Introducing the Multidisciplinary Design of a Visualisation-Oriented Natural Language Interface

Ecem Kavaz[1], Francesca Wright[1], Montse Nofre[1], Anna Puig[1,2], Inmaculada Rodríguez[1,3] and Mariona Taulé[1,3]

[1]*CLiC, Centre de Llenguatge i Computació, Universitat de Barcelona (UB)*
[2]*IMUB, Institut de Matemàtica, Universitat de Barcelona (UB)*
[3]*UBICS, Institute of Complex Systems, Universitat de Barcelona (UB)*

Abstract

This paper introduces the demo for an innovative Data Visualisation in Linguistics platform tailored for the analysis of hierarchical multivariate data (DVIL). It is a Visualisation-oriented Natural Language Interface (V-NLI) that seamlessly integrates both direct manipulation, featuring diverse visualisation types and glyphs, and conversational interaction styles. Moreover, it incorporates a chatbot especially designed to facilitate user-guided visual analysis, a VisChatbot, enhanced by linguistic improvements. We showcase DVIL's efficacy in a practical case study focused on the analysis of toxic language within online news platforms, particularly highlighting its suitability for dissecting conversations structured as threads.

Keywords

Data visualisation, multivariate hierarchical data, natural language processing, chatbot, visualisation chatbot, hate speech

1. Introduction

Recent advances in Natural Language Processing (NLP) have favoured the development of Visualisation-oriented Natural Language Interfaces (V-NLIs)[1], which allow users to interact with data visualisations using Natural Language (NL). These V-NLIs usually integrate a chatbot (VisChatbot) that coexists with WIMP-based (Windows, Icons, Menus, Pointer) interaction, ultimately aiming at enhancing the user experience (UX) of visualisation analysis. In this article, we present DVIL (Data VIsualization in Linguistics), a V-NLI intended for linguists who need to analyse annotated datasets. Here, we extend the description in [2] that mainly focused on basic functioning and the platform's software architecture. Moreover, this paper presents the multidisciplinary work done by computer scientists and linguists to design a VisChatbot.

As case study, we present the analysis of toxic language, based on NewsCom-TOX [3]. This corpus consists of annotated comments from Spanish digital media news, organised into threads, thus forming complex hierarchical structures of multivariate data (i.e. each data point has several attributes). In this case, each comment of a news article is a data point and is labelled with a set of linguistic features, such as argumentation, sarcasm or insult, among others.

Note that data visualisation is useful throughout the entire process of the data analysis when linguists feed the NLP learning models, from the individual annotations through the definition of the Gold Standard, i.e the agreement achieved by several annotators (see steps (1) to (3) in Figure 1) to the visualisation of the automatic classification (step (4)). Specifically, the visualisations shown in this paper correspond to the first part of the process, i.e. analysis of data resulting from the Gold Standard.

2. Context

2.1. The NewsCom-TOX Corpus

Our research uses a specific data model, the NewsCom-TOX corpus [3], the aim of which is to study toxic language in the comments of news items appearing in Spanish digital media. The corpus consists of 4,359 comments posted in response to news articles extracted from online newspapers. These articles were manually selected taking into account their controversial subject matter, their potential toxicity, and the number of comments posted. We used a keyword-based approach to search for articles related mainly to immigration. The comments were manually annotated for toxicity, to analyse and identify

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain.

ekavazka27@alumnes.ub.edu (E. Kavaz); frwrighw11@alumnes.ub.edu (F. Wright); montsenofre@ub.edu (M. Nofre); annapuig@ub.edu (A. Puig); inmarodriguez@ub.edu (I. Rodríguez); mtaule@ub.edu (M. Taulé)
https://clic.ub.edu/e_kavaz (E. Kavaz); https://clic.ub.edu/m_nofre (M. Nofre); https://webgrec.ub.edu/webpages/personal/ang/001040_annapuig.ub.edu.html (A. Puig); https://webgrec.ub.edu/webpages/000006/ang/inmarodriguez.ub.edu.html (I. Rodríguez); https://clic.ub.edu/m_taule (M. Taulé)
0000-0002-3973-2247 (E. Kavaz); 0000-0001-6883-7657 (M. Nofre); 0000-0002-2184-2800 (A. Puig); 0000-0001-5931-7713 (I. Rodríguez); 0000-0003-0089-940X (M. Taulé)

CEUR Workshop Proceedings (CEUR-WS.org)

messages with racial and xenophobic content. Therefore, a specific set of labels, corresponding to features of toxic language, was designed to analyse and identify messages with racial and xenophobic content.

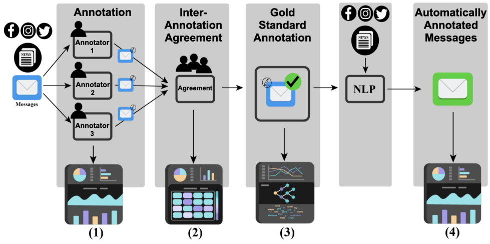


Figure 1: Visualisation along data annotation: (1) Individual annot., (2) Inter-annotator agreement, (3) Gold-Standard annot., and (4) Automatic annot.

Detecting toxic language is a difficult task because this type of language has a high and unavoidable subjectivity. In fact, new approaches are now being developed to model conflicting perspectives and opinions coming from people with different personal and demographic backgrounds [4]. In our case, we follow the model used so far for annotation, inter-annotator agreements and definition of a gold standard corpus. This corpus has also been used in the DETOXIS (DEtection of TOxicity in comments In Spanish) task [5] and, partially task [6].

2.2. Annotation Process

The NewsCom-TOX corpus is multi-level annotated with different binary linguistic categories taking into account the information conveyed in each comment and also the whole discourse thread in which the comment occurs. Therefore, the comments are hierarchically structured in the form of threads, with comments that refer directly to the news item and others that are responses to previous comments. Figure 2 is an example of hierarchical structure: the root of the hierarchy is the news item (at the top of the figure), Comment 1 and Comment 4 are direct comments to the news, and Comments 2 and 3 are responses to Comment 1 and 2, respectively.

The linguistic features we annotate are: argumentation, constructiveness, stance, target, stereotype, sarcasm, mockery, insult, improper language, aggressiveness and intolerance (gray squares in Figure 2). All these features have a binary value, indicating its presence or absence. Furthermore, some of the features can be correlated, for argumentation and constructivity, insult and improper language, and these correlations are useful to assign the level of toxicity. As a result of the annotated features, we classify each comment as 'toxic' or 'not toxic', and we assign different levels of toxicity (1=mildly toxic, 2=toxic, 3=very toxic) to those that are annotated first as toxic, depending on type and quantity of features we mark in each comment. We hypothesise that the combination of these categories helps to determine the level of toxicity more objectively.

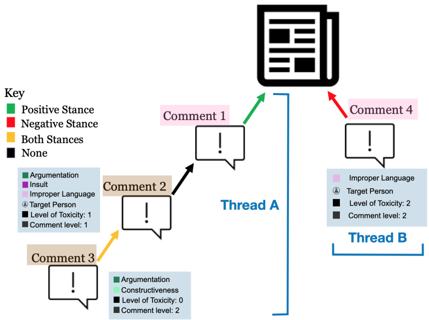
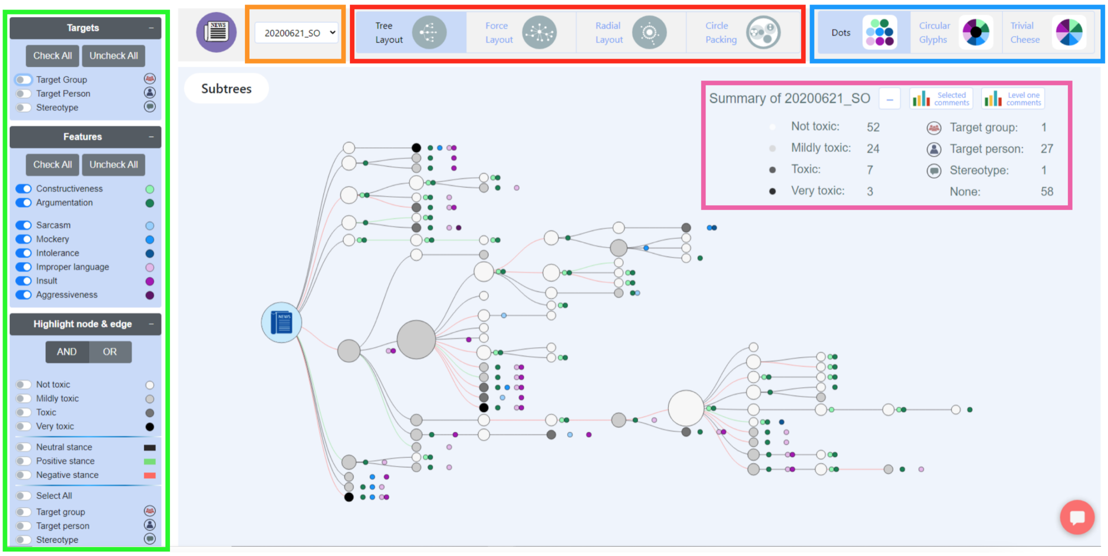


Figure 2: Example of annotation of the NewsCom-TOX corpus: level and threads of comments, features and stances.

We also annotate the contextual information: the conversational thread in which the comment occurs. This information is very useful for the annotators since it helps them to better interpret and understand the content of the message [7]. The contextual information includes a number that indicates the chronological order in which the comment was posted in the time thread on the website (number of comments in Figure 2), and an identifier of the thread in which the comments are embedded (Comments 1 to 3 belong to Thread A and Comment 4 belongs to Thread B). A comment may directly refer to the news itself or a previous comment; in the latter case, a conversation or discussion between different users can emerge. A comment is categorised as a level 1 comment when it refers directly to the news article itself (in Figure 2 Comments 1 and 4, highlighted in pink). Otherwise, if the comment does not directly relate to the news but instead addresses a preceding comment, it is classified as a level 2 comment (Comments 2 and 3, coloured in brown). Finally, the term "stance" refers to the position that a comment takes in relation to the news or the comment it refers. For example, if a comment aligns with and supports the argument made in the news or the comment it refers, it is said to have a positive stance, indicating a continuity in the line of reasoning. Also, the stance can be negative, if a comment disagrees the news or the comment it refers, and it is considered neutral when a comment neither supports nor opposes news or the previous comment. Understanding the stance helps to analyse the flow of conversation and identify patterns of agreement or disagreement in the argumentation presented.

In summary, each comment is annotated following these criteria by three annotators in parallel (step (1) in Figure 1), and an inter-annotator agreement test is carried out once all the comments on each article have been



Summary of 20200621_SO

Not toxic:	52	Target group:	1
Mildly toxic:	24	Target person:	27
Toxic:	7	Stereotype:	1
Very toxic:	3	None:	58

Figure 3: Visualisation platform showing a tree layout with dots displaying features.

annotated (step (2)). Then, disagreements are discussed by the annotators and a senior annotator until an agreement is reached. The team of annotators involved in the task consisted of two expert linguists and two trained annotators, who were linguistics students.

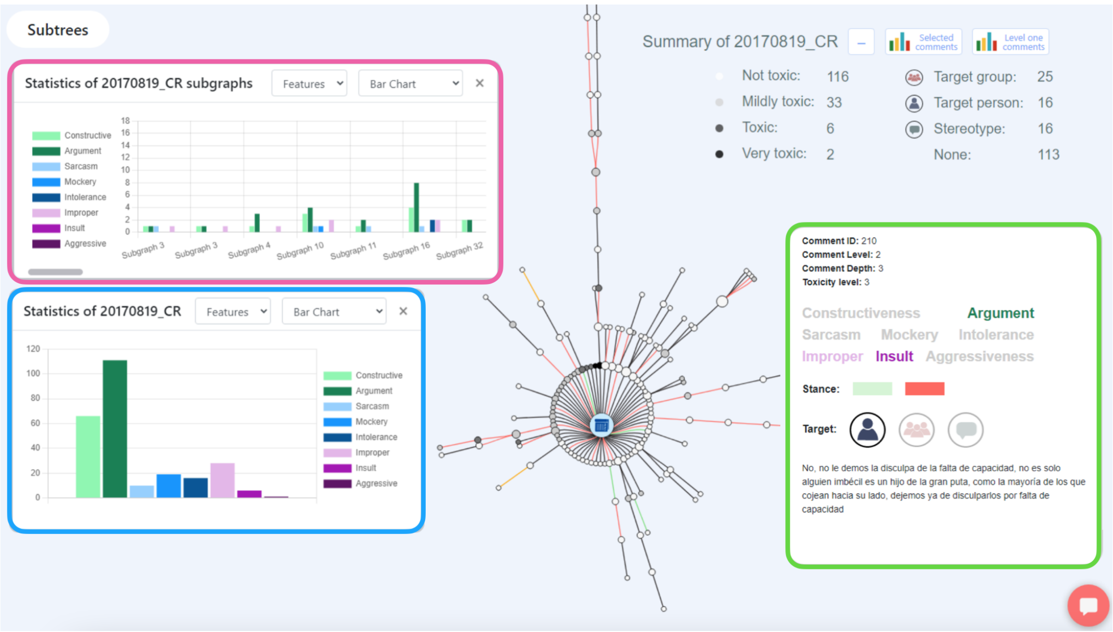
3. Visualisation Platform

This section first describes how hierarchical multivariate data is represented in the DVIL platform, then the elements of the WIMP-based interface such as the filters and others, and finally the VisChatbot interaction.

3.1. Main Visualisation

As NewsCom-TOX contains hierarchical and multivariate data, our goal is to visualise the data structure and features cohesively, ensuring no detail is lost. Each hierarchical structure can have different characteristics such as nodes at different depths of the hierarchy (forming a tree-shaped elongated structure, see Figure 3), or having more nodes connected to the root node directly (forming a star-shaped compact structure, Figure 4). To facilitate the analysis of such a variety of shapes of hierarchies, the DVIL platform includes several visualisation types (layouts): Tree, Radial, Force, and Circular Packing (see red frame in 3). Moreover, we developed an algorithm for categorisation to automatically decide the layout of the opened hierarchical visualisation, ensuring the selected one conveys the most informative presentation possible [8]. Note that we also give control to the users so that they can decide to change the layout at any time.

Each layout begins with a root node that represents the news article (see big blue node in Figure 3), and connected nodes representing comments on it or comments to other comments. The size of the nodes corresponds to the number of child nodes each node has. We visualise the level of toxicity directly on the layouts, employing a colour range, in which white denotes non-toxic and black denotes very toxic. We clustered the comments' features in three groups (stances, targets and abstract features) to visualise them on demand in the most informative way. We visualise stances on the edges that connect comments in the hierarchical structure as green for positive stance, red for negative and orange for both. As targets are more concrete features, we decided to visualise them with icons (see Target Group, Target Person and Stereotype in the Summary area, pink frame in Figure 3). Finally, to visualise more abstract features like Sarcasm, Mockery, Intolerance and others, we designed three glyphs. The former is an one-by-one glyph which shows features side by side (Dots). The last two glyphs show all of features together, i.e. all-in-one, and shows the features in a circular way or in a cheese-shaped glyph (see the three icons in the blue frame, and note that the main visualisation shows nodes' features as (coloured) Dots). Specifically, we used green shades for positive features (i.e. constructiveness), blue for neutral features (i.e. sarcasm), and magenta for negative features (i.e. insult), which can be seen in next to the nodes (i.e comments) in Figure 3.



Not toxic: 116
Mildly toxic: 33
Toxic: 6
Very toxic: 2
Target group: 25
Target person: 16
Stereotype: 16
None: 113

Comment ID: 210
Comment Level: 2
Comment Depth: 3
Toxicity level: 3

Constructiveness Argument
Sarcasm Mockery Intolerance
Improper Insult Aggressiveness

Stance: 🟩 🟥

Target:

No, no le demos la disculpa de la falta de capacidad, no es solo alguien imbécil es un hijo de la gran puta, como la mayoría de los que cojean hacia su lado, dejemos ya de disculparlos por falta de capacidad

Figure 4: Additional bar charts showing the statistics for the whole visualisation and subgraphs in a force layout visualisation.

3.2. WIMP-Based Interface

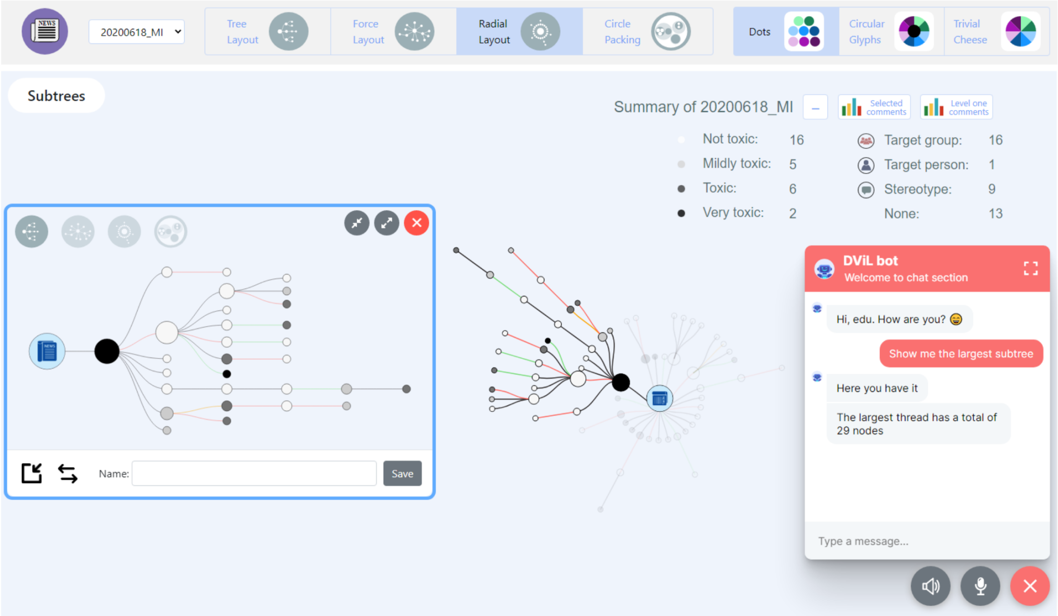
There are various options available for interacting with the visualisations to analyse them. Located on the side menu (see the green frame in Figure 3), there are filters which allow you to highlight nodes based on their objective (target), characteristics (features), orientation (stance) and toxicity (levels of toxicity). The top menu provides options to navigate back to the main page, switch between the four implemented layouts (Tree, Force, Radial, and Circle Packing), and select glyph types. Additionally, positioned in the upper right corner, beneath the top menu, is a summary of graph statistics (targets and levels of toxicity) that can be expanded or collapsed according to the user's preference for viewing statistics or solely interacting with the graph. Buttons for statistic graphs (bar and pie) for visualising features for the whole graph or subgraphs are displayed in this section as well. Moreover, the user can analyse the statistics of features in additional charts shown in pop-ups windows (Figure 4), in particular, the statistics of all the features of the whole visualisation (bar chart in the blue frame), and the statistics of subgraphs, i.e. subparts of the hierarchical structure (bar charts in the pink frame).

The statistics graphs, shown in pop-ups, allow for quick analysis and the ability to establish correlations between features. For example, the correlation between the level of comments and toxicity level can help to support the hypothesis that comments of level 1 (those that refer to the news directly, tend to be less toxic than comments of level 2. Furthermore, with a tooltip we visualise all the details about a comment including the actual comments, it's Comment id, Thread id, features, stances and targets (green frame in Figure 4).

3.3. VisChatbot Interface

VisChatbot knowledge and functions are specialised towards the DVIL interface with the goal of facilitating the visualisation of data and the statistical analysis of the NewsCom-TOX corpus. Linguists can analyse their corpus by requesting the chatbot to carry out functions that they would otherwise have to carry out "manually", i.e. through several interactions with filters and buttons of the WIMP; to carry out actions that are not accessible on the interface itself; and to ask for help and explanations about the domain of the data, in our case study toxic language, e.g. "what is mockery?". The chatbot interface is the usual website chat widget , and its interaction possibilities include text-based interaction, multimedia interaction (through additional charts as responses to users' queries), and speech interaction.

In the first user-VisChatbot interaction (see Figure 5), the chatbot greets the user and explains how it can help the user by asking if they would like a whole tutorial or any help. Afterwards, the user can interleave natural



Summary of 20200618_MI

Not toxic:	16	Target group:	16
Mildly toxic:	5	Target person:	1
Toxic:	6	Stereotype:	9
Very toxic:	2	None:	13

DViL bot
Welcome to chat section

Hi, edu. How are you? 😄

Show me the largest subtree

Here you have it

The largest thread has a total of 29 nodes

Type a message...

Figure 5: Example of a pop-up displaying a subtree obtained from a force layout by a query in the chatbot.

language (NL) queries and WIMP interactions. Note that the chatbot can maintain the context of the conversation but also the context of WIMP interactions. For example, the user selects a filter using the mouse, and queries to the chatbot "unselect the filter". Specifically, the chatbot is capable of the following:

- Generating/updating a visualisation in response to a query (e.g. "Please highlight constructive and argumentative comments from the news").
- Able to interact with all GUI elements (filter, glyphs, statistics charts, layouts).
- Performing common operations on the platform (logging in or logging out, opening a dataset).
- Understanding follow-up queries (e.g. 1st query: "Show me mockery", 2nd query: "remove it").
- Explaining how to interact with interface elements ("How can I use glyphs?").
- Explaining the characteristics of the dataset (such as the levels of toxicity) and providing external links to our articles related to NewCom-TOX.

When the VisChatbot encounters difficulty understanding a query, it offers disambiguation widgets or requests the user to rephrase their intent. Furthermore, the chatbot offers textual feedback to the user confirming the request is done, and/or offer additional information and visual feedback by flashing a green light on the elements of the interface when necessary (e.g., user asked how to see glyphs, flashing will occur on the menu where glyphs can be activated). Additionally, our chatbot possesses the ability to comprehend both low-level queries, which typically involve simple one-turn interactions between the user and the bot, and high-level queries, which are more complex and cannot be executed using the WIMP , often requiring multiple turns (follow-up queries). For example, we integrated the functionality to extract subtrees (the most toxic, the longest... thread) via queries within the chatbot, presenting them on a re-sizable and repositionable pop-ups. This feature serves as a "zoom-in" on the subtree, facilitating a detailed examination of a particular section of the overall structure and enabling analysis of node characteristics without the necessity of focusing on the entire set (see blue frame in Figure 5).

The VisChatbot has been implemented using Rasa conversational framework [9], that consists of the i) *Natural Language Understanding (NLU) Component*, which analyses user input, identifies the intents (what the user wants) and extracts entities (names, dates) from their messages, with the possibility of using synonyms. ; ii) the *Dialogue Management Component* that determines, using rules and stories, the conversation flow based on the NLU output and the current conversation context; and finally, the *Actions Component* that allows developers to tailor the chatbot's functionality to specific needs. For example, in our case study, an action communicates with DVIL's frontend to update the current visualisation in response to a user's query.

4. VisChatbot Design

The VisChatbot was designed by linguistic experts, who together with computer scientists, aimed to offer natural, nuanced and fluent conversations related to the visualisation. To do so, we focused the linguistic aspects as indicated next:

- We defined the training examples taking into account not only synonyms but also paraphrases.
- We configured stories and rules to account for complex conversational pathways.
- We established the VisChatbot's responses to be appropriate and provide useful and concise information following the principle of minimisation in general conversations.
- We considered the conversational context allowing the use of co-references and ellipsis.
- We included buttons and charts in the VisChatbot's responses with some synchronisation with the WIMP interface (such as highlighting and annotations in the visualisation).

Finally, we carried out an exploratory study with five linguistics students [10] with encouraging results. Successful interactions significantly outnumbered failures. Analysing failed attempts revealed areas for improvement, such as missing training data, user errors (misspellings, poorly phrased queries), or limitations of the chatbot's capabilities. We'll use this data to enhance our V-NLI.

5. Conclusions

This paper presents DVIL, a conversational platform for data visualisation in Linguistics. DVIL integrates WIMP and conversational interaction styles, and enables different visualisation types (tree, radial, force, circle packing), glyphs and additional basic charts. The platform incorporates a VisChatbot designed to interact with users and guide them through the different visualisation options, enhanced with a variety of linguistic improvements. We also show the functionality of the platform in a case study related to the analysis of toxic language in digital news media, highlighting its usefulness for analysing structured conversations such as threads. However, the model is extendable to other scenarios.

6. Acknowledgments

This work was supported by the SGR CLiC project (2021 SGR 00313, funded by the Generalitat de Catalunya), FairTransNLP-Language (PID2021-124361OB-C33, MICIU/AEI/10.13039/501100011033/FEDER,UE) and ACISUD (PID2022-136787NB-I00 funded by MICIU/AEI/10.13039/501100011033).

References

[1] L. Shen, E. Shen, Y. Luo, X. Yang, X. Hu, X. Zhang, Z. Tai, J. Wang, Towards natural language interfaces for data visualization: A survey, IEEE Transactions on Visualization and Computer Graphics 29 (2023) 3121–3144. doi:10.1109/TVCG.2022.3148007.

[2] E. Kavaz, A. Puig, I. Rodríguez, A conversational data visualisation platform for hierarchical multivariate data, in: C. Gillmann, M. Krone, S. Lenti (Eds.), EuroVis 2023 - Posters, The Eurographics Association, 2023. doi:10.2312/evp.20231053.

[3] M. Taulé, M. Nofre, V. Bargiela, X. Bonet, Newscomtox: a corpus of comments on news articles annotated for toxicity in spanish, Language Resources and Evaluation (2024). doi:10.1007/s10579-023-09711-x.

[4] S. Akhtar, V. Basile, V. Patti, Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection, 2021. doi:https://doi.org/10.48550/arXiv.2106.15896. arXiv:2106.15896.

[5] M. Taulé, A. Ariza, M. Nofre, E. Amigó, P. Rosso, Overview of detoxis at iberlef 2021: Detection of toxicity in comments in spanish, Procesamiento del Lenguaje Natural 67 (2021) 209–221. doi:10.26342/2021-67-18.

[6] A. Ariza-Casabona, W. S. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, P. Rosso, Overview of detests at iberlef 2022: Detection and classification of racial stereotypes in spanish, Procesamiento del Lenguaje Natural 69 (2022) 217–228. doi:10.26342/2022-69-19.

[7] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, I. Androutsopoulos, Toxicity detection: Does context really matter?, in: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4296–4305. doi:https://doi.org/10.18653/v1/2020.acl-main.396.

[8] E. Kavaz, A. Puig, I. Rodríguez, R. Chacón, D. De-La-Paz, A. Torralba, M. Nofre, M. Taule, Visualisation of hierarchical multivariate data: Categorisation and case study on hate speech, Information Visualization 22 (2023) 31–51. doi:10.1177/14738716221120509.

[9] Rasa, Rasa conversational platform, 2023. URL: https://rasa.com/.

[10] F. Wright, Enhancing the dvil chatbot through linguistic expertise. [degree thesis. facultat de filologia i comunicació, universitat de barcelona (ub)], 2023. URL: http://hdl.handle.net/2445/203511.