uncivil or deliberative attributes would prevail in the comments. Their results revealed that incivility predominated in race-related comments. Also, incivility varied across different types of news outlets, which matches the findings of Su et al. (2018).

2.2. Engagement, news sources, and topic analysis

The engagement has also been studied. Shugars and Beauchamp (2019) analyzed Twitter interactions between users to develop a model to predict the decision to re-engage or exit after the initial decision to interact. To measure how the tweet's topic influenced engagement, they used different measures as favorite count, retweet count, time of the day, and the presence of one or more topics, among others. Their results showed that users tend to engage with others who are different and over ideological divides. Topics that draw more engagement were racism and humanitarian crises in Puerto Rico.

Coe et al. (2014) also analyzed the engagement with other discussion participants. Engagement was defined as when a user directly replied to another participant's message, and in most cases, the reply included a quotation to the original message. Of all the comments, they found engagement in 42.5%. Their main finding was that comments that presented greater engagement were more civil since they were less likely to include incivility in their messages. They found that only 15.5% of these comments included some form of incivility, against 25.6% of disengaged comments. There were no analyzes regarding the news topics.

Ksiazek et al. (2016) analyzed commenting interactions (user-content) and replying to other's interactions (user-user) among YouTube news videos. User-user interactions can also be understood as conversations between users. One hypothesis tested whether soft news videos (sports, science, health, and technology) exhibit more conversations than hard news videos (politics, elections, foreign affairs, and crime). To test the hypothesis, they compared the ratio of replies between the two aforementioned news types. Their results revealed no differences in the probability of engaging associated with this factor.

Finally, Cheng et al. (2017) presents that troll behavior engages other users in the same conduct and persists across them. This suggests that there might be a correlation between incivility (as it may include troll attitude) and engagement.

2.3. Automatic incivility classification

Several attempts have been made to achieve adequate results on automatic incivility detection. We can divide the studies into two groups: traditional techniques like SVMs and logistic regressions and more recent and novel techniques like neural networks and, more specifically, deep learning. We briefly explain each study and the results they achieved for their classifiers, which can be compared using three metrics: the precision (the proportion of the positively classified instances that are truly positive), the recall (the proportion of the truly positive instances that are classified as positive), and the F1 score (the harmonic mean of precision and recall). Through this subsection, we assume familiarity with standard tools and algorithms for data-mining. For further details, we refer the reader to (Rajaraman & Ullman, 2011; Friedman, Hastie, & Tibshirani, 2001).

Theocharis, Barberá, Fazekas, and Popa (2020) build an automatic binary classifier of incivility using logistic regression. The data was extracted from Twitter and focused on tweets about Congress members in the United States. The labeling process was outsourced as they used an online platform for crowd-coding (CrowdFlower, now called Figure Eight). They only coded whether the tweet was civil or uncivil, resulting in 26% of incivility

throughout the training set. The best results were obtained when their model was trained using a much bigger labeled dataset, in which additional tweets were labeled according to Google's perspective API (which is much more error-prone). With this augmented dataset, they achieved a precision of 73% and a recall of 61% for the uncivil class.

Stoll, Ziegele, and Quiring (2019) analyzed different traditional techniques for detecting uncivil and impolite comments. They annotated a dataset of 10114 comments from Facebook sites of German media, which resulted in 16.6% of incivility. They present that Naive Bayes (NB) algorithms outperformed logistic regressions, decision trees, and SVMs (by, on average, 3%). The best results of NB were an F1 score of 0.069, recall of 0.80, and precision of 0.61.

Maity et al. (2018) tested different baseline classifiers using SVMs, k-nearest neighbors, and logistic regressions against two proposed deep learning models: Bi-directional Long short-term memory and Character Convolutional neural network. The data was extracted from Twitter, and they managed to annotate 2427 tweets. The criteria for labeling messages as uncivil were: blackmail or threads, insult, cursing, and sexual harassment. They outperformed the baselines by 6.5% on F1-score and achieved the best results with the Character Convolutional neural network model combined with an opinion conflict metadata. The F1-score of this model was 0.82.

Sadeque et al. (2019) developed a multi-label classifier using bidirectional GRUs (gated recurrent units). The input they used was the labeled data provided by Coe et al. (2014). Their classifier detected name-calling with a precision of 45.76% and a recall of 50.63% (F1 = 48.07%). Vulgarity was detected with a precision of 48.72% and a recall of 57.57% (F1 = 52.77%). Finally, they found that adding the metadata present in the dataset as auxiliary features at the last layer of the model (sigmoid) had virtually zero effects on improving the results.

On the same line, Davidson et al. (2020) recently developed binary incivility classifiers using BERT and DistilBERT. Their dataset consisted of Reddit posts and their comments

(5000 in total), annotated by experts on four categories: name-calling, aspersion, pejorative or disparaging remarks, and vulgarity. They pre-trained the models on 3 million unlabeled Reddit posts and then fine-tuned them with the 5000 labeled comments for the classification task. Their best results were obtained with DistilBERT, achieving a precision of 0.936, recall of 0.702, and F1 of 0.802. Authors also propose a more lightweight option based on performing logistic regression on a dataset previously labeled with their DistilBERT classifier and Twitter data annotated by Theocharis et al. (2020). Post and comments were fed to the regression using an embedding based on TF-IDF. With that model, they achieved an F1 score of 0.782, which is competitive with their previously mentioned classifiers.

3. RESEARCH QUESTIONS AND HYPOTHESES

While studies about incivility in news comments have flourished in recent years, observing several issues in different contexts (e.g., Humprecht, Hellmueller, & Lischka, 2020; Kim & Herring, 2018; Rossini & Maia, in press), research on engagement in news comments is not as common, especially in contexts other than the U.S. As such, this study aims to tackle both concepts together. Our first research question asks:

RQ1. How much a) incivility and b) engagement are present in online news discussions posted to Chilean news outlets?

The literature described in Section 2 suggests a relationship between incivility and news topics, while the relationship between engagement and news topics is not as clear. Consequently, we pose a hypothesis and a research question to address these concepts:

H1. Hard-news topics will concentrate more incivility than soft-news topics.

RQ2. Is there any association between engagement and news topic?

Finally, we aim to observe if there is a relationship between incivility, engagement and news topics. We ask:

RQ3. What is the relationship between incivility, engagement and news topic?

4. METHODS

4.1. Obtaining the data

The data was obtained by querying news posts on the Facebook account of *Radio Bío-Bío*, one of the most trusted radio stations in Chile[1], together with all the comments and replies made by Facebook users to these posts[2]. Every day, the station posts over two hundred news articles online, and users can express their opinion and get involved in discussions through comments and replies to the post. Our analysis contemplated 231 days of data, in which we got more than forty thousand news and almost five million comments and replies.

From each post, we got the original article, the reactions to the post (which included the number of clicks and impressions and the number of "likes", "loves", "wows" and "hahas"), the comments, and it's reactions, which included an identifier for the author and the replies to each message (with the same data as the comments). The list of all fields retrieved by the API is given by the schema representing the shape of each post's JSON document and is shown in Figure 4.1.

4.2. Annotating uncivil comments

The next step is to train a neural network-based NLP machine to classify uncivil comments. By previous results using different techniques on this task (Theocharis et al., 2020; Stoll et al., 2019; Maity et al., 2018; Sadeque et al., 2019; Davidson et al., 2020), we build our machine using google's BERT framework (Devlin et al., 2019). More specifically, we use BETO (Cañete et al., 2020), which provides an initial pre-training in Spanish

[1]See Reuters Institute's Digital News Report `http://www.digitalnewsreport.org`

[2]Per the access methods provided by Facebook, some of the comments made after 12-24 hours of the original news post may have been missed, as these comments had to be scrapped together with the news and comments of the next day and could have been deleted from the data that is available for querying. However, we were routinely doing manual inspections, which assured that these lost comments, if occurring, were just a tiny fraction of the total comments that were obtained. To start with, we saw that over 90% of the comments are always issued on the same day the news is posted.

```
{
  "data": {
    "id": "string",
    "created_time": "timestamp",
    "description": "string",
    "message": "string",
    "link": "string",
    "from": { "name": "string",  "id": "string" },
    "insights": { "post_clicks": "integer", "post_impressions_unique": "integer" },
    "reactions": { "love": "integer", "wow": "integer", "haha": "integer", "like": "integer" }
  },
  "article": {
    "publish_date": "timestamp",
    "text": "string"
  },
  "comments": [
    {
      "id": "string",
      "message": "string",
      "from": { "name": "string", "id": "string" },
      "reactions": { "love": "integer", "wow": "integer",  "haha": "integer", "like": "integer" },
      "replies": [
          {
              "id": "string",
              "message": "string",
              "from": { "name": "string", "id": "string" },
              "reactions": { "love": "integer", "wow": "integer", "haha": "integer", "like": "integer" }
          }
      ]
    }
  ]
}
```


Figure 4.1. Schema of the data collected from each post.

for BERT. From BETO, our next task is to produce a high-quality dataset of manually-classified Facebook comments to fine-tune it and adapt it to the task of classifying incivility.

Our training dataset consists of a constructed week, containing 17,000 comments and replies, sampled from our own Facebook data. We assembled a group of three trained coders, and they were given the task of classifying each comment and reply against the following categories proposed by Chen (2017). For examples of comments of each category, see Table 4.1.

(i) Whether the text contained name-calling or not.

(ii) Whether the text contained rudeness or not, such as vulgar language.

(iii) Whether the text contained stereotypes that denigrate gender, ethnic or sexual minorities.

Table 4.1. Examples of Facebook comments categorized as name-calling, rudeness, and stereotypes.

Category	Examples
Name-calling	1. *"Ahí se nota tu falta de cerebro"*
	2. *"Y qué dijo Chanchelet?"*
	3. *"Eres un imbécil"*
Rudeness	1. *"A tu hermana le meten varios goles por semana"*
	2. *"Esa Bachelet es una yeta de mierda"*
Stereotypes	1. *"A tu mujer hay que agrandarle la cocina para que esté contenta"*
	2. *"Cuidado que es mapuche – no te vaya a quemar la casa"*
	3. *"No esperaba menos de una comunista en todo caso. Lo raro sería que propusiera trabajar más"*

It is worth mentioning that the data is heavily unbalanced for machine-learning standards, but it is consistent with the known distribution of incivility from previous studies (Coe et al., 2014): about 74% of the comments did not have any category, and only 3% contained two or three incivility forms. For this reason, when training BERT, although we implemented a multi-label model, the final task was simplified to just classifying comments as uncivil whether they satisfied any of the three categories above or not uncivil. For a more detailed explanation of the model and the experiments made, see Appendix A.

As is common with machine-learning methods, the constructed week for training the classifier was divided into three sets: train, evaluate, and test (72%, 13%, and 15%, respectively). The division was stratified, as our goal was to preserve the same proportions of examples in each class as observed in the original dataset. The results of our classifier are quite promising and are shown in Table 4.2.

Table 4.2. Classifier results.

Accuracy	Balanced Accuracy	Precision (uncivil)	Recall (uncivil)	F1-score (uncivil)	Precision (civil)	Recall (civil)	F1-score (civil)
92.00%	88.39%	87.34%	80.90%	84.00%	93.48%	95.89%	94.67%

Combining a multi-label model with binary classification gave outstanding results, compared to a binary model and classifier. Notably, we achieved an F1 score of 84.00%. The sensitivity was 80.90% (i.e., 80.90% of all uncivil comments were classified as such), and the specificity was 95.89% (i.e., 95.89% of all civil comments were classified as such), which sums up to a 92.00% of accuracy, or 88.39% of balanced accuracy.

4.3. Network of interactions

From the data, we also constructed a multi-directed, labeled network of all the interactions generated through the posts to observe the dynamic of them. We define an interaction from a user A and a user B as when there is a direct message from A to B (A mentions B in the comment or reply) or when A made the first reply to B's comment, similar to how engagement is defined in Coe et al. (2014).

Every Facebook user that posted a comment in *Radio Bío-Bío* is a node of this network. For each interaction from user A to user B in a piece N of news, we added an edge from A to B with label N. The database was made more manageable by removing those users

with less than 5 comments through all the days of data. This strategy of discarding non-relevant data is also used by Shugars and Beauchamp (2019). We generated a network of 283 thousand users, 26329 articles, and 1433104 comments and replies with those rules. We remark that there may be multiple edges between users, representing multiple interactions within the same post and in different posts and dates. A reduced example can be seen in Figure 4.2.

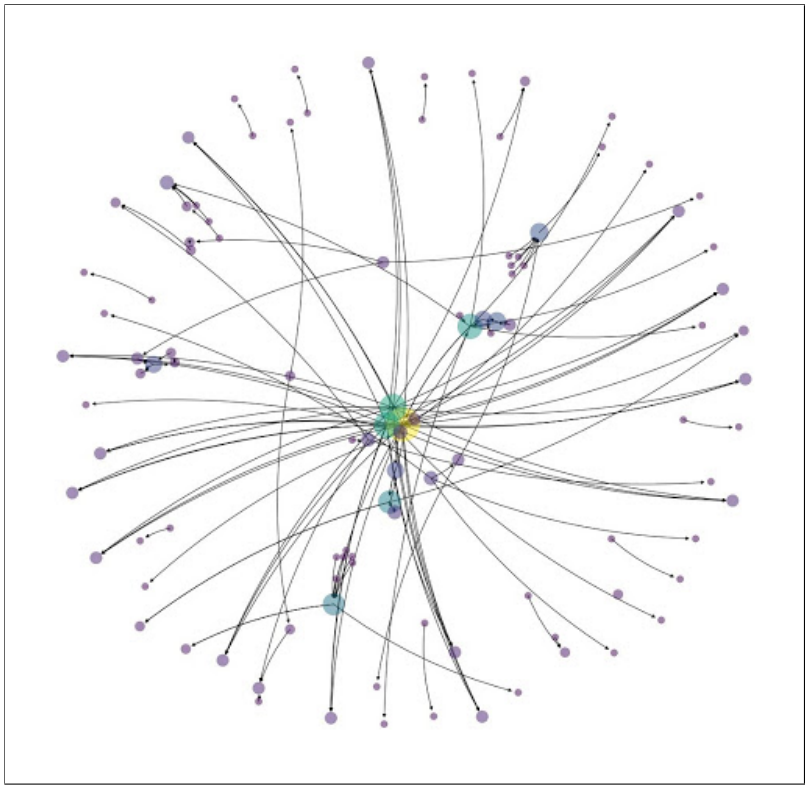

Figure 4.2. The network of conversations of a specific post. On the image, each point represents a user who is participating in discussions. The point's size and color represent the degree centrality of each user, which is the fraction of nodes it is connected to. The bigger the point, the more connected the user is to others. Each line represents a direct message from one user to another.

4.4. Analyzing the data

Our analysis treats each piece of news as a different object. For each such piece of news, we computed several indicators measuring both incivility and several statistics associated with the sub-network of interactions given only by comments to this piece of news. The statistics for each news are then aggregated either across all news or divided by topics. Let us dig deeper into these statistics and how to aggregate them.

4.4.1. Incivility

Our measure of interest is the ratio of uncivil comments: the number of uncivil comments divided by the total comments posted for that piece of news, which, as explained, is then further aggregated across several sets of news.

However, we do not use the raw ratios as computed by our classifier but instead look to correct them, accounting for the false positives and negatives of the classifier. We use a simple bayesian correction commonly used in medical contexts when dealing with exams with false positive and false negative rates (See, e.g., Havers et al., 2020). The correction is based on the following equation, which relates the observed ratio Robs of the incivility in the comments of a news piece, with the corrected ratio R, by means of the sensitivity (Sens) and specificity (Spec) of the classifier.

$$Robs = R \cdot Sens + (1 - R) \cdot (1 - Spec)$$

$$Robs = R \cdot 0.809 + (1 - R) \cdot (1 - 0.959)$$

With this correction, the incivility rate augmented in most cases, which is in line with the results according to Table 4.2: the probability that our classifiers assigns the label "civil" to an uncivil comment is greater than the probability that our classifier assigns the label "uncivil" to a civil comment. Furthermore, this correction can only be applied to the news in which Robs is at least 0.041 and at most 0.809 (the false-negative rate). News

with a lesser observed ratio were corrected by setting the ratio to 0, and news with a greater observed ratio were corrected by setting the ratio to 1.

4.4.2. The network of interactions

For each news post, we extract the sub-network given only by those edges that are labeled with this post, representing only those interactions occurring within that post. Next, we compute a series of measures associated with that sub-network, which are then aggregated and averaged just as we do for the incivility ratio. The measures extracted are listed below (see Table 4.3 for further details). Note that some metrics were obtained from network-based approaches for understanding dynamics in medical contexts (Kannampallil, Awadalla, Jones, & Abraham, 2019).

- First, we extract the proportion of nodes (users) that participate in an interaction with at least one more user in a news post against the total number of users posting a comment or reply in that news post. We denote this measure as rate nodes, as seen in Table 4.3. We also extract the proportion of those comments and replies in the news post representing an interaction with another user against the total number of comments and replies of that post, which we denote by rate edges (see again Table 4.3). These two measures serve to understand users' engagement in interactions within that post; the higher rate nodes is, the more likely it is for users commenting in that post to direct this comment to another user. Likewise, higher rate edges measures indicate that more comments were aimed at other users. We used ratios instead of a raw number of users and nodes as it is a more conservative measure and has been used before by Ksiazek et al. (2016).
- Next, we compute the average degree of nodes and the largest strongly connected component's size. These measures are standard in social network analysis and give us a way to quantify how connected is the network of interactions of given news.

- Finally, we measure the number of loops of lengths 2, 3, 4 and the sum of the three present in the graph, which corresponds to measures L2, L3, L4, and L2+. In our network, loops represent conversations, that is, bidirectional interactions in which the original node A posting an interaction to a node B is then subject of an interaction with B (for the case of L2), or subject of an interaction with a node C that is also connected to B via a chain of interactions, see Figure 4.3.

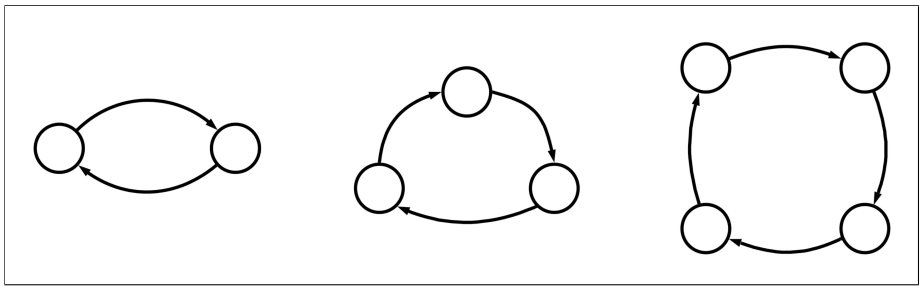

Figure 4.3. Loops between two, three, and four users.

Table 4.3. Measures used for the network analysis.

	Network measure	Definition	Description of the measure
General Measures	Rate Nodes	Vertex points in the network over all the possible points.	Represents the fraction of users that are involved in conversations for each news.
	Rate Edges	Connections between nodes in the network over all possible interactions.	Represents the fraction of comments and replies that are involved in conversations between users for each news.
Connectivity Measures	Average Degree	The average number of in and out edges per node in the network.	Represents the level of connectivity of the users of the network for each news. It measures engagement in conversations, how many messages are sent and received per user.
	Largest Strongly Connected Component	The number of nodes in the largest connected sub-network, where all nodes are connected to each other.	Represents the "core" users involved in discussions per news.
Recurrence-related Measures	Loop L2	The number of loops between two nodes. A loop L2 is defined as when the two nodes have an edge over the other. A graphic example can be seen in Figure 4.3.	Represents the number of direct discussions between two users in a post. Each user tags the other on a comment or reply.
	Loop L3	The number of loops with the same origin and end nodes, with two intermediary nodes. A graphic example can be seen in Figure 4.3.	Represents the number of direct discussions between three users. Each user tags the following on a comment or reply.
	Loop L4	The number of loops with the same origin and ending nodes, with three intermediary nodes. A graphic example can be seen in Figure 4.3.	Represents the number of direct discussions between four users, where each user tags the following on a comment or reply.
	Loop L2+	Sum of L2, L3, and L4 loops.	Represents the number of direct discussions between two, three, or four users in a post.

4.4.3. Aggregating over sets of news

As we mentioned, all of our incivility and network measures are then aggregated over-all news in our dataset to compute general statistics. But we also aggregate news in different ways, mostly per news topics, but also by date.

Given the sheer amount of news we collected, assigning topics manually to each of them is out of the question, so we have to resort to automatic labeling once again. However, this time, we departed from a fully automatic topic-modeling label, such as, e.g., Deep Pyramid Convolutional Neural Networks for Text Categorization proposed by Johnson and Zhang (2017) because such automatic labels are often uncorrelated with human-oriented topics such as health, sports, politics, etc. Instead, we defined several categories and assigned them to news based on a boolean combination of occurrences of specific keywords or phrases, which were fine-tuned with further manual inspection. It is important to empathize that keywords were selected before the measures were computed, so there was no bias in selecting specific keywords to get a desired result. Table 4.4 contains a list of the topics analyzed and a summary of the main keywords sought after. Details (in Spanish) are found in Appendix B.

Table 4.4. Keywords for news classification.

Topic	Keywords
Indigenous and native people	*mapuche, mapudungun, lonco, machi, aymara, kaweskar, yagan, quechua, diaguita, kolla*
Venezuela	venezuela, venezuelan, maduro
Political figures	politicians, piñera, bachelet, chadwick, mañalich, hutt, vallejos, lavín, rubilar, jiles, ossandón, kast, jackson, boric, jadue, matthei, sichel, briones, presidential campaign, cabinet, overflows
Military/Police	armed forces, police, soldier, army, military, air force
Immigration	immigrant, immigration, migrants, migration, refugees
Haiti	haiti, haitian
Education	education, *SIMCE*, *MINEDUC*, student, teacher, school, *JUNJI*, preschool, *PSU*, *DEMRE*, basic education, secondary education, municipal schools, subsidized private, university selection, council of rectors, private universities, traditional universities
Social Security	*ISAPRE*, *FONASA*, *AFP*, pensions
Sexual abuse	sexual assault, was raped, sexual harassment, rape, sexual abuse, consent
Crime	crime, drug trafficking, narco, illegal possession of weapons, arms trafficking
Taxes	*SII*, vat, tax, treasury, tributary, income
Women	feminism, feminine, feminist, *8m*, *lastesis*, gender roles, international women's day, free abortion, women's ministry, three causes, rapist on your way
LGBTQ	sexual minorities, gender identity, LGBT, LGBTQ, LGBTQ+, lesbian, transgender, cisgender, gay, homosexual, transvestite, bisexual, queer
Science	science, STEM, laboratory, astronomer, astronomy, *ANID*, *CONICYT*, biology, computing, scientist, scientific, mathematician, mathematics, physicist, chemical, chemistry
Health	ministry of health, nutrition, vaccine, antibiotic, wellness, cancer, doctor's office, medicines, pharmacy, mental health
Neighboring countries	brazil, argentina, peru, bolivia, colombia, ecuador, uruguay, paraguay
Sports	football, footballer, sport, sportsman, athlete, athletics, marathon, triathlon, tennis, swimming, rugby, hockey, basketball, basketball, volleyball, handball, cycling, fencing, gymnast, skiing, olympic games

For each such topic, news posts are then divided into two sets: those that belong to the topic and those that do not belong. We compute measures for each news post in each group, aggregate them over those groups, and analyze each of these measures' distribution. Thus, apart from the average of the measures for a given news topic, we could also compare their distribution against the measures for each other news in the dataset, allowing us to question the relative impact of the news topic in the behavior of comments and replies. For example, we can answer questions such as whether "posts that contain words related to immigration generate a higher rate of incivility". Whenever measures were different, statistical significance was tested through a Welch's t-test, using a significance level of $\alpha = 0.05$.

4.4.4. Other divisions

Apart from news topics, we dig deeper into the incivility ratio of news in a specific timeframe (October 2019 - December 2019) corresponding to a period of intense political turmoil in Chile. We also provide a comparison of the gender of the commentators. Once again, the statistical difference of the measures was tested using Welch's t-test.

5. RESULTS

5.1. Overall incivility and engagement

RQ1 aimed to answer how much incivility is present in online news discussions and how engaged users are in them. Overall results on incivility showed that out of the more than 4.5 million comments and replies, 25.68% of them are uncivil. Analyzing by post, 71.59% of them include at least one uncivil message, and on average, 24.51% of the messages per post are uncivil. If we look at the length of the messages, the average is 82.68 characters. On uncivil comments, the average rises to 98.39 characters, and on the opposite, the average length of civil comments decreases to 77.25 characters. The variation of 27.37% ($p = 0.0$) confirms that lengthier comments have a bigger probability of being uncivil. This result matches the findings of Chen et al. (2020), as they present that although longer comments can make room to arguments with deliberative attributes, it also gives space to change mid-comment and become uncivil.

If we look at the users who comment on the posts, on average, each user posted 7.31 messages through all year of data under analysis. However, the distribution is uneven. We have a few users that concentrate most of the messages, where the user that most posted through our study has 3217 messages, and almost 50% of the users only posted one time. Finally, on average, users posted 4.06 uncivil comments and 6.22 civil comments.

Overall results on engagement showed that 31.64% of online discussions' messages represent direct mentions to other users. Analyzing by post, 50.13% of users tag or reply to other users in their messages, and 35.82% of all the messages per post represent interactions. The willingness to participate in discussions can be measured by the L2+ metric, which for posts is, on average, 4.95 loops.

Finally, there is a greater incivility rate in messages representing interactions, as it is, on average, 28.63%.

5.2. Incivility in article topics

Next, to test H1, we focus on the distribution of uncivil comments when news are aggregated into topics. There were, in total, 27668 articles and 17 topics. Table 5.1 shows the results on the percentage of incivility present on each news topic analyzed. As expected, the level of incivility varies from topic to topic. We also note that topics with higher incivility in our study differ from previous work (Coe et al., 2014), probably because the local culture influences this distribution.

The topic with the higher difference is indigenous and native people, with an increase of 34,19% (p = 8.82e-59) concerning the rest of the news. Other hard news topics such as sexual abuse, political figures, and military/police also have higher incivility rates in the discussions. On the contrary, soft topics such as sports science and health have lower percentages of incivility. Therefore, H1 is supported. According to Coe et al. (2014), sports was the topic with the highest incivility, contrary to what we found. This highlights that some uncivil topics are different between country and country, suggesting that internal factors influence incivility.

The topic of immigration also has a higher level of incivility (p = 1.37e-10). Still, when we look for news about different countries from which migrants in Chile come from, we see that the incivility is quite uneven within these groups. To be more precise, the topic neighboring countries does not attract more incivility (p = 0.89), but the topics Venezuela and Haiti do (p = 6.17e-47 and p = 0.03 respectively). This agrees with the fact that these two countries are among those who contribute most to immigration: For the year 2019, Venezuela was the one with the most immigrants (30.5)% and Haiti the third (12.5%)[1].

[1]See Estimation of foreign people habitual residents in Chile as of December 31, 2019, `https://www.ine.cl/`

Table 5.1. Incivility by article topic.

Topic	Mean Incivility	Number of comments	Number of articles
Indigenous and native people	32.60%*	168203	720
Sexual abuse	31.17%*	74198	552
Venezuela	29.84%*	252762	1172
Political figures	29.33%*	1128082	4970
Military/Police	28.98%*	1151751	5605
Crime	28.68%*	92843	683
Immigration	28.15%*	96164	640
Taxes	27.82%*	133997	861
Haiti	26.90%*	27108	137
Education	25.55%*	234176	1360
Social Security	25.47%*	202498	1259
Women	24.55%	148205	758
Neighboring countries	24.48%	458900	3768
LGBTQ	24.23%	106941	420
Sports	23.45%*	307808	4066
Science	22.96%*	299963	2109
Health	22.61%*	193112	1209
All topics	24.51%	4513231	27668

Note. The detailed results by topic are found in Appendix C.
$*p < .05$

5.3. Engagement in article topics

To answer RQ2, we study how engagement varies across different article topics. We relied mostly on three network measures: L2+, rate nodes, and rate edges, but we use the remaining measures to complement our analysis. We focus mostly on these measures because they are the most convincing metrics for determining engagement. On the one hand, the number of loops present in the discussions, measured through the L2+ metric, represents the amount of multi-way discussions held in the comments. Thus more discussions naturally correlate with engagement. And on the other hand, rate nodes and rate edges also represent more engagement, as they constitute the rate of users participating in discussions and the rate of direct messages inside the post, respectively. Thus, we will say that a certain news topic exhibits high engagement whenever L2+ or both rate nodes and rate edges are statistically higher than the average news. We examined the whole database against the previously defined topics, as shown in Table 5.2.

Our results follow different findings to those of incivility and news topics. Hard subjects like indigenous and native people, LGBTQ, and Venezuela have the largest amount of loops of 2,3 or 4 users. About rate edges, Venezuela and indigenous and native people have lower rates, which means that although there were fewer direct messages, loops increased. Also, for these three topics, average degree and largest strongly connected component augmented, which means users are more connected to others (there are more in and out edges per node in the network) and that there are more "core" users involved in the network discussing with others. Science is also a topic with more engagement, as rate nodes and rate edges are 13.12% (p = 2.10e-5) and 10.51% (p = 5.23e-3) respectively bigger in contrast to non-science news. These results begin to suggest that higher engagement is not always accompanied by higher incivility as Science is a topic with lower incivility.

Some topics do not raise more engagement in discussions, such as neighboring countries and sports. Also, for both of these topics, the average degree and largest strongly connected component are lower compared to news not containing these topics, which

Table 5.2. Engagement by article topic.

Topic	Mean L2+	Mean Rate nodes	Mean Rate edges	Number of comments	Number of articles
Indigenous and native people	9.31*	0.51	0.39*	60940	702
LGBTQ	8.81*	0.57	0.43	36160	408
Venezuela	7.99*	0.48*	0.38*	87731	1134
Haiti	7.90*	0.51	0.41	10447	130
Women	7.57*	0.55	0.41	52786	732
Political figures	7.18*	0.43*	0.33*	335428	4762
Military/Police	7.17*	0.47*	0.36	379163	5413
Education	5.91*	0.53*	0.38*	82897	1312
Social Security	5.72*	0.49	0.37	65934	1193
Immigration	5.54	0.56*	0.41*	36243	631
Science	4.65	0.56*	0.39*	109062	2019
Health	5.24	0.52	0.37	67070	1164
Crime	4.87	0.55	0.41	32612	651
Taxes	4.64	0.48	0.37	39284	806
Sexual abuse	5.50	0.46*	0.33*	26049	516
Neighboring countries	3.89*	0.47*	0.34*	142537	3478
Sports	2.29*	0.50	0.34*	97046	3736
All topics	4.95	0.50	0.36	1433104	26329

Note. The detailed results by topic are found in Appendix C.
$*p < .05$

strengthens the idea that the engagement is less. For these topics, lower engagement goes in hand with lower incivility.

Lastly, other hard news topics like crime, taxes, and sexual abuse neither raise more engagement. This suggests that engagement behaves differently than incivility, and there is no direct association between engagement and news topics (RQ2).

5.4. Incivility and engagement patterns

As one of our main results, we observe no direct relation between incivility and engagement in discussions for different news topics (RQ3). On the contrary, our results indicate that some higher incivility topics show high engagement, while other topics show low engagement, and the same holds for topics with low incivility. Thus, we can divide news topics into four different categories, or quadrants, according to users' behavior commenting news on those topics. Figure 5.1 depicts the four quadrants; here, the X-axis represents incivility, and the Y-axis represents engagement.

The first quadrant stands for higher incivility and higher engagement. News topics in this quadrant represent topics for which users are more propense to replying, tagging, or involving in two-way discussions. At the same time, comments feature higher rates of incivility. Some topics of this category are indigenous and native people, Venezuela, social security, and military/police.

The second quadrant corresponds to lower incivility, and higher engagement, representing that users are more propense to replying and tagging, but not in an uncivilized way. An example is LGBTQ, where there is no statistical difference in this topic's incivility rate concerning the others, but there is more engagement. Science and women are also in this category.

Figure 5.1. Quadrants of behavior. Each topic was given a point of the form (incivility, engagement). The points were determined first if there was a statistical difference and then by each variable's specific value. A comprehensive explanation of the algorithm used to place each topic in their respective quadrant is found in Appendix D.

The third quadrant points for lower incivility and lower engagement. News topics in this quadrant represent the less heated conversations, as users dialogue in a civilized way and do not tend to reply to others. Examples are health, sports, and neighboring countries.

The final quadrant represents higher incivility and lower engagement, which speaks for messages more likely to be uncivil, but they do not trigger discussions between users. Crime, sexual abuse, and taxes are in this category. Note that topics in this quadrant include various types of crimes.

5.5. Other

Although this investigation focuses on the most part on understanding incivility and engagement through different news topics, we did a minimal exploration of gender and sexual abuse topic and on specific social events. About gender, our user database is divided into 52.27% men, 44.12% women, and 3.61% users without identified gender. Regardless of comments, men concentrate 60.60% of messages, while women 35.15% and users without identified gender 4.26%. These proportions coincide with the literature, which states that women are less likely to enter a discussion (Van Duyn et al., 2019). Nevertheless, this difference in participation does not translate into large incivility differences, as results above all news showed that men are only 2.86% (p = 1.01e-22) more uncivil than women, having rates of 36.73% and 35.74% respectively. Distributions in Figures 5.2 and 5.3 show that there are almost no differences between them. Regarding sexual abuse, there are no statistical differences in incivility rate between men and women (p = 0.74), and the distributions between them follow the same pattern (similar to Figures 5.2 and 5.3). In short, these results suggest that gender is not a determinant contextual factor in incivility rates.

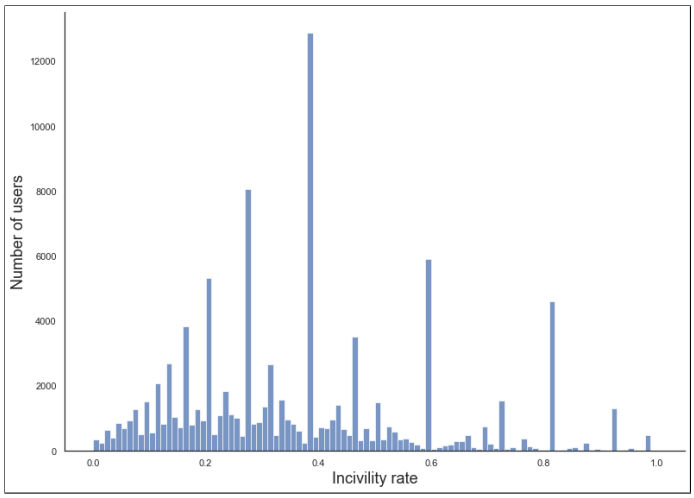

Figure 5.2. Male distribution of incivility rate per user

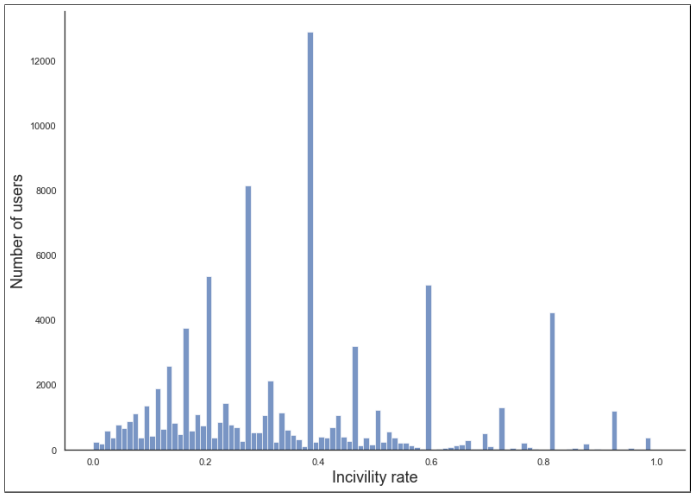

Figure 5.3. Female distribution of incivility rate per user

Finally, about specific time frames, our data is within a complicated period in Chile, as a social outbreak occurred against the political class on October the 18th of 2019. In general terms, incivility augmented 6.55% during the social outbreak (p = 3.90e-6). But when we analyzed the political figures topic, surprisingly, the incivility rate did not augment for this period as expected, and even decreased by 5.50% (p = 0.003). What augmented was the number of posts and messages, given that in a period of time of approximately one third compared to the pre-outbreak, a very similar amount of civil and uncivil interactions were generated. These results begin to suggest that incivility is not triggered by specific social milestones. On the contrary, it is something that is already part of daily discussions on social networks.

6. DISCUSSION

Incivility has been studied several times, but to our understanding, only a few times with the amount of data that we have presented throughout our investigation. With the use of a machine learning technique for natural language processing, we could train a model that could learn to classify incivility automatically. This model allowed us to support our findings by millions of comments, a task that would have been impossible with only expert coders. However, there is a trade-off as we lose precision when relying on automatic annotation. One example is that all incivility forms were grouped and studied as one, as the number of comments per category was not good enough for automatic classification.

Another strength of our study is that we used expert coders for our training data and did not rely on mechanical turkers. Our method of coding the comments strengthens the results obtained, as there is no unintended bias in the dataset product of differences in perceptions (Sadeque et al., 2019). For example, Theocharis et al. (2020) used an online platform to label their data.

About the classifier, our first attempts were made with traditional techniques like SVMs and linear regressions, and we encountered similar results as Theocharis et al. (2020) and Stoll et al. (2019). These results were not sufficient as the model failed to classify a significant number of civil and uncivil comments correctly. There are no direct baselines for us to use since there are no studies of automatic classification of incivility in Spanish. Nevertheless, to our understanding, we present higher scores among all existing investigations. We achieved an F1-score of 0.84, 4.73% more than the state-of-the-art classifier for detecting incivility (Davidson et al., 2020). We are in knowledge that achieving similar results in a multi-label classifier is challenging due to the imbalance of the classes and the finer-grained task (Sadeque et al., 2019). Comparing our results on this task (go to Appendix A, Tables A.2 and A.3) to the ones of Sadeque et al. (2019), our multi-label classifier outperforms theirs, as we achieved an F1-score of 0.61 (26.90% higher) for name-calling and an F1-score of 0.63 (19.39% higher) for vulgarity or rudeness.

Our findings on incivility allow us to compare this phenomenon with results from other cultures. We had anticipated that uncivil comments were about one in every five (Coe et al., 2014). However, on average we found that one in every four comments were uncivil (25.68%) for our local context. Also, Coe et al. (2014) found that 55.5% of the articles included at least one uncivil comment, which in our case that number raised to 71,59%. Finally, Stoll et al. (2019) reported 16.6% of uncivil comments on Facebook sites of German media. This suggests that incivility varies among different contexts and cultures.

Name-calling was the prevalent form of incivility in our training dataset. It was present in 20% of all comments. These results support the findings of Coe et al. (2014), insisting on the idea that if participants end to engage in demeaning attacks directed to an individual or group, incivility in discussions would be considerably less.

Topic analysis for incivility raised differences from the ones presented by Coe et al. (2014). While some topics prevailed as the most uncivil in both studies (politics, crime, and taxes), others like sports, immigration, and social security behaved differently. This reinforces our finding that incivility not only varies by topic but also by different cultural contexts. A topic that draws our attention is indigenous and native people, as it is the subject with the highest incivility rate. It is important to remark that Chile is in the midst of an upsurge in violent conflicts related to indigenous –specifically, Mapuche– claims, which surely permeates the uncivil reaction we witness. In fact, if we further filter out the indigenous and native people category to contain news only related to this specific conflict, the incivility is even higher. Thus, what we see is probably a mix of incivility generated by a hard political topic and the predominance of incivility in comments related to race, as presented by Chen et al. (2020), but further research is needed to understand further why this is the most uncivil topic in this Chilean news outlet.

Finally, the incivility rates reported for political figures are similar to the findings of Saldaña and Rosenberg (2020), where the ratio is higher than 28% in both cases. Yet, Global North studies reported less incivility in this topic (Coe et al., 2014; Theocharis et

al., 2020). This is consistent with the assertion that cultural problems influence the civility of discussions, as Chile presents low levels of political trust (Saldaña & Rosenberg, 2020).

Regardless of engagement, we rely on network theory to propose more informative and fine-grained metrics that go beyond general measures as suggested by Ksiazek et al. (2016). With this approach, we support the findings of their study. Factors like the article's topic affect engagement, but not associated to the softness or hardness of the news topic. Higher engagement is not always associated with soft news topics (social security, military/police, political figures, and indigenous and native people reported high engagement levels). Nor is it associated only with hard news, since science also presented high levels of engagement. Our method allowed us to examine each post in their exhaustive conversational context as proposed by Su et al. (2018), since we studied comments collectively and not isolated (measured through L2, L3, L4, and L2+).

The high levels of engagement for the topics indigenous and native people, immigration, Haiti, and Venezuela agree with the findings of Shugars and Beauchamp (2019), as racism draws more engagement.

As far as we are aware, our proposal to divide user behavior into four quadrants, dividing both the incivility and engagement dimensions, represents the first framework to understand how incivility and engagement behave together. Notably, our findings differ from the statement that comments with greater engagement are more civil (Coe et al., 2014), as we showed that some news topics might exhibit more engagement but less incivility, on average.

Moreover, we also found topics with high incivility but low engagement (see crime, taxes, and sexual abuse topics). To our understanding, these findings suggest that the established notion that troll behavior engages other users in the same conduct (Cheng et al., 2017) deserves a more detailed look: a manual investigation of news comments in these categories make us think that the incivility here is mostly pointed at third parties

such as the perpetrator of the crimes, or the establishment, and thus it is also important to understand who is being trolled by users.

Finally, we can affirm that negative discussion context increases the engagement and probability of a user trolling back (Cheng et al., 2017) for some topics like immigration and indigenous and native people. For other topics like LGBTQ, women, and science, we find what could be called a more productive mode of engagement, where users take part in conversations civilly (Shugars & Beauchamp, 2019).

7. CONCLUSIONS AND FUTURE WORK

To summarize our results, first, we present a machine learning model that can predict whether a comment is uncivil or not with a high F1-score. Our model presents the higher scores among all existing studies in the field (Davidson et al., 2020; Sadeque et al., 2019; Maity et al., 2018).

Second, we find that incivility varies among different news topics, but this pattern also depends on the local context on which users are discussing. Topics like sports presented the higher incivility rate in the Global North (Coe et al., 2014)), wherein contrast, our investigation presented this topic as one with the lowest incivility rate among all.

Third, engagement results follow similar findings from previous research but with finer metrics. Higher engagement is not always associated with soft news or hard news topics (Ksiazek et al., 2016). Also, we provide a novel way of measuring engagement, as we use network theory to propose more informative and specific metrics.

Finally, we propose a novel framework to understand how incivility and engagement behave together and discard the idea that comments with greater engagement are more civil (Coe et al., 2014). We find topics for each of the defined quadrants, which allows us to get a raw idea of how news topics affect the phenomena studied.

About limitations, we present similar obstacles as in Coe et al. (2014), since both studies were restricted to comments of only one news forum. Future work should consider extending this investigation to a wide range of sources, as other studies have proven that incivility rates vary across different types of news outlets (Su et al., 2018). It is also important to build a more robust method for classifying news into topics. Our method is acceptable, but it has room for improvement.

Future research should also consider improving the automatic classification of incivility. Although our results were outstanding among previous research, it remains the debt of studying incivility in each of its forms, and not in a grouped way as we had to do. To

achieve that, one possibility is to have more labeled data for the training process. Other possibilities include trying other BERT adaptations like DistilBERT, which presented better results than BERT for Davidson et al. (2020).

Also, it would be interesting to dig deeper into gender analysis. Our minimal findings suggest that gender is not a determinant factor in incivility rates. However, Pierson (2015) presents that women and men tend to comment on different topics, where comments from women are more present on forums related to parenting, fashion, and health. Also, Van Duyn et al. (2019) exhibit that women are more likely to engage in discussions of local matters (against state, national, or international news). This realization may suggest that engagement and rates of incivility may be different among different topics and gender. Future work should explore this area.

Finally, our network of interactions presents the duality of also being a signed network. These graphs are characterized because the edges have annotations about the positive or negative connotation of the interaction (Bonchi, Galimberti, Gionis, Ordozgoiti, & Ruffo, 2019). As each message in our network is labeled as civil or uncivil, the connotation can be given by this value. Bonchi et al. (2019)) presented a novel algorithm for detecting polarized communities in signed networks. This algorithm could be used to study the polarization present in each post of our dataset.

REFERENCES

Antoci, A., Delfino, A., Paglieri, F., Panebianco, F., & Sabatini, F. (2016). Civility vs. incivility in online social interactions: An evolutionary approach. *PloS one*, *11*, e0164286. doi: 10.1371/journal.pone.0164286

Bonchi, F., Galimberti, E., Gionis, A., Ordozgoiti, B., & Ruffo, G. (2019). Discovering polarized communities in signed networks. In *Proceedings of the 28th acm international conference on information and knowledge management* (pp. 961–970). doi: 10.1145/3357384.3357977

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. In *Pml4dc at iclr 2020.*

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk.* Austin TX: Palgrave Macmillan. doi: 10.1007/978-3-319-56273-5

Chen, G. M., Fadnis, D., & Whipple, K. (2020). Can we talk about race? exploring online comments about race-related shootings. *Howard Journal of Communications*, *31*(1), 35–49. doi: 10.1080/10646175.2019.1590256

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing* (pp. 1217–1230). doi: 10.1145/2998181.2998213

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. doi: 10.1111/jcom.12104

Davidson, S., Sun, Q., & Wojcieszak, M. (2020). Developing a new classifier for automated identification of incivility in social media. In *Proceedings of the fourth workshop on online abuse and harms* (pp. 95–101). doi: 10.18653/v1/2020.alw-1.12

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York.

Havers, F., Reed, C., Lim, T., Montgomery, J., Klena, J., Hall, A., ... Thornburg, N. (2020). Seroprevalence of antibodies to sars-cov-2 in 10 sites in the united states, march 23-may 12, 2020. *JAMA Internal Medicine*, *180*(12), 1576–1586. doi: 10.1001/jamainternmed.2020.4130

Humprecht, E., Hellmueller, L., & Lischka, J. A. (2020). Hostile emotions in news comments: A cross-national analysis of facebook discussions. *Social Media + Society*. doi: 10.1177/2056305120912481

Hwang, H., Kim, Y., & Huh, C. U. (2014). Seeing is believing: Effects of uncivil online debate on political polarization and expectations of deliberation. *Journal of Broadcasting & Electronic Media*, *58*(4), 621–633. doi: 10.1080/08838151.2014.966365

Johnson, R., & Zhang, T. (2017). Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 562–570).

Kannampallil, T., Awadalla, S. S., Jones, S., & Abraham, J. (2019). A graph-based approach for characterizing resident and nurse handoff conversations. *Journal of biomedical informatics*, *94*, 103178.

Kim, Y., & Herring, S. C. (2018). Is politeness catalytic and contagious? effects on participation in online news discussions. *Proceedings of the 51st Hawaii International*

Conference on System Sciences, 1955—1964. doi: 10.24251/HICSS.2018.247

Ksiazek, T. B., Peer, L., & Lessard, K. (2016). User engagement with online news: Conceptualizing interactivity and exploring the relationship between online news videos and user comments. *New media & society*, *18*(3), 502–520. doi: 10.1177/1461444814545073

Lysak, S., Cremedas, M., & Wolf, J. (2012). Facebook and twitter in the newsroom: How and why local television news is getting social with viewers? *Electronic News*, *6*(4), 187–207. doi: 10.1177/1931243112466095

Maity, S. K., Chakraborty, A., Goyal, P., & Mukherjee, A. (2018). Opinion conflicts: An effective route to detect incivility in twitter. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–27. doi: 10.1145/3274386

Pierson, E. (2015). Outnumbered but well-spoken: Female commenters in the new york times. In *Proceedings of the 18th acm conference on computer supported cooperative work & social computing* (pp. 1201–1213).

Rains, S. A., Kenski, K., Coe, K., & Harwood, J. (2017). Incivility and political identity on the internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication*, *22*(4), 163–178. doi: 10.1111/jcc4.12191

Rajaraman, A., & Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.

Rosenberg, A. (2018). ¿Qué me estás queriendo decir? Un nuevo acercamiento metodológico para entender la incivilidad de usuarios en comentarios de noticias online. *Index Comunicación*, *8*(3), 87–104.

Rossini, P., & Maia, R. C. M. (in press). Characterizing Disagreement in Online Political Talk: Examining Incivility and Opinion Expression on News Websites and Facebook in Brazil. *Journal of Public Deliberation.*.

Sadeque, F., Rains, S., Shmargad, Y., Kenski, K., Coe, K., & Bethard, S. (2019). Incivility detection in online comments. In *Proceedings of the eighth joint conference on lexical and computational semantics (* SEM 2019)* (pp. 283–291). doi: 10.18653/v1/S19-1031

Saldaña, M., & Rosenberg, A. (2020). I don't want you to be my president! incivility and media bias during the presidential election in chile. *Social Media+ Society*. doi: 10.1177/2056305120969891

Shugars, S., & Beauchamp, N. (2019). Why keep arguing? predicting engagement in political conversations online. *Sage Open*. doi: 10.1177/2158244019828850

Stoll, A., Ziegele, M., & Quiring, O. (2019). Detecting incivility and impoliteness in online discussions. classification approaches for german user comments.
doi: 10.31235/osf.io/a47ch

Su, L. Y.-F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., & Brossard, D. (2018). Uncivil and personal? comparing patterns of incivility in comments on the facebook pages of news outlets. *New Media & Society*, *20*(10), 3678–3699. doi: 10.1177/1461444818757205

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on twitter. *SAGE Open*. doi: 10.1177/2158244020919447

Van Duyn, E., Peacock, C., & Stroud, N. J. (2019). The gender gap in online news comment sections. *Social Science Computer Review*. doi: 10.1177/0894439319864876

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37–52. doi: 10.1016/0169-7439(87)80084-9

APPENDIX

A. BERT BASED CLASSIFICATION MODEL

BERT is a language representation model that has achieved state-of-the-art results on different NLP tasks such as question answering and sentiment classification (Devlin et al., 2019). Its architecture consists of a multi-layer bidirectional transformer encoder with an attention-based approach. The framework provides two steps: pre-training and fine-tuning. The model is trained on unlabeled data over two pre-training tasks: next sentence prediction and masked language modeling. For fine-tuning, the model is initialized with the pre-trained parameters, and the other parameters are fine-tuned with the label data of the specific supervised downstream task (Devlin et al., 2019).

Using BERT, we experimented with multi-label classification, binary classification, and multi-label model combined with binary classification. The main difference between binary and multi-label classification is that instead of one out feature in the final layer of linear classification, we have tree out features with sigmoid activation function.

The experiments were conducted with different BERT models: BERT multilingual cased, BERT multilingual cased fine-tuned with our unlabeled data, and BETO. The last one is a model trained over BERT for 2 million steps with a Spanish corpus by Cañete et al. (2020). To deal with the classes' unbalance, we stratified the data set to ensure that each subset of training, evaluation, and testing had an even distribution of each category. We also gave the model weights to assign each label for loss calculation based on the number of occurrences. Finally, we fine-tuned the hyperparameters over manual inspection, using as a guideline the range of values proposed by Devlin et al. (2019). The best results for the experiments are presented above.

A.1. Multi-label classifier

For each model, the best results were achieved with a sequence length of 256, a learning rate of 3e-05, and 4 epochs.

Table A.1. Global results of the multi-label classifier.

Exp.	Model	Batch size	LRAP	Hamming Loss
1	BERT multilingual	16	0.97	0.79
2	BERT multilingual fine-tuned	16	0.97	0.74
3	**BETO**	**8**	**0.98**	**0.81**

Table A.2. Name-calling results.

Exp.	F1- score (uncivil)	Precision (uncivil)	Recall (uncivil)	F1- score (civil)	Precision (civil)	Recall (civil)	Balanced accuracy
1	0.58	0.58	0.58	0.89	0.90	0.89	0.74
2	0.46	0.46	0.49	0.86	0.87	0.86	0.67
3	**0.61**	**0.64**	**0.59**	**0.91**	**0.90**	**0.92**	**0.75**

Table A.3. Rudeness results results.

Exp.	F1- score (uncivil)	Precision (uncivil)	Recall (uncivil)	F1- score (civil)	Precision (civil)	Recall (civil)	Balanced accuracy
1	0.63	0.67	0.60	0.97	0.97	0.98	0.79
2	0.42	0.48	0.37	0.96	0.96	0.97	0.67
3	**0.63**	**0.61**	**0.65**	**0.97**	**0.98**	**0.97**	**0.81**

Table A.4. Stereotype results.

Exp.	F1- score (uncivil)	Precision (uncivil)	Recall (uncivil)	F1- score (civil)	Precision (civil)	Recall (civil)	Balanced accuracy
1	**0.24**	**0.34**	**0.19**	**0.98**	**0.98**	**0.99**	**0.59**
2	0.18	0.39	0.12	0.99	0.98	1.00	0.56
3	0.21	0.33	0.15	0.98	0.98	0.99	0.56

Overall, BETO gives the best results. About recall, for all categories, only a fraction of the pertaining comments were classified as such. These results are not enough for our task since it would introduce a significant error to the data. Finally, Table A.4 shows that stereotype is the form of incivility most hard to label. This is because only 2.8% of the dataset has this category.

A.2. Binary classifier

For these experiments, we only used the model BETO based on the previous results. We trained for 4 epochs with a learning rate of 3e-5. The results are in Table A.5. As expected, we got better results than the multi-label classifier. However, the sensitivity is not yet enough, as only 66% of all uncivil comments are classified as such.

Table A.5. Results of the binary classifier.

Batch size	Sequence length	Accuracy	Balanced accuracy	Precision	Recall
8	256	0.83	0.77	0.70	0.64
8	512	0.84	0.76	0.71	0.61
32	**256**	**0.83**	**0.78**	**0.67**	**0.66**

A.3. Multi-label model with a binary classifier

We used the models previously trained on the multi-label classifier experiments. With this modification, we achieved outstanding results with the BETO model, which we ended up using for the following tasks. Results are shown in Table A.6.

Table A.6. Results of the multi-label model with a binary classifier.

Model	Accuracy	Balanced accuracy	Precision	Recall
BERT multilingual	0.84	0.78	0.67	0.69
BERT multilingual fine-tuned	0.85	0.80	0.70	0.71
BETO	**0.92**	**0.88**	**0.81**	**0.87**

B. TOPIC KEYWORDS IN SPANISH

Table B.1. Spanish keywords for news classification.

Topic	Keywords
Indigenous and native people	*mapuche, mapudungun, lonco, machi, aymara, kaweskar, yagan, quechua, diaguita, kolla*
Venezuela	*venezuela, venezolano, maduro*
Political figures	*políticos, piñera, bachelet, chadwick, mañalich, hutt, vallejos, lavín, rubilar, jiles, ossandón, kast, jackson, boric, jadue, matthei, sichel, briones, campaña presidencial, gabinete, desbordes*
Military/Police	*fuerzas armadas, carabineros, militar, ejercito, militares, fuerza aérea*
Immigration	*inmigrante, inmigracion, migrantes, migración, refugiados*
Haiti	*haiti, haitiano*
Education	*educación, SIMCE, MINEDUC, alumno, alumna, profesor, escolar, JUNJI, preescolar, PSU, DEMRE, enseñanza básica, enseñanza media, colegios municipales, particular subvencionado, selección universitaria, consejo de rectores, universidades privadas, universidades tradicionales*
Social Security	*ISAPRE, FONASA, AFP, pensiones*
Sexual abuse	*agresión sexual, fue violada, acoso sexual, violación sexual, abuso sexual, consentimiento*
Crime	*crimen, narcotrafico, narco, tenencia ilegal de armas, tráfico de armas*
Taxes	*SII, IVA, impuesto, fisco, tributario, renta*
Women	*feminismo, femenino, feminista, 8m, lastesis, roles de género, día internacional de la mujer, aborto libre, ministerio de la mujer, tres causales, violador en tu camino*
LGBTQ	*minorías sexuales, identidad de género, LGBT, LGBTQ, LGBTQ+, lesbiana, transgenero, cisgenero, gay, homosexual, travesti, bisexual, queer*
Science	*ciencias, STEM, laboratorio, astrónomo, astronomía, ANID, CONICYT, biología, computación, científico, científica, matemático, matemáticas, físico, químico, química*
Health	*ministerio de salud, minsal, nutrición, vacuna, antibiótico, bienestar, cáncer, consultorio, medicamentos, farmacia, salud mental*
Neighboring countries	*brazil, argentina, perú, bolivia, colombia, ecuador, uruguay, paraguay*
Sports	*futbol, futbolista, deporte, deportista, atleta, atletismo, maratón, triatlón, tenis, natación, rugby, hockey, basquetbol, baloncesto, voleibol, balonmano, ciclismo, esgrima, gimnasta, esquí, juegos olímpicos*

C. ALL RESULTS FOR EACH NEWS TOPIC

Table C.1. Results for news related to indigenous and native people.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.33*	0.24	34.19%	8.82e-59
Rate Nodes	0.51	0.50	1.44%	4.48e-01
Rate Edges	0.39*	0.36	9.49%	1.47e-05
Average Degree	1.08*	0.89	21.33%	1.13e-43
Largest Strongly Connected Component	4.31*	2.65	62.76%	1.31e-29
Loop L2	8.77*	4.61	90.36%	9.90e-27
Loop L3	0.38*	0.16	139.05%	1.28e-10
Loop L4	0.16*	0.07	136.0%	2.98e-05
Loop L2+	9.31*	4.83	92.61%	8.06e-27

$*p < .05$

Table C.2. Results for news related to immigration from Venezuela.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.3*	0.24	22.93%	6.17e-47
Rate Nodes	0.48*	0.50	-4.05%	1.81e-02
Rate Edges	0.38*	0.36	5.96%	1.21e-02
Average Degree	1.08*	0.89	20.98%	3.43e-56
Largest Strongly Connected Component	3.97*	2.64	50.61%	8.01e-33
Loop L2	7.49*	4.59	63.04%	2.21e-28
Loop L3	0.34*	0.16	115.5%	5.35e-14
Loop L4	0.16*	0.07	141.18%	1.41e-07
Loop L2+	7.99*	4.82	65.83%	5.03e-29

$*p < .05$

Table C.3. Results for news related to political figures.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.29*	0.23	25.03%	2.63e-161
Rate Nodes	0.43*	0.52	-16.54%	1.53e-31
Rate Edges	0.33*	0.36	-8.84%	1.76e-06
Average Degree	1.01*	0.88	14.84%	4.03e-108
Largest Strongly Connected Component	3.49*	2.52	38.57%	2.50e-72
Loop L2	6.79*	4.26	59.58%	4.75e-66
Loop L3	0.26*	0.14	82.87%	9.43e-26
Loop L4	0.12*	0.06	110.48%	8.05e-14
Loop L2+	7.18*	4.46	60.99%	2.86e-66

$*p < .05$

Table C.4. Results for news related to military/police.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.29*	0.23	23.95%	7.36e-155
Rate Nodes	0.47*	0.51	-7.54%	1.33e-11
Rate Edges	0.36	0.36	-0.81%	5.60e-01
Average Degree	1.0*	0.87	15.05%	3.01e-115
Largest Strongly Connected Component	3.45*	2.50	37.97%	1.47e-82
Loop L2	6.76*	4.19	61.61%	2.22e-72
Loop L3	0.28*	0.14	103.51%	5.17e-39
Loop L4	0.13*	0.05	138.68%	1.21e-16
Loop L2+	7.17*	4.38	63.87%	2.98e-74

$*p < .05$

Table C.5. Results for news related to immigration.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.28*	0.24	15.23%	1.37e-10
Rate Nodes	0.56*	0.50	11.08%	1.47e-02
Rate Edges	0.41*	0.36	16.07%	9.54e-03
Average Degree	0.97*	0.90	8.66%	9.17e-08
Largest Strongly Connected Component	3.07*	2.68	14.45%	8.58e-04
Loop L2	5.2	4.70	10.53%	9.77e-02
Loop L3	0.22*	0.16	35.56%	1.04e-02
Loop L4	0.11	0.07	64.57%	6.35e-02
Loop L2+	5.54	4.94	12.12%	6.31e-02

$*p < .05$

Table C.6. Results for news related to immigration from Haiti.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.27*	0.25	9.8%	2.62e-02
Rate Nodes	0.51	0.50	1.56%	7.68e-01
Rate Edges	0.41	0.36	14.07%	7.08e-02
Average Degree	1.08*	0.90	20.11%	8.19e-07
Largest Strongly Connected Component	3.9*	2.69	45.16%	6.93e-05
Loop L2	7.48*	4.70	59.17%	4.68e-04
Loop L3	0.29*	0.17	76.5%	4.09e-02
Loop L4	0.12	0.07	77.95%	1.41e-01
Loop L2+	7.9*	4.94	60.01%	4.29e-04

$*p < .05$

Table C.7. Results for news related to education.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.26*	0.24	4.47%	8.60e-03
Rate Nodes	0.53*	0.50	5.57%	3.41e-03
Rate Edges	0.38*	0.36	5.12%	1.86e-02
Average Degree	0.93*	0.90	3.42%	2.26e-03
Largest Strongly Connected Component	3.04*	2.67	13.68%	1.17e-04
Loop L2	5.6*	4.67	19.9%	1.08e-04
Loop L3	0.21*	0.16	26.34%	1.73e-02
Loop L4	0.11*	0.07	59.37%	7.94e-03
Loop L2+	5.91*	4.90	20.66%	8.02e-05

$*p < .05$

Table C.8. Results for news related to social security.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.25*	0.24	4.09%	1.56e-02
Rate Nodes	0.49	0.50	-1.54%	5.41e-01
Rate Edges	0.37	0.36	3.71%	2.56e-01
Average Degree	0.97*	0.90	8.05%	4.28e-11
Largest Strongly Connected Component	3.13*	2.67	17.04%	8.27e-06
Loop L2	5.41*	4.68	15.42%	2.76e-03
Loop L3	0.21*	0.16	27.63%	1.87e-02
Loop L4	0.11*	0.07	58.64%	1.54e-02
Loop L2+	5.72*	4.92	16.43%	2.03e-03

$*p < .05$

Table C.9. Results for news related to sexual abuse.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.31*	0.24	27.89%	1.68e-21
Rate Nodes	0.46*	0.50	-8.64%	2.58e-04
Rate Edges	0.33*	0.36	-6.62%	1.97e-02
Average Degree	0.94*	0.90	4.71%	7.75e-03
Largest Strongly Connected Component	3.22*	2.68	20.08%	1.76e-03
Loop L2	05.09	4.71	8.17%	2.94e-01
Loop L3	0.26*	0.16	55.56%	9.97e-03
Loop L4	0.15*	0.07	117.0%	9.73e-03
Loop L2+	5.5	4.94	11.24%	1.72e-01

$*p < .05$

Table C.10. Results for news related to crime.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.29*	0.24	17.52%	3.98e-13
Rate Nodes	0.55	0.50	9.05%	3.28e-01
Rate Edges	0.41	0.36	15.71%	2.14e-01
Average Degree	0.92	0.90	2.73%	7.36e-02
Largest Strongly Connected Component	2.75	2.69	2.34%	5.21e-01
Loop L2	4.65	4.72	-1.54%	7.93e-01
Loop L3	0.15	0.17	-8.72%	4.21e-01
Loop L4	0.07	0.07	1.82%	9.34e-01
Loop L2+	4.87	4.95	-1.74%	7.69e-01

$*p < .05$

Table C.11. Results for news related to taxes.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.28*	0.24	13.98%	2.09e-11
Rate Nodes	0.48	0.50	-4.54%	1.97e-01
Rate Edges	0.37	0.36	3.77%	5.76e-01
Average Degree	0.93*	0.90	3.51%	1.28e-02
Largest Strongly Connected Component	2.7	2.69	0.17%	9.55e-01
Loop L2	4.44	4.72	-5.94%	2.46e-01
Loop L3	0.15	0.17	-10.74%	3.09e-01
Loop L4	0.05*	0.07	-30.97%	4.32e-02
Loop L2+	4.64	4.96	-6.45%	2.10e-01

$*p < .05$

Table C.12. Results for news related to women.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.25	0.25	0.14%	9.42e-01
Rate Nodes	0.55	0.50	10.18%	1.28e-01
Rate Edges	0.41	0.36	14.79%	9.56e-02
Average Degree	0.99*	0.90	10.75%	1.43e-11
Largest Strongly Connected Component	3.74*	2.66	40.36%	1.04e-10
Loop L2	7.07*	4.65	51.98%	1.82e-10
Loop L3	0.32*	0.16	100.18%	1.74e-07
Loop L4	0.18*	0.07	172.16%	3.88e-05
Loop L2+	7.57*	4.88	55.21%	6.78e-11

$*p < .05$

Table C.13. Results for news related to LGBTQ.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.24	0.25	-1.17%	6.34e-01
Rate Nodes	0.57	0.50	14.41%	2.63e-01
Rate Edges	0.43	0.36	20.76%	2.28e-01
Average Degree	0.99*	0.90	10.02%	8.65e-08
Largest Strongly Connected Component	3.8*	2.68	42.09%	3.92e-09
Loop L2	8.31*	4.66	78.37%	2.55e-09
Loop L3	0.34*	0.16	105.28%	4.76e-06
Loop L4	0.16*	0.07	141.72%	2.28e-03
Loop L2+	8.81*	4.89	80.15%	1.55e-09

$*p < .05$

Table C.14. Results for news related to science.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.23*	0.25	-6.83%	3.43e-07
Rate Nodes	0.56*	0.50	13.21%	2.08e-05
Rate Edges	0.39*	0.36	10.51%	5.23e-03
Average Degree	0.88*	0.90	-2.71%	7.53e-04
Largest Strongly Connected Component	2.58*	2.70	-4.48%	2.69e-02
Loop L2	4.42	4.74	-6.72%	7.56e-02
Loop L3	0.15	0.17	-7.58%	3.13e-01
Loop L4	0.08	0.07	9.98%	5.09e-01
Loop L2+	4.65	4.98	-6.52%	8.70e-02

$*p < .05$

Table C.15. Results for news related to health.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.23*	0.25	-8.08%	1.37e-06
Rate Nodes	0.52	0.50	4.18%	6.56e-02
Rate Edges	0.37	0.36	4.58%	7.57e-02
Average Degree	0.93*	0.90	4.08%	1.28e-03
Largest Strongly Connected Component	2.88*	2.68	7.15%	2.06e-02
Loop L2	4.95	4.71	5.15%	2.78e-01
Loop L3	0.2	0.16	21.01%	5.67e-02
Loop L4	0.09	0.07	35.75%	2.47e-01
Loop L2+	5.24	4.94	6.1%	2.10e-01

$*p < .05$

Table C.16. Results for news related to neighboring countries.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.24	0.25	-0.15%	8.87e-01
Rate Nodes	0.47*	0.51	-6.36%	3.93e-05
Rate Edges	0.34*	0.36	-7.06%	1.46e-04
Average Degree	0.88*	0.90	-2.43%	3.36e-04
Largest Strongly Connected Component	2.5*	2.72	-8.09%	6.55e-06
Loop L2	3.69*	4.87	-24.26%	8.81e-21
Loop L3	0.15*	0.17	-13.73%	2.11e-02
Loop L4	0.05*	0.07	-31.32%	6.14e-04
Loop L2+	3.89*	5.11	-24.02%	1.16e-19

$*p < .05$

Table C.17. Results for news related to sports.

Measure	Mean (related news)	Mean (not related news)	Variation	P-value
Incivility	0.23*	0.25	-5.02%	2.00e-06
Rate Nodes	0.5	0.50	-0.16%	9.27e-01
Rate Edges	0.34*	0.36	-5.41%	1.09e-02
Average Degree	0.81*	0.91	-11.37%	2.53e-72
Largest Strongly Connected Component	1.96*	2.81	-30.34%	9.29e-138
Loop L2	2.2*	5.13	-57.04%	2.05e-190
Loop L3	0.07*	0.18	-63.68%	1.80e-54
Loop L4	0.02*	0.08	-71.6%	6.28e-29
Loop L2+	2.29*	5.39	-57.48%	7.91e-189

$*p < .05$

D. QUADRANT CLASSIFICATION ALGORITHM

To classify the topics on a specific quadrant, we gave each of them a pair of values (*x*, *y*). *X* represents the incivility and *y* the engagement. For incivility, we used the mean incivility of the topic and whether it represented a statistical difference or not. We interpolated the incivility mean in ranges [0, 1] if the p-value was lower than 0.05 (which means there is a statistical difference in distributions) and the variation of incivility from news without the topic was greater than zero. In any other case, we interpolated the mean incivility in ranges [-1, 0]. Algorithm 1 shows a pseudo-code.

Algorithm 1: Incivility x coordinate

```
if (incivility_p_value < 0.05) and (incivility_variation > 0) then
    interpolate_function = Interpolate1D([min_incivility, max_incivility], [0,
      1]);
    x = interpolate_function(incivility_mean);
else
    interpolate_function = Interpolate1D([min_incivility, max_incivility], [-1,
      0]);
    x = interpolate_function(incivility_mean);
end if
return x
```

Calculating the *y* value for engagement was not direct because we had three metrics instead of one (L2+, rate nodes, and rate edges). The first thing needed was to reduce the dimensionality of these three metrics. For that chore, we used principal component analysis (PCA) (Wold, Esbensen, & Geladi, 1987), a technique commonly applied in machine learning problems for different tasks, one of them being to help the visualization of data in fewer dimensions. The first step was to scale the three metrics onto a unit scale (mean = 0 and variance = 1). After that, we used the PCA algorithm on our scaled data, which consists of applying an orthogonal linear transformation, where the greatest variance between the data is sought. Note that after reduction, there is no particular meaning assigned

to this new value. The result was a value that represents one possible pooled measure of engagement for our specific data points. Algorithm 2 shows a pseudo-code.

Algorithm 2: Engagement principal component

```
scaled_values = StandardScaler(l2+_values, rate_nodes_values,
  rate_edges_values);
pca = PCA(n_components = 1);
principal_component_values = pca.transform(scaled_values);
return principal_component_values
```

With the principal component value of engagement for each topic, we applied a similar algorithm to incivility. We interpolated these values in ranges [0, 1] if one of these three conditions were met:

(i) L2+ p-value was lower to 0.05, and the variation of L2+ from news without the topic was greater than zero.

(ii) Rate nodes p-value was lower to 0.05, and the variation of rate nodes from news without the topic was greater than zero.

(iii) Rate edges p-value was lower to 0.05, and the variation of rate edges from news without the topic was greater than zero.

In any other case, we interpolated the principal component value in ranges [-1, 0]. Algorithm 3 shows a pseudo-code.

Algorithm 3: Engagement y coordinate

```
1 if (l2+_p_value < 0.05 and l2+_variation > 0) or (nodes_p_value < 0.05 and
    nodes_variation > 0) or (edges_p_value < 0.05 and edges_variation > 0)
    then
2 |   interpolate_function = Interpolate1D([min_engagement, max_engagement],
  |     [0, 1]);
3 |   y = interpolate_function(engagement_principal_component);
4 else
5 |   interpolate_function = Interpolate1D([min_engagement, max_engagement],
  |     [-1, 0]);
6 |   y = interpolate_function(engagement_principal_component);
7 end if
8 return y
```