Data Visualisation for Supporting Linguists in the Analysis of Toxic Messages

Ecem Kavaz[1], Anna Puig[1], Inmaculada Rodriguez[1], Mariona Taule[2], and Montserrat Nofre[1]

[1]Department of Mathematics and Computer Science, University of Barcelona
[2]Department of Catalan Philology and General Linguistics, University of Barcelona
Barcelona, Spain
ekavazka27@alumnes.ub.edu

Keywords: Data visualisation, Corpus Annotation, Toxicity, Hate Speech

Abstract: The goal of this research is to provide linguists with visualisations for analysing the results of their hate speech annotation. These visualisations consist of a set of interactive graphs for analysing the global distribution of annotated messages, finding relationships between features, and detecting inconsistencies in the annotation. We used a corpus that includes 1,262 comments posted in response to different Spanish online new articles. The comments were annotated with features such as sarcasm, mockery, insult, improper language, constructivity and argumentation, as well as with level of toxicity ('not-toxic', 'mildly toxic', 'toxic' or 'very toxic'). We evaluated the selected visualisations with users to assess the graphs' comprehensibility, interpretability and attractiveness. One of the lessons learned from the study is the usefulness of mixed visualisations that include simple graphs (Bar, Heat map) - to facilitate the familiarisation with the results of the annotated corpus together with more complex ones (Sankey, Spider or Chord) - to explore and identify relationships between features and to find inconsistencies.

1 INTRODUCTION

Social media have become a powerful tool for many people for self-expression as they can share their voices and opinions freely even anonymously if desired. These platforms let people to use their freedom of speech very actively and effortlessly from the comfort of their homes and at any time (Paschalides et al., 2019). As the use of social media has increased, the multimedia data available has also increased, which provides researchers with greater opportunities to examine these data. On a daily basis, millions of messages are created and shared on different online platforms (Chen et al., 2017). The news comment section offered by some online newspapers is one of the possible spaces in which readers can express their opinions, although sometimes these opinions can be conveyed in an aggressive, offensive or inappropriate manner, especially when they are given anonymously or under a false name. This offensive, abusive or toxic language can be labelled as hate speech. It can be simply described as a kind of speech that attacks a person or a group based on characteristics such as race, religion, ethnic origin, national origin, gender, disability, sexual orientation, or gender identity (Gagliardone et al., 2015). In this context, frameworks and tools for automatically classifying messages are becoming ever more essential for detecting the trends and spreading patterns that will help to identify anomalous behaviour and hate speech (Florio et al., 2020).

In recent years, methods for the automatic classification of hate speech messages have been widely studied in different social networks and areas (Paschalides et al., 2019), (Grosman et al., 2020). The quality of these frameworks depends greatly on the algorithms used in NLP (Natural Language Processing), but also on having access to a sufficiently large corpus of annotated messages in the training steps of these algorithms (Frénay and Verleysen, 2013). This training dataset usually consists of messages (including tweets and comments), which are manually annotated by humans. Indeed, the quality of this man-

DOI:10.24132/CSRN.2021.3101.7

ISBN 978-80-86943-34-3

ual annotation is a key point to ensure the success of the whole process. Annotation involves processing a large number of messages and tends to become a difficult and time-consuming task plagued by errors and inconsistencies. Linguists usually follow a well-controlled methodology in which a single message is annotated by several annotators (preferably experts). Afterwards, agreement must reached for all the annotations. Detecting errors, trends and inconsistencies efficiently in the individual and the agreed-upon annotations can be helpful to speed up and to guarantee the quality and reliability of the final annotation.

Considering the aforementioned aspects, which make annotation a complex and challenging task, data visualisation can be a helpful method that allows linguists to analyse the results. Viewing data as mere numbers conveys little meaning, whereas data visualisation helps people to process information more easily(Knaflic, 2015). Well-designed interactive data visualisations can appeal to people effortlessly (Wu et al., 2016). However, the design of the most suitable data visualisation in a particular context is not an easy task.

The goal of this research is to provide annotators with a set of visualisations for analysing the results of their hate speech annotation. We use the NewsComTOX corpus, which consists of comments posted in response to different Spanish online news articles annotated with toxicity. Concretely, we contribute with: (1) a set of interactive graphs to allow the annotators to see the global distribution of comments, find relationships between features and detect possible errors and inconsistencies in the annotation, and (2) the lessons learned from a preliminary user evaluation of the proposed visualisations to detect the most useful graphs for linguists.

2 RELATED WORK

In this section we first present research works aimed at using data visualisation for the monitoring of automatic annotation systems. We place the focus on the visualisation techniques used by the authors. We then consider works that are aimed at supporting linguistic annotators through meaningful visualisations.

2.1 Visualisations for monitoring automatic annotation systems

Visual analytics for automatic annotation systems aims to identify valuable information in social data. Concretely, the following research works analyse anomalous user behaviour, anomalous information spread and the use of toxic language. (Shi et al., 2019) carried out a survey on the visual analytics of anomalous user behaviours. The survey revealed four types of user behaviours, including social interaction, travel, network communication and transactions. For each of the four types of user behaviours, the authors analysed trends in common data types, anomaly detection techniques, visualisation techniques and interaction methods. Our research is focused on the first type of user behaviour, social interaction, i.e. the communication of ideas and opinions between people in online newspapers.

Fluxflow (Zhao et al., 2014) is an interactive data visualisation system designed to display and evaluate anomalous information spread (rumours and misinformation) on Twitter. The novel visualisation design consisted of packed circles (retweets) arranged along a timeline showing how an original message spreads among people over time. The size of the circles symbolised the power of the influence of a user and the colour represented an anomaly score. Similarly to FluxFlow, Episogram (Cao et al., 2015) was designed to analyse retweeting behaviours on Twitter. It showed the activity of each person separately and every message from each person separately in the form of single lines on a timeline. Nevertheless, this visualisation caused cluttering, making it hard to understand at first sight.

RumorLens (Resnick et al., 2014) was created to help journalists to detect the diffusion of rumours on online social media and, once again, Twitter was the source of data for this research. The authors used a Sankey diagram to effectively summarise the diffusion of a rumour since it makes it very simple to follow and see people's decisions regarding posting or not posting the rumour.

Mandola (Paschalides et al., 2019) was designed with NLP and ML techniques to monitor, and detect online-hate speech on online social media. The Mandola dashboard included the so called Hate-map and Hotspot Map visualisations, both showing findings on a world map. While Hate-Map displays hate data with heat spots in certain countries, Hotspot Maps have a colour scale representing six levels of hate speech. There was also a Heat map that displays hate speech in five topic areas, separated by years. Mandola is close to our research because of its analysis of hate speech. The difference resides in that our goal is to use visualisations to support linguists in the annotation process, whereas Mandola aimed to use visualisations for the monitoring of an automatic annotation system.

Overall, previous research focused on novel approaches for monitoring automatic annotation systems using different visualisation techniques, such as Sankey diagrams and Heat maps. In this paper, we use these and other types of graphs to help linguists to visualise the results of their annotations.

2.2 Visualisations for supporting linguist annotators

Several tools and platforms (i.e. set of tools) are available to support the task of annotating a corpus. All of them provide basic functionality for data annotation. Nevertheless, some of them do not support more advanced functionalities, such as the management of the inter-annotator agreement. The inter-annotator agreement is a measure of how well two (or more) annotators can make the same annotation decision for a certain feature. This measure may impact the quality and efficacy of the annotation process.

For instance, Brat (Stenetorp et al., 2012) is a mainstream annotation tool that does not allow for several annotators, with a consequent non-support of inter-annotator agreement. Another tool is MAT, which supports the annotation and the management of multiple annotators through a web interface (MAT, 2020). IBM Watson Knowledge Studio, which is integrated in the well-known Watson platform, includes an annotation tool, for creating a training corpus that is well designed and documented (Watson, 2020).

The above-mentioned tools provide support to the annotators but they do not use data visualisations. In contrast, Eras (Grosman et al., 2020) and WebAnno (Yimam et al., 2013), use data visualisation to show, for example, the results of annotators agreement through a Heat map. The visualisations we present in this paper are in line with these two frameworks since we also aim to facilitate and improve annotators' work by means of new, meaningful visualisations.

3 USED DATA

In this section we describe the NewsCom-TOX corpus, the dataset used for developing the visualisations, and the annotation tagset used.

The NewsCom-TOX corpus consists of 1,262 comments posted in response to different articles extracted from Spanish online newspapers from August 2017 to May 2019 annotated with toxicity. The articles selected cover four different topics -economy, politics, religion and immigration- and the comments were selected in the same order in which they appear in the time thread in the web. Those comments that were duplicated were removed. Table 1 shows the distribution of comments per topic and the corresponding newspaper from which they were obtained.

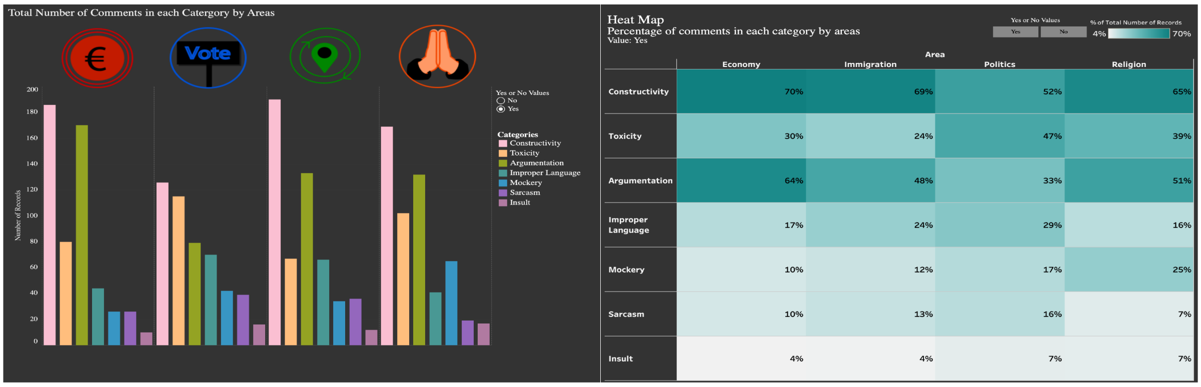
Topic	Comments	Newspaper
Economy	309	La Información, El País
Politics	239	Huffpost, La Vanguardia
Religion	298	Xataca Ciencia
Inmigration	416	El Confidencial
Total	1262	

Table 1: Distribution of comments per topic

In order to have a balanced representation of comments per topic, two different news articles were needed in the case of economy and immigration-related topics. Articles were selected to potentially lead to controversy with the aim of finding comments with opposing opinions and examples of toxic language. Toxicity is difficult to define, possibly because it can be expressed at different levels and in different ways (Ross et al., 2017), (Davidson et al., 2017), (Fortuna and Nunes, 2018). In order to reflect this diversity in the expression of toxicity, the proposal is to assign different levels of toxicity, indicating whether the comment is 'not toxic', 'mildly toxic', 'toxic' or 'very toxic'. With the aim of reducing the subjectivity in the annotation and, therefore, also the disagreement between annotators, we propose first annotating different linguistic features such as sarcasm, mockery, insult, improper language, constructivity and argumentation. These binary features allow us to discriminate the level of toxicity of the comments. Furthermore, some of these features can be correlated, for instance argumentation and constructivity, insult and improper language, and these correlations are also useful when assigning the level of toxicity. Our hypothesis is that the combination of these features helps to determine the level of toxicity in a more objective way. The tagset used for the annotation of comments with toxicity is the following:

- <**argumentation**>: indicates that the comment gives arguments or reasoned explanations or grounds opinions with evidences.
- <**constructivity**>: a comment is constructive when it is respectful and polite (regardless of whether it is in favour or against the content of the article or of another comment)[1], when it intends to create an enriching and useful dialogue, when it contributes with new knowledge, ideas and pro-

[1] We can find two types of comments, those that comment on some specific or general aspect of the article, or those that are responses to another comment.



	Area			
	Economy	Immigration	Politics	Religion
Constructivity	70%	69%	52%	65%
Toxicity	30%	24%	47%	39%
Argumentation	64%	48%	33%	51%
Improper Language	17%	24%	29%	16%
Mockery	10%	12%	17%	25%
Sarcasm	10%	13%	16%	7%
Insult	4%	4%	7%	7%

Figure 1: Visualisations of Global Data Distribution. Bar chart and Heat map.

posals and offers new perspectives and insights to approach the subject.

- <**sarcasm**>: a comment is sarcastic when the content is ironic -that is, when the writer uses words that mean the opposite of what he really wants to say- and when it is accompanied by a harsh, sharp and negative criticism and made in bad faith. Ironic comments without intention to cause pain (without a negative load) are not considered toxic and are tagged as <sarcasm=no>.
- <**mockery**>: indicates that the comment ridicules, mocks or humiliates a person or group.
- <**insult**>: indicates that the comment contains one or more insults or slurs with the intention to offend a person or group.
- <**improper language**>: indicates that the comment contains language not consider to be proper or that is vulgar and impolite and/or which includes rude words.
- <**toxicity**>: a comment is toxic when it attacks, denigrates or disqualifies a person or group on the basis of certain characteristics such as race, ethnicity, nationality, religion, gender and sexual orientation, among others. This attack can be expressed in different ways -directly (through insult, mockery and inappropriate humour) or indirectly (for instance through sarcasm)- and at different levels of intensity, that is at different levels of toxicity (the most aggressive being those comments that incite hate or even physical violence).

It should be noted that all these tags have binary values (value= yes/no) except the toxicity tag, which has four values (<1= non-toxic>; <2= mildly toxic>; <3= toxic> and <4: very toxic>). The level of toxicity is determined by the presence and combination of the features presented above. In fact, these features are different ways or mechanisms to express the toxicity and, therefore, they also help to define what is meant by toxicity. The more negative features appear in the comment, the higher the level of toxicity. For instance, we tag as 'mildly toxic' comments in which only one feature appears, the most frequent being <sarcasm>, <mockery> and <improper language>, whereas in comments tagged as <very toxic> the combination of features is higher than two, an especially frequent combination is <improper language>, <mockery> and <insult>. This annotation allows us to establish fine grained criteria for analysing and better defining what can be considered a comment with toxic language or hate speech.

4 GRAPHS TO VISUALISE TOXIC MESSAGES

We have used multiple types of graphs and diagrams to visualise the annotated data, which allow to analyse and measure these kind of data. We have used Tableau[2] (an interactive data visualisation software), DisplayR[3] (an online visualisation tool) and lastly SankeyMatic[4] (an online Sankey diagram builder) to create data visualisations.

4.1 Visualisations of global distributions.

All the visualisations are combined in a Tableau story, which offers an individual page for each visualisation. The first page of our Tableau project is an introduction including four icons representing the four topics (economy, politics, immigration and religion), a brief explanation that provides an overview of the dataset

[2] https://www.tableau.com
[3] https://www.displayr.com
[4] http://www.sankeymatic.com

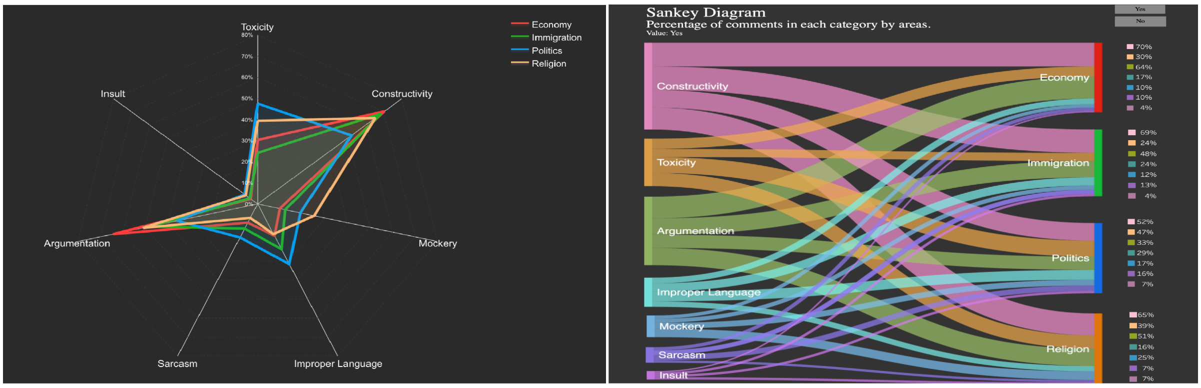


Figure 2: Visualisations of Global Data Distribution. Spider Chart and Sankey diagram.

(such as the origin of the messages and the questions to be answered using the visualisations) and the text visualisation of the seven features used in the annotation of toxicity. Each feature is represented by size, depending on the total number of ‘yes’ comments they have. On the tooltips, there is an example comment about the feature and the number of comments annotated as ’yes’ for that feature. This text visualisation is a summary of upcoming visualisations.

Created graphs are then separated into three groups: (1) visualisations to analyse the global distributions of messages by topics (2) visualisations of the relationships between features, and (3) visualisations by features.

Global data distribution visualisations aim to explore data in terms of topics and features of the messages (see Figures 1 and 2). The first graph in this section is a simple Bar chart. The Bar chart has four sections represented by icons for four topics. Features are colour coded and displayed in each topic individually. The Bar chart displays features in terms of their total numbers. There is a filter that allows users to select values (’yes’ or ’no’). Please note that every message has been annotated with tags ’yes or no’ for every feature. For example, a comment tagged with ’yes’ in the argumentation, mockery and sarcasm features, can be tagged as ’no’ in constructivity and insult Secondly, a Heat map[5] is created to show data as a whole and in a simple way without other complex design elements. The Heat map illustrates the features for each topic including a filter to choose ’yes’ or ’no’.

The Spider[6] graph is created to present the global distribution of toxic messages for each topic. The Spider web has seven sides representing the seven features that appear in the annotated messages. Topics are represented by their assigned colours. On the Spider chart, each topic has a unique shape made up of points combined together with lines. Each point shows the selected percentage value (’yes’ or ’no’) percentage of a feature and lines are used to combine the points to form the shape. In the page of the Spider graph, there is a toxicity symbol which displays the mean level of toxicity for each topic.

Another approach for examining the global distribution of data is to create a Sankey diagram[7]. In this case, features are presented on the left side, splitting to form areas on the right side. The Sankey diagram also allows users to choose ’yes’ or ’no’ values. Next to the topics, there are keys that illustrates the percentage of each feature in terms of the selected value.

4.2 Visualisations of relationships between features

Two visualisations have been created to observe relationship between features (see Figure 3). These two visualisations aim to show a comparison between the features annotated in the comments, which will mainly allow annotators/linguists to analyse their hypothesis.One of them is a Scatter plot with icons for topics, representing one feature on the x-axis and one on the y-axis. Each axis shows the possible values of the feature represented on that axis (yes-no). There is a filter where users can choose the desired feature for each axis to compare with each other. Thus, with this graph, users can examine the number of occurrences of each combination of values

(’no-no’, ’no-yes’, ’yes-no’, and ’yes-yes’) in selected features. For example, the combination ’no-no’

[5] A heat map displays data values with colour, usually by intensity or hue.

[6] A spider graph displays multivariate data with three or more quantitative variables represented on axes starting from the same point.

[7] Sankey diagram is a type of flow diagram where the width of the links are proportional to the size of the data.

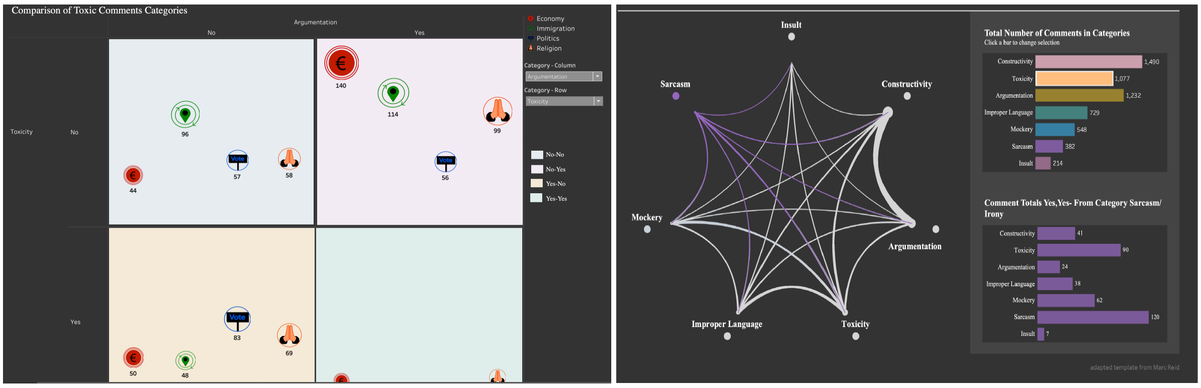


Figure 3: visualisations of relationships between features. Scatter plot and Chord diagram.

for constructivity and argumentation refers to messages tagged as ’no’ for the constructive feature and ’no’ for the argumentation feature.

The other visualisation is a Chord diagram[8] (Reid, 2020). While a Scatter plot analyses data in topics and features, a Chord diagram does not separate data into topics but rather analyses the data as a whole in the toxic comment features. The Chord diagram also has value selection buttons for ’no-no’, ’no-yes’, ’yes-no’ and ’yes-yes’ comments in features. The Chord diagram lets users select a feature and then highlights the arcs of that selected feature, which allows users to examine features individually in relation to the other features. The thickness of the arc that connects two features represents the number of messages annotated by those two features with the value selection. The total number of comments in the selected feature and others are also displayed on a separate Bar chart, which displays the total amount of comments in each feature.

4.3 Visualisations by features

One of the main objectives of this study is to measure the level of toxicity. There are four levels of toxicity, namely, non-toxic (1), mildly toxic (2), toxic (3) and very toxic (4). The aim of these graphs include, analysing annotated corpus, hypotheses and most importantly to identify inconsistencies. Although the mean of toxicity of each topic is illustrated on the Spider, and the Sankey diagram, three additional graphs are created to analyse the level of toxicity in features. Firstly, in the left part of Figure 4 there is a Bar chart in which each level of toxicity is represented with a separate bar for each feature. In this graph, the ’yes’ and ’no’ values are displayed separately, and can be selected from the list, while a tooltip displays the total number of messages at each level of toxicity.

A Sankey diagram is proposed to show the level of toxicity for each feature (see the right hand side of Figure 4). In this diagram, the data come from the level of toxicity (non-toxic to very toxic) and go into features with lines in which the thickness of the lines represents the total number of comments in each level and its connected feature. The same toxicity icon is used to show the total number of messages on the tooltip for each features at each level of toxicity. Lastly, in Figure 4 there is a Treemap in which the level of toxicity is represented for a selected feature. The Treemap is more complex as it has many components to explore and is highly interactive. There are four different maps for each topic. All levels of toxicity (1 to 4) are divided into ’yes’ and ’no’ are represented by colours blue to red. The darkest blue represents the lowest toxicity and the darkest red shows the highest toxicity. For example, if the chosen feature is constructive, we can see the number of messages at the level of toxicity 2 separately as ’yes’ or ’no’ in the economy topic.

There is another small Treemap which represents the total number of messages by topic and, when the topic is chosen, the Treemap shows the total number of messages at each level of toxicity without dividing them into ’yes’ or ’no’. There is a filter to change the range of levels. Treemap allows users to compare the levels of toxicity in terms of features and also topics. There is another Treemap that represents the total number of messages in a topic and, when the topic is chosen, the Treemap shows the total number of messages at each level without dividing them into ’yes’ or ’no’. There is a filter to change the range of levels. In this graph ’yes’ or ’no’ values are displayed separately and they can be selected from the list, while the tooltip displays the total number of messages at each level of toxicity.

[8] A chord diagram shows the connection between features. Thickness of the links represents the data size.

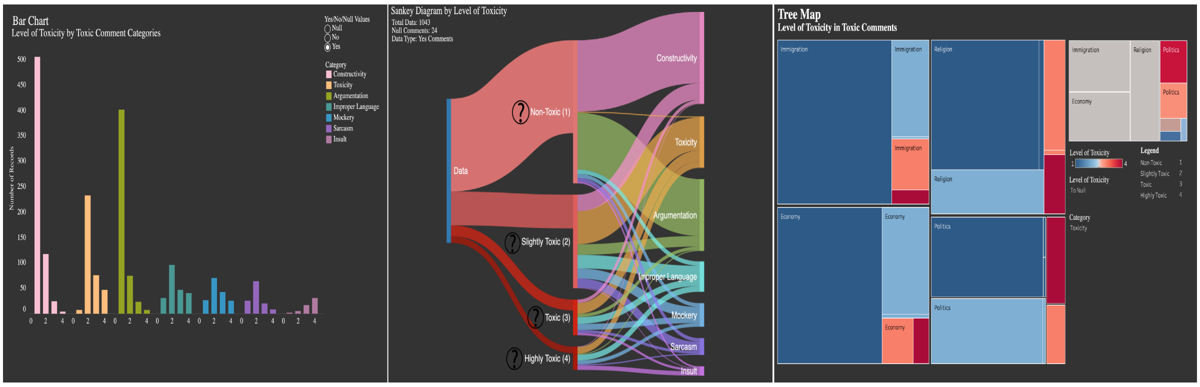


Figure 4: Visualisations by features. Left to right, Bar chart, Sankey diagram and Treemap.

5 EVALUATION

5.1 Methodology

The goal of the evaluation was to assess comprehensibility (how well the graph communicates the information to the user), interpretability (how well the user can extract meaning from the visualisation), attractiveness (to what extent the visualisation is visually appealing for the user), and to gather users' opinions.

The evaluation was exploratory, aimed at obtaining participants' perceptions. It was unmoderated, performed through an online questionnaire. A total number of eight participants were recruited, including five females and three males. The questionnaire consisted of demographic questions followed by three visualisation tasks lasting up to 30-35 minutes. Demographic questions included, gender, age, and two questions related to the participants' prior degree of experience in message annotation and visual analytics. Most of the participants had prior experience in data visualisation and more than half of the participants had expertise in message annotation.

The visualisation tasks intertwined links to our Tableau visualisations and questions related to them:

- Task 1: Please, follow the tableau link of ”visualisations of global distribution” (Bar Chart, Heat map, Spider Chart, and Sankey Diagram).
  - *Q1-Comprehensibility*: ”Score from 1 (the most negative) to 5 (to most positive) how well the graph communicates the global distribution.” Please, justify the best and the worse scores.
  - *Q2-Interpretability*: ”Score from 1 (very difficult) to 5 (very easy) how easy is to interpret the graph.”
  - *Q3-Attractiveness*: ”Score from 1 (very bad) to 5 (very good) the visual appeal of the graph.”
- Task 2: Please, follow the tableau link of ”visualisations of relationships between features” (Scatter Chart and Chord Diagram).Same questions as TASK1: Q1-Q3.
- Task 3: Please, follow the tableau link of ”visualisations by features” (Bar Chart, Sankey Diagram, and Treemap). Same questions as TASK1: Q1-Q3.

The questions were chosen to explore the three dimensions of visualisations including: the communication of information (how well the visualisations communicate), the interpretation of graphs (how easy or difficult it is to understand the graphs, and the attractiveness (how appealing the visualisations). There were nine closed-questions (i.e. score 1 to 5), three of which were followed by three open-questions (i.e. justify your answer) presented in the questionnaire. Data were collected anonymously.

5.2 Results

For each task, we first analysed answers to questions (Q1-Q3), then we presented the qualitative data arising from users' comments.

Task 1 : visualisations of global distribution

According to the results in Figure 5, the Heat map received the highest scores in *Q1-Comprehensibility*, followed by the Bar chart. The Spider graph received relatively good scores in this question, whereas the Sankey diagram was the least favourite graph. The results show that the Bar chart is the easiest to understand in *Q2-Interpretability*, with 75 percentage of participants scoring it as very easy and the Sankey diagram as the hardest to understand, with relatively low scores. The Spider chart and the Sankey diagram scored the highest in *Q3-Attractiveness*.

In the first task, the Bar chart was the most favoured graph in the dimensions of comprehensibility and attractiveness. The Bar chart was the second favourite graph in the communication of information dimension, scoring slightly lower than the Heat map. The Bar chart was described by multiple participants as very easy to understand and use. A couple of participants commented that they favoured the Bar chart as they are used to analysing this type of graph. Another positive comment was that the colours of the bars were both appealing and made the visualisation clearer. However, various participants commented that the Bar chart should have included percentages along with the actual total number of comments in the features. The Heat map received the highest scores for *Q1-Comprehensibility*. The results also show that the Heat map was easy to understand and communicates information well, however, visually it was not as attractive as the other three visualisations.

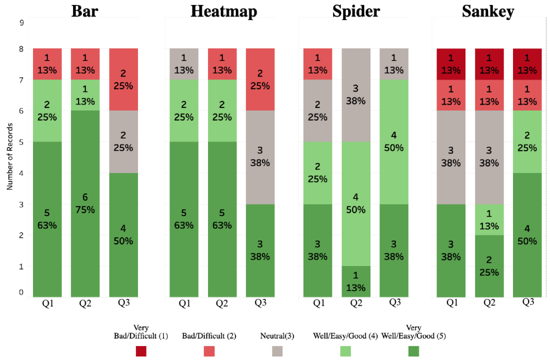


Figure 5: Stacked bar chart displaying the results in Task 1. Q1-Comprehensibility, Q2-Interpretability and Q3-Attractiveness.

The Spider graph received mixed reviews from users. One participant stated that *"I think the Spider gives a very clear idea of the distribution of attributes by features, with its isolated and superimposed surfaces, it reflects very clearly what has been annotated"*. Another participant stated that *"The Spider chart is also a good way of displaying the data for a general comparison across topics. However, the comparison among features within a specific topic is less clear when only one axis contains the percentage indicator."* Another participant agreed with this comment by stating that comparisons between the features were a little bit difficult when the values matched. The results suggested that the Spider graph was visually very appealing but slightly more difficult to understand than the Bar chart and the Heat map.

The Sankey diagram was the least favoured of all the graphs, even though it received high scores in the visual appeal section. The majority of the participants commented that the shape of the graph was confusing, difficult to understand and that it was difficult to compare features of messages and topics. One participant stated that *"The Sankey diagram seems rather chaotic compared to the other ones"*. Some participants commented that they needed to pay extra attention to the Sankey diagram because of its complexity.

Multiple participants commented that to improve communication, all of the graphs should have included both percentage values and actual total numbers. The interactivity of the graphs was an important element as results show that participants preferred the Bar chart since it was very interactive while they did not like the fact that the Sankey diagram was static. The comments suggest that participants favoured graphs on which they could spend the least possible time as they did not want to waste time trying to understand the graphs.

Task 2: visualisations of relationships between features

In the *Q1-Comprehensibility*, (see Figure 6), the Scatter plot and the Chord diagram achieved similar scores. The Scatter plot is favoured slightly more by the participants. The Scatter plot achieved higher scores in the *Q2-Interpretability* than the chord diagram, with a total of seven high scores, while the Chord diagram received only two high scores. For the *Q3-Attractiveness*, the Chord diagram was considered visually more attractive than the Scatter plot and it obtained significantly high scores.

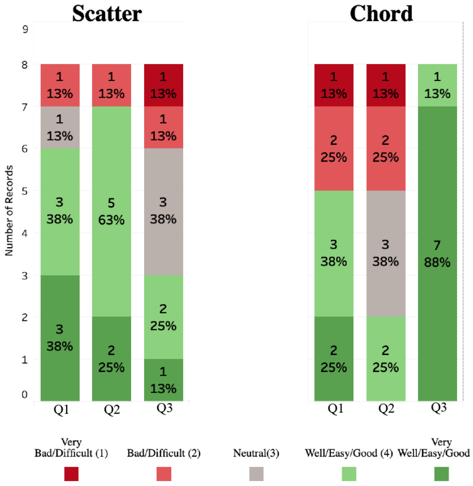


Figure 6: Stacked bar chart displaying the results in Task 2. Q1-Comprehensibility, Q2-Interpretability and Q3-Attractiveness.

In the second task, most of the participants agreed that the Scatter plot was easier to understand than the Chord diagram, and the Chord diagram was visually more appealing than the Scatter plot. There were mostly positive views about both graphs in terms of their ability to communicate the information and the Scatter plot was described as more intuitive than the Chord diagram. Participants commented that the Scatter plot provided a clear view for comparing topics and features, and that it was also very engaging thanks to its interactivity. A participant commented that *"with the Scatter plot it is easier to understand the correlation between the different features, an explanatory legend appears that also includes an example, the relationship is quickly associated with the number of examples, it displays the relationships between features according to the topic. It is more intuitive"*. On the other hand, some participants have found the Chord diagram, relatively easy to understand and useful, especially in terms of the overall view of the data. One participant stated that *"Chord Diagram helps to understand the different relationships between features very well and allows interaction by focusing on various elements and their intersections, providing very valuable information."*

Task3: visualisations by features

The Bar chart received the highest scores in the *Q1-Comprehensibility*, (see Figure 7). The Sankey diagram and the Treemap obtained similar scores. In the *Q2-Interpretability*, the Bar chart was considered to be the easiest and the Treemap the hardest to understand. As in task 1, the Sankey diagram given the highest scores in the *Q3-Attractiveness*, followed by the Bar chart which received the second highest scores in Task 3. The Treemap was the least preferred graph in this task. In the Task 3 (see Figure 7), the most preferred graph was the Bar chart in the dimensions of comprehensibility and interpretability. This is in line with other comments since the Bar chart is very commonly used, it is easier to interpret, and comparing the features with each other was therefore easier. Also analysing a Bar chart does not require prior knowledge of the visual analytics, which was another reason for the popularity of the Bar chart.

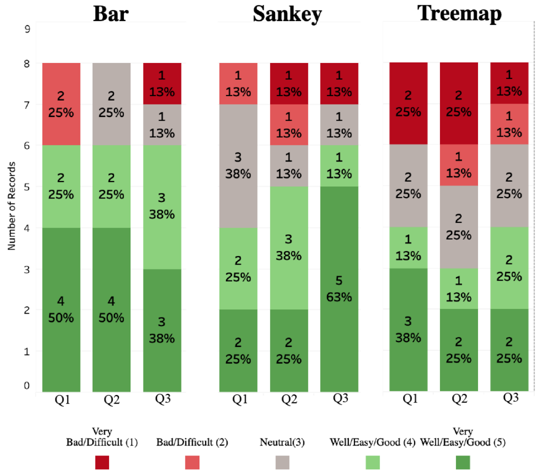


Figure 7: Stacked bar chart displaying the results in Task 3. Q1-Comprehensibility, Q2-Interpretability and Q3-Attractiveness.

The Sankey diagram had the highest score in the visual appeal section. There were mixed observations about the Sankey diagram as some participants found it very clear while some did not. One participant stated that *"The Sankey Diagram illustrates very well the distribution of features according to the level of toxicity and vice versa, it is very easy to understand and provides a lot of information, it would be useful if the information could be isolated interactively"*. Various participants agreed with the statement and they have described the Sankey diagram as being rather difficult to understand at first but afterwards it was clear as the information was globally displayed. Having more interactive elements could solve the problems of interpretation as users can filter down to explore features separately. On the other hand, some participants described it as chaotic and confusing.

Lastly, the results showed that the Treemap was the least liked graph in this task. An interesting finding was that participants described the Treemap as difficult to follow and understand though it stored more information than the Sankey diagram and the Bar chart, such as comparisons in topics and comparisons of 'yes' and 'no' comments together. For example, a participant commented that *"The Treemap, which at first glance seems more unpleasant, when you look at it closely, it gives very interesting information, which is when we find the same attribute labelled two different ways, helping then in the finding of inconsistencies in the annotation. For example, in the Economy topic, toxicity level 2 is represented by annotated comments such as Toxic = Yes and Toxic = No"*. However, most of the participants commented that they did not understand the Treemap.

Overall, results show that participants were attracted by the graphs that were easier to understand. The appeal of the graphs was also an important element, though, not as important than the ease of use. Many participants were not attracted to the graphs they did not understand and did not want to spend time on them, even though they liked their appearance more. Another finding suggests that participants would have benefited from greater guidance in the complex visualisations with various elements, which would have facilitated the comprehension, thereby proving more attractive. An idea is to combine simple graphics with more complex ones to create simple graphics like the Bar chart to guide graphs like the Treemap.

6 CONCLUSIONS AND FUTURE WORK

This paper presents a study related to data visualisation of hate speech (toxicity) annotations. To do so, we proposed various data visualisations, that includes different diagrams to assist annotators in the detection of inconsistent annotations, the analysis of the global data distribution and the discovery of relationships between features. We used a complex corpus composed by comments posted in different Spanish online new articles. All the comments were annotated using a new tagset that combines several features to determine the level of toxicity in a more objective way. The challenge in the proposed data visualisations was providing annotators with a wide spectrum of diagrams that highlight trends and relationships in an easy and comprehensible way. We proposed a diversity of visualisations (from those that were well-known to some others that could be new or unfamiliar for annotators). We conducted a preliminary evaluation of the proposed visualisations from collected qualitative and quantitative data obtained from a small but representative group of annotators. That is, they have different degree of expertise on visual analytic tools, and different knowledge in the annotation of corpus. We evaluated several dimensions of the visualisation experience (comprehensibility, interpretability and attractiveness).

In the following, we share our lessons learned, including design recommendations useful for future visualisation studies on corpus annotation. Regarding the comprehension of visualisations, the first consideration is that the participants mostly prefer the simple graphics (such as the Bar chart or Heat map), probably because annotators already acquainted with them on their daily annotation. However, the more expert participants in using visual analysis the more they valued visualisations that show more complex details (like the Sankey, the Spider or the Chord Diagram). Moreover, we also observed that in some cases the perception of the same diagram differed depending on which task and when it was visualised. For example, participants who rated the Sankey diagram as difficult to understand in Task 1, they found it easier in Task 3. Keeping this understanding in mind, the use of mixed visualisations that include easiest graphs -to facilitate ”the landing” in the annotated corpus-, together with the more complex ones - to explore more complex relations - could help to enhance the visualisation effectiveness. In relation to the interpretation of the visualisations, we found that participants appreciated having redundant information in the graphs (for instance, percentages and absolute values), and also they highly valued the interaction offered by some graphs, particularly in the ones that were more complex to understand (such as the Scatter Plot and the Tree Map). The use of interactivity by prioritising the most important attributes such as features, topics, Yes/No values, etc. to select the data to be shown is also an important fact to remind take into account in future visualisations. Last but not least, participants gave rather positive comments in open questions in those graphs which they scored high in the attractiveness dimension (such the Scatter Plot and the Spider Diagram). Making visualisations attractive and clear, using suitable colours and icons, is engaging and, more importantly, instructive and enlightening.

In the future, the first step is to design novel visualisations tailored according to linguists' needs with the findings of this paper and validate them by a larger group of annotators. Moreover, we plan to integrate the visualisations in earlier stages of the annotation process. Visualisation will serve as a tool to lead the annotation, letting users examine corpus as well as finding and editing inconsistent data. Additionally, we plan to introduce more valuable information related to the context in which the comments are obtained. Collecting information such as, where the message is post, who is the user, location and timestamps, could also help annotators to detect diffusion of hate speech.

ACKNOWLEDGEMENTS

To MISMIS-Language (PGC2018-096212-B-C33), CI-SUSTAIN (PID2019-104156GB-I00) and Accio COMRDI18-1-0010.

REFERENCES

Cao, N., Lin, Y.-R., Du, F., and Wang, D. (2015). Episogram: Visual summarization of egocentric social interactions. *IEEE Computer Graphics and Applications*, 36(5):72–81.

Chen, S., Lijing, L., and Yuan, X. (2017). Social media visual analytics. *Computer Graphics Forum*, 36(3):563–587.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Florio, K., Basile, V., Polignano, M., Basile, P., and Patti, V. (2020). Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Frénay, B. and Verleysen, M. (2013). Classification in the presence of label noise: a survey. *IEEE trans. on neural networks and learning systems*, 25(5):845–869.

Gagliardone, I., Gal, D., Alves, T., and Martinez, G. (2015). *Countering online hate speech*. Unesco Publishing.

Grosman, J. S., Furtado, P. H., Rodrigues, A. M., Schardong, G. G., Barbosa, S. D., and Lopes, H. C. (2020). Eras: Improving the quality control in the annotation process for natural language processing tasks. *Information Systems*, page 101553.

Knaflic, C. N. (2015). *Storytelling with data: A data visualization guide for business professionals*. John Wiley & Sons.

MAT (2020). The MITRE annotation toolkit. [Online; accessed 2-sep-2020].

Paschalides, D., Stephanidis, D., Andreou, A., Orphanou, K., Pallis, G., Dikaiakos, M. D., and Markatos, E. (2019). Mandola: A big-data processing and visualisation platform for monitoring and detecting online hate speech. *ACM Trans. on Internet Technology*, 37(4):1–21.

Reid, M. (2020). Creating a Chord Diagram with Tableau Prep and Desktop. [Online; accessed 13-sep-2020].

Resnick, P., Carton, S., Park, S., Shen, Y., and Zeffer, N. (2014). Rumorlens: A system for analyzing the impact of rumors and corrections in social media.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Shi, Y., Liu, Y., Tong, H., He, J., Yan, G., and Cao, N. (2019). Visual analytics of anomalous user behaviors: a survey. *arxiv:1905.06720v2*.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation.

Watson (2020). Watson knowledge studio. [Online; accessed 2-sep-2020].

Wu, Y., Cao, N., Gotz, D., Tan, Y.-P., and , Keim, D.-A. (2016). A survey on visual analytics of social media data. *IEEE trans. on multimedia*, 18(11):2135–2148.

Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations.

Zhao, J., Cao, N., Wen, Z., Song, Y., Lin, Y.-R., and Collins, C. (2014). fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE trans. on Visualization and Computer Graphics*, 20(12):1773–1782.