A Data Viz Platform as a Support to Study, Analyze and Understand the Hate Speech Phenomenon

Arthur T.E. Capozzi
Università degli Studi di Torino
Dipartimento di Informatica
Turin, Italy
arthur.capozzi@gmail.com

Viviana Patti
Università degli Studi di Torino
Dipartimento di Informatica
Turin, Italy
patti@di.unito.it

Giancarlo Ruffo
Università degli Studi di Torino
Dipartimento di Informatica
Turin, Italy
ruffo@di.unito.it

Cristina Bosco
Università degli Studi di Torino
Dipartimento di Informatica
Turin, Italy
bosco@di.unito.it

ABSTRACT

In this paper we present a data visualization platform designed to support the Natural Language Processing (NLP) scholar to study and analyze different corpora collected with the purpose to understand the hate speech phenomenon in social media.

The project started with the creation of a corpus which collects tweets addressed to specific groups of ethnic minorities considered very controversial in the Italian public debate. Each tweet has been manually tagged with a series of attributes in order to capture the different features used to characterize the hate speech phenomenon. This corpus is mainly built to be used for training an automatic classifier and helping us in its testing and validation, before being it adopted to detect tweets targeted as hate speech on larger scale datasets. As opposed as many other traditional machine learning tasks, to build a good classifier achieving high scores in terms of accuracy is very challenging in such scenario, because of the intrinsic ambiguity of the language, the lack of a proper and explicable context in social media, and the attitude of on line users of being sarcastic and ironical. Therefore, in order to properly validate an effective feature selection process, correlations between selected attributes must be studied and analyzed. This motivated us to build an interactive platform to explore data in our corpora across the dimensions that have been used to characterize collected tweets.

In our paper, after a brief introduction of the hate speech dataset, we will show how the dashboard can fit into the NLP pipeline, and how its architecture can be structured. Finally, we will present some of the challenges we have faced to visualize data with spatial, temporal and numerical attributes.

CCS CONCEPTS

• **Human-centered computing → Visualization systems and tools**; • **Computing methodologies →** *Natural language processing*;

KEYWORDS

Data Visualization, dashboard, hate speech

ACM Reference Format:
Arthur T.E. Capozzi, Viviana Patti, Giancarlo Ruffo, and Cristina Bosco. 2018. A Data Viz Platform as a Support to Study, Analyze and Understand the Hate Speech Phenomenon. In *International conference on Web Studies (WS.2 2018), October 3–5, 2018, Paris, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3240431.3240437

INTRODUCTION

The political and social debate has nowadays been largely shifted to the Web, especially in social media. Politicians have found a new way to communicate directly with the citizens; journalists and scholars can monitor news spreading and opinion formation almost in real time; citizens have the opportunity of being more engaged in the debate and of being able to express their own ideas directly to the actors involved in the decisional processes.

When we focus in particular on the many opportunities that are available to scientists, data streaming from social media is a powerful source to detect trends and opinion shifts among citizens and also a testbed to validate social science theories. For example, using data from social media, scholars have tried to predict electoral results [9] and the stock market [1] as well as to compare traditional polls with the preferences expressed on line by citizens [7]. De Choudhury et al. [5] have instead tried to predict tendency to depression on the basis of tweets from selected individuals.

This kind of data can be very valuable, but we must also stress that other scholars have underlined the limitations of approaches that try to predict with absolute certainty the opinion of citizens expressed through Twitter [2, 4]. It is also important, in drawing conclusions through the analysis of data collected from social media, to consider various limitations arising, for example, from the spread of social media in particular countries or regions.

WS.2 2018, October 3–5, 2018, Paris, France

ACM ISBN 978-1-4503-6438-6/18/10…$15.00
https://doi.org/10.1145/3240431.3240437

Hate speech is considered today a phenomenon to be studied with particular attention, because of the availability of user generated content and the implications in the digital society, where the interplay between opinions formed and expressed on line and actions that individuals may decide to adopt on and off line can escalate extremely quickly. This may trigger cascade and feedback effects that are highly valuable for marketing, but that can also be difficult to stop when the gossip is unwanted or even dangerous. Moreover, the predictability of viral phenomena is still an hard scientific problem to be handled, and this may cause many different uncontrollable scenarios.

With the advent of new forms of communication and the social media revolution, hate speech is likely to have changed its characteristics and ways to be expressed. The motivations can be many, but among them, the feeling of a presumed anonymity that users have in social media, can be taken into account.

An important debate is currently ongoing with policy makers that are trying to support the principle that hate speech on line should be stopped or fought because of its uncontrollability; on the other side, some others discuss that the phenomenon must be considered as an instance of the many different manifestations of free speech, and that any attempt to stop hate speech is indeed an act of repression that must be condemned and avoided.

We think that there is still a lack of understanding of the phenomenon itself, and that a lot of efforts must be devoted to develop new automatic tools to analyze and to detect hate speech among the many other different figurative expressions that arise from social media. A way to understand hate speech could be to quantify and qualify it by studying its temporal evolution, the hypes corresponding to some specific events or targets, and the correlations between different attributes that characterize the phenomenon. The huge availability of data could be an important tool to perform extensive analyses, but before that such large scale studies will be considered reliable, we need to trust a new generation of classifiers whose accuracy on big corpora must be at least as good as the manual annotation performed by human beings.

In this paper we present a data visualization platform designed to support the NLP scholar to study and analyze different corpora collected with the purpose to understand the hate speech phenomenon in social media.

The project started with the creation of a corpus which collects tweets addressed to specific groups of ethnic minorities considered very controversial in the Italian public debate. Each tweet has been manually tagged with a series of attributes in order to capture the different features used to characterize hate speech [10]. The main purpose of this corpus is to be used as a training set for learning an automatic classifier. In a forthcoming phase of the project, this classifier will be tested and validated before being adopted to detect tweets targeted as hate speech on larger scale datasets.

As opposed as many other traditional machine learning tasks, to build a good classifier achieving high scores in terms of accuracy is very challenging in such scenario [11], because of the intrinsic ambiguity of the language, the lack of a proper and explicable context in social media, and the attitude of on line users of being sarcastic and ironical. Therefore, in order to properly validate an effective feature selection process, correlations between selected attributes must be studied and deeply understood. This motivated us to build an interactive platform to explore data in our corpora across the dimensions that have been used to characterize collected tweets.

In the following sections, after an overview of related work and a brief introduction of the dataset, we will show how our dashboard can fit into the NLP pipeline, and how its architecture is structured. Finally, we will present some of the challenges we faced to visualize data with spatial, temporal and numerical attributes.

RELATED WORK

As observed by many experts of the field (Marti Hearts, Stephen Few, Isabel Meirelles, to cite just a few), the main difficulty of visualizing nominal or categorical data graphically is that there is neither an obvious nor an objective way to quantify such information. Texts, naturally represented by means of lexical, syntactic, and semantic transformations, are communication devices of a categorical nature, and they have no inherent ordering, lacking of a formal well defined structure, and suffering of high dimensionality.

Nevertheless, many different attempts have been made to develop textual data visualizations methodologies, in order to support computational linguistic scholars, or to give users a more intuitive way to navigate through their own emails, files, and generic textual information.

As pointed out by Isabel Meirelles in [6], the awareness of the difficulty of representing graphically textual information can be dated back to surprisingly ancient times. The Codex of St. Peter, XIV century, and any other attempts to deliver a visual interpretation of the bible to illiterate believers, is an evidence of this effort.

Generally speaking, we can classify textual data visualizations in three different types: (i) visualization of *connections* among entities within and across documents in a corpus; (ii) visualization of document *concordances* and word *frequencies*; and (iii) visualization of *relationships* between words when used in given languages and lexical ontologies.

As an example of the first type, Stephen Few's "on the Origin of the Species: The Preservation of Favoured Traces" visualization[1] allows to detect the different edits in the several revisions of Charles Darwin's "The Origin of the Species", since the first version published in 1859. The original interactive version was designed to serve exploratory and teaching purposes, and users were able to see changes at both the macro level, and word-by-word. This idea constitutes a formidable tool for text mining, that enables the users to follow the evolution of scientific ideas during the lifetime of the scientist himself.

Google, over the years, delivered a number of services that can be presented as instances of the second type of textual data visualization tools: *Google Trends*, *Google Correlate*, and *Google Ngram Viewer* are widely accessible to everyone who wants to perform some kind of literature analysis. These services have become fundamental tools to understand how the language evolved by means of word frequencies and concordances within a huge collection of digitalized books, originally written across centuries.

Visualizations of the above mentioned third type usually provide tools for literacy and citation analysis; e.g., PhraseNet [13], Word

[1] https://fathom.info/traces/

Tree [16], Web Seer [2], and Themail [14] introduced many different ways to generate visual overviews of unstructured texts. Many of these projects were connected to *ManyEyes*, that was launched in 2007 by Fernanda B. Viégas, Martin Wattenberg, and al. [15] at IBM, and closed in 2015 to be included in IBM Analytics. ManyEyes represented probably a step forward in the exploitation of relationships among different artworks: it was designed as a web site where people could upload their data, to create interactive visualizations, and to establish conversations with other authors. The ambitious goal was to create a social style of data analysis so that visualizations can be tools to create collaboration and carry on discussions.

Another visualization that should be cited here is Moritz Stefaner's *Revisit* [3]. Revisit has some similarities with the purpose of our project, although it serves a very different purpose. Its main task is to provide a real-time visualization of tweets around a specific topic. We also focus on some given topics (namely targets, see following sections), even if we are not interested on real-time displaying; in fact, Revisit has been designed to create Twitter walls at conferences or other environment. However, differently to other Twitter stream tools, it focuses on temporal dynamics in the stream. We also want to focus on temporal dynamics presenting an interactive time line to let our users look for hypes contrasting a white-noise of continuous adoption of hate speech under different conversational scenarios. Such hypes may be caused by external events that trigger on line discussions between polarized social media communities. One of the goal of our platform is to look for evidences of these (probably not linear) cause-effect relationships on topics targeted by the scholar that created the tweets' corpora.

Finally, an important source of inspiration is the Bubble Chart introduced by the visualization pioneer Hans Rosling and his Gapminder foundation[4]. Bubble Charts are largely used today to look for patterns within different sources of numerical information made of many dimensions, some of which can be categorical. The challenge is to adopt this widely adopted visualization paradigm when visualizing textual information aggregated in terms of targets. Our proposal is described in the following sections.

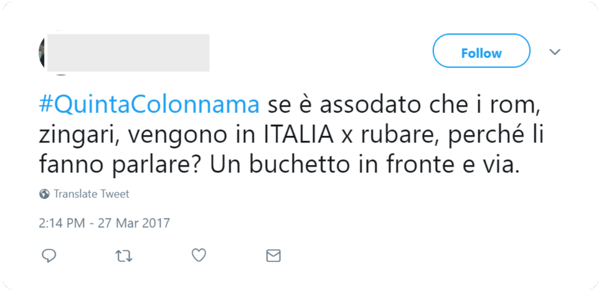
DATA MODEL

Our platform is tailored to deal with a specific data model that has been adopted. Such model is in line with the semantic scheme designed and applied recently in developing a set of Italian corpora of tweets, created with the aim to build a pool of reference datasets on different targets of hate for an automatic system of hate speech monitoring [5]. At the moment the efforts where devoted to develop a first corpus annotated for hate speech against immigrants [10], but also other corpora focusing on the identification of other targets of hate and abusive behavior are currently under development. This is the reason we decided to implement this data visualization dashboard, that will be suitable to explore also the other datasets that will be produced in the future. The first corpus that motivated the design of our interactive platform is described in detail in [8, 10] and is composed of a list of tweets manually annotated by a pool of expert annotators and by CrowdFlower contributors. All the tweets of the corpus refer to three different targets polarizing the Italian public debate about the migration phenomenon.

The corpus consists of about 6,000 annotated tweets, and has been created as a selection of the Italian tweets posted by Twitter users from October 1st, 2016 to April 25th, 2017 and streamed down via official APIs. The overall dataset was filtered by the presence of keywords referring to one of the selected targets: Ethnic Group, Religion, and Roma. From this dataset we randomly selected a subset to be annotated. A detailed description of the entire pipeline of the data collection and annotation can be found in [8, 10].

As previously mentioned, tweets have been tagged with specific attributes to describe the facets of the hate speech. The various attributes, their value domain and a brief description are listed below:

- **target**: [Roma, Ethnic Group, Religion] the target towards which the tweet is directed;
- **hate speech**: [yes, no] a boolean value that identifies the presence of hate speech within the tweet;
- **intensity**: [0, 1, 2, 3, 4] it can be higher than 0 only if hate speech occurs to shows its intensity;
- **offensiveness**: [no, weak, strong] it describes the degree of offensiveness of the message; like the remaining attributes, offensiveness may occur even if hate speech doesn't;
- **aggressiveness**: [no, weak, strong] it identifies the degree of aggressiveness of the tweet;
- **irony**: [yes, no] it shows if a figurative use of language occurs in the message;
- **stereotype**: [yes, no] it identifies the presence of stereotyping within the tweet.



[redacted]

#QuintaColonnama se è assodato che i rom, zingari, vengono in ITALIA x rubare, perché li fanno parlare? Un buchetto in fronte e via.

2:14 PM - 27 Mar 2017

Figure 1: Example: the tweet above, whose target is 'Roma' is part of the corpus.

Fig. 1 shows one tweet in the corpus, that can be translated as it follows: "#QuintaColonna but if it is common knowledge that Roma, Gypsies, come to ITALY to steal, why do they make them talk? A little hole in the forehead and we're good.". The hashtag *#QuintaColonna* refers to a popular talk show in Italy.

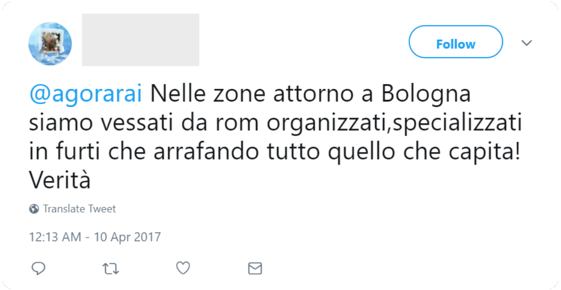
After the annotation process, the tweet is associated with a series of tags and it appears as reported in Table 1.

It is illustrative to recall that hate speech is indeed a speech, gesture, writing, or display that incites violence or prejudicial action against a person or a group on the basis of attributes such as race, religion or ethnic origin. This means that the tweet reported in Fig. 1 contains hate speech because of the explicit call for violence.

[2] http://hint.fm/seer/
[3] http://truth-and-beauty.net/projects/revisit/
[4] https://www.gapminder.org
[5] http://hatespeech.di.unito.it

Table 1: Attributes' values, after the annotation process, characterizing tweet in Fig. 1. The id field can be used to get the original tweet using Twitter API, if it is still available.

id	target	hate speech	intensity	offensiveness	aggressiveness	irony	stereotype
#	Roma	yes	3	strong	strong	yes	yes



[redacted]

@agorarai Nelle zone attorno a Bologna siamo vessati da rom organizzati,specializzati in furti che arrafando tutto quello che capita! Verità

12:13 AM - 10 Apr 2017

Figure 2: Example: the tweet above contains a violent message, but cannot be annotated as hate speech.

On the contrary, tweet shown in Fig. 2 cannot be annotated as 'hate speech', although the presence of negative anti-Roma stereotypes can be identified. For the sake of completeness, the annotation of this tweet is reported in Table 2.

The complete resource is going to be made freely available and accessible for non-commercial use by the end of 2018[6].

As a side note, we completed the dataset adding to every single annotated tweet two other attributes, namely the *date* when the tweet was created and posted on line, and the geographic *location* of the user's device at the moment of writing (even if the latter information is not always available). This allows us to explore data at different spatial and temporal granularities, for example aggregating annotated tweets by day, week, or month, or by geographical region.

THE DASHBOARD

In this section we introduce the dashboard designed for exploring corpora adopting the annotation schema described above. Our platform is a web application implemented on top of a mongodb/node.js/ express.js stack, using d3.js as the main visualization front end library. It allows the exploration of all the annotated tweets in the corpus stored in back end, and it follows the popular interactive pattern known as the Shneiderman's mantra [12]: Overview first, zoom and filter, details on demand. There are basically three views that return different levels of details on data, that are shown in Figures 3, 4, and 5. The different views will be described later on.

Introduction to the dashboard

As discussed in the related work section, the visualization of nominal and categorical data, particularly of textual ones, is a topic raising an increasing interest. Moreover, the higher computational power of computers is allowing better performance of NLP and data analysis techniques.

After applying some basic transformations to a corpus, it may be useful to visualize the obtained annotated dataset to facilitate the search for patterns, outliers or correlations between attributes: quantification can be an easier task when applied to discrete attributes than directly to the original texts where they are involved.

The dataset obtained from the application of NLP functions can be characterized by a large number of attributes and may show different and complex facets. Furthermore, these attributes can belong to many different domains, making the task of feature selection for an automatic classifier also more complicated. In this situation, data visualization can be of great help as a support for the scholar.

A dashboard can be useful before and after the training phase: (i) a scholar can visually interact with the annotated tweets to better understand the information included in the corpus that will be used as training and test set for a classifier; (ii) after that such automatic classification task has been validated and proven to be effective, its execution on large scale corpora can help the NLP scientist to get rid of manual annotation of tweets. In this second and forthcoming domain, an interactive data visualization platform will become a powerful tool for the actual analysis of the hate speech phenomenon. When an analysis on large scale datasets will be feasible and reliable, other sources of data can be integrated to the already available streams, and more computational social science research questions would be addressable.

Both these uses of the dashboard require the exploration of the data and in the next section we will understand more in detail what are the tools provided for this purpose, even if the lack of a fully reliable automatic classifier restricts the use cases domain to the first scenario we described in the previous paragraph.

From now on, tweets have temporal attributes (the creation date), geographical attributes (only for a small portion of the dataset), numerical attributes (the intensity attribute) and categorical attributes (for example aggressiveness can assume the values no, weak and strong).

However, both numerical and nominal attributes can be converted to discrete domains, and their values can therefore be represented in scales.

Tools provided

For the exploration of the dataset, we wanted to show the temporal evolution of the attributes in the clearest and most effective way possible. In addition to this, the dashboard provides various tools that should facilitate the following operations:

- search for correlations between the various attributes used to label tweets
- geographic analysis on several levels of detail
- the search for correlations between the increase in the number of tweets related to the targets of interest or an increase in the values of hate speech and news events.

[6] https://github.com/msang/hate-speech-corpus

Table 2: Attributes' values, after the annotation of the tweet in Fig. 2. Observe that in this example we do not have hate speech, despite the presence of aggressiveness and offensiveness.

id	target	hate speech	intensity	offensiveness	aggressiveness	irony	stereotype
#	Roma	no	0	weak	weak	no	yes

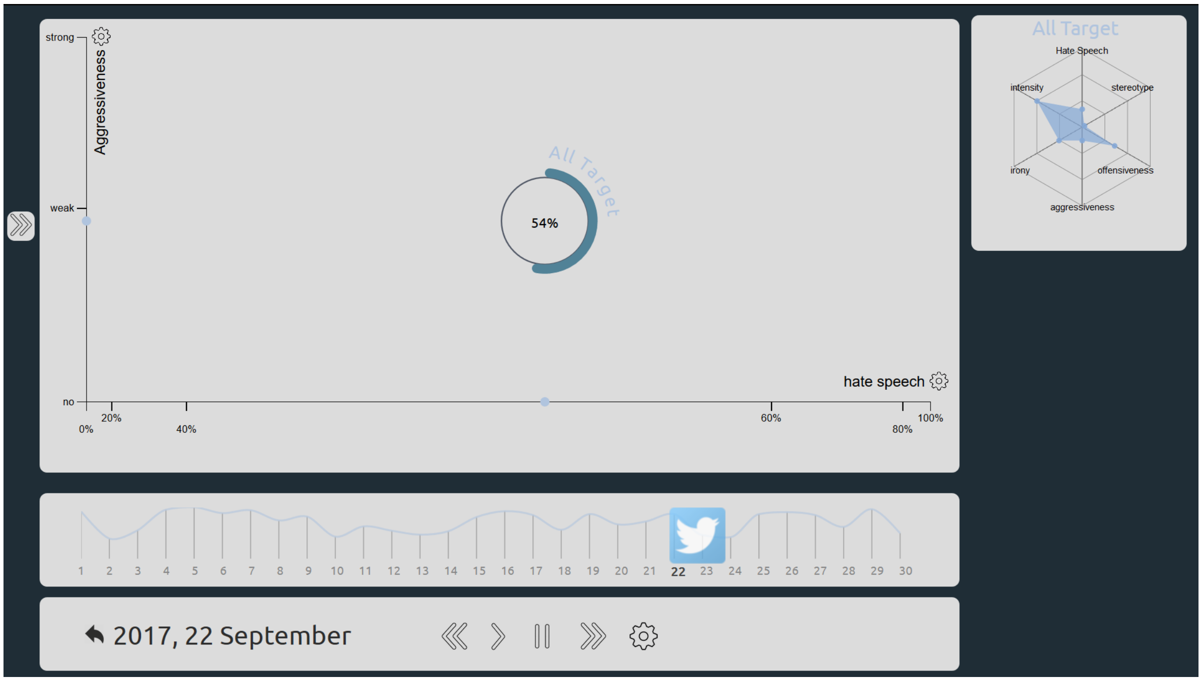


Figure 3: Main page

It is very important to keep in mind that the choice of sources, from which to extract the data to be displayed[7], can undoubtedly be very influential in the considerations that can be inferred by exploring the dashboard.

It is also important to specify that, using the functions provided by the dashboard, we could draw incorrect conclusions due to what could be called background noise. This limit will be exceeded when we will be able to compare the hate speech data related to our targets with data related to at least another hot topic of totally different nature.

We could perhaps be impressed by a percentage of hate speech of 15% in tweets aimed at a particular target, but maybe that percentage is the same that we would find for any other topic, from city traffic to too long periods of rain, in the same time period of the same geographical regions.

For these reasons we left a high level of interactivity to our platform. Users can select which attributes to compare each other using "bubble charts"-like diagrams (Fig. 3), and also focus on a particular day or time period, as well as switching to a geographical map to better observe where the majority of tweets containing hate speech are aggregated, without neglecting to observe the dynamic nature of these phenomena.

Users

Because of the two purposes we imagined for the dashboard, the users can be multiple. On the one hand a data analyst or an NLP expert can use the dashboard as a support for his/her work. On the other hand a data journalist can explore the annotated tweets in the corpus to estimate the impact on social media debates of some particular news event.

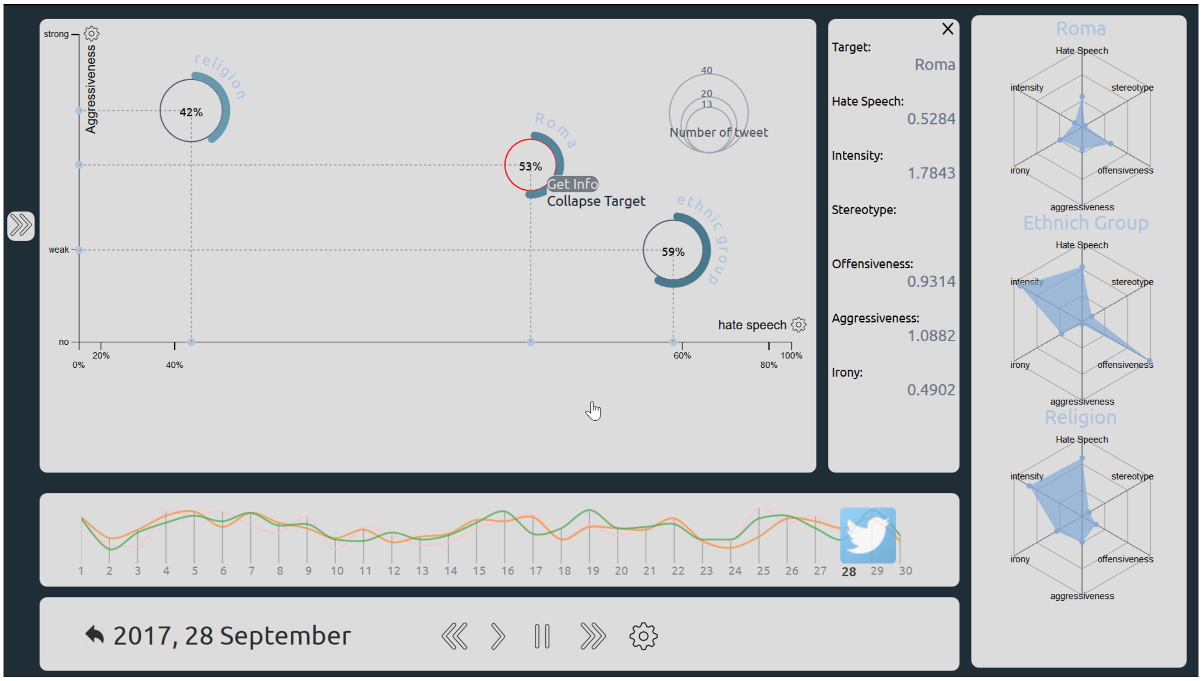
Once the platform will be completed, it will be freely accessible on line via the Web; it could also be used as a tool to raise in users the awareness of the hate speech phenomenon, for example in schools, but also to know its extent and virality. This is one of the reasons that led us to implement the dashboard exclusively with Web technologies and open standards[8] so as to make it executable in a simple browser window.

The dashboard in detail

Below we will present in detail the implementation choices adopted for the dashboard. The main page of the dashboard allows us a

[7] The case of the corpus we currently use, that is based on Twitter, not necessarily reflects other possible behavioral patterns.

[8] The visualizations are all made with the d3.js JavaScript library [3].



Target: Roma

Hate Speech: 0.5284

Intensity: 1.7843

Stereotype:

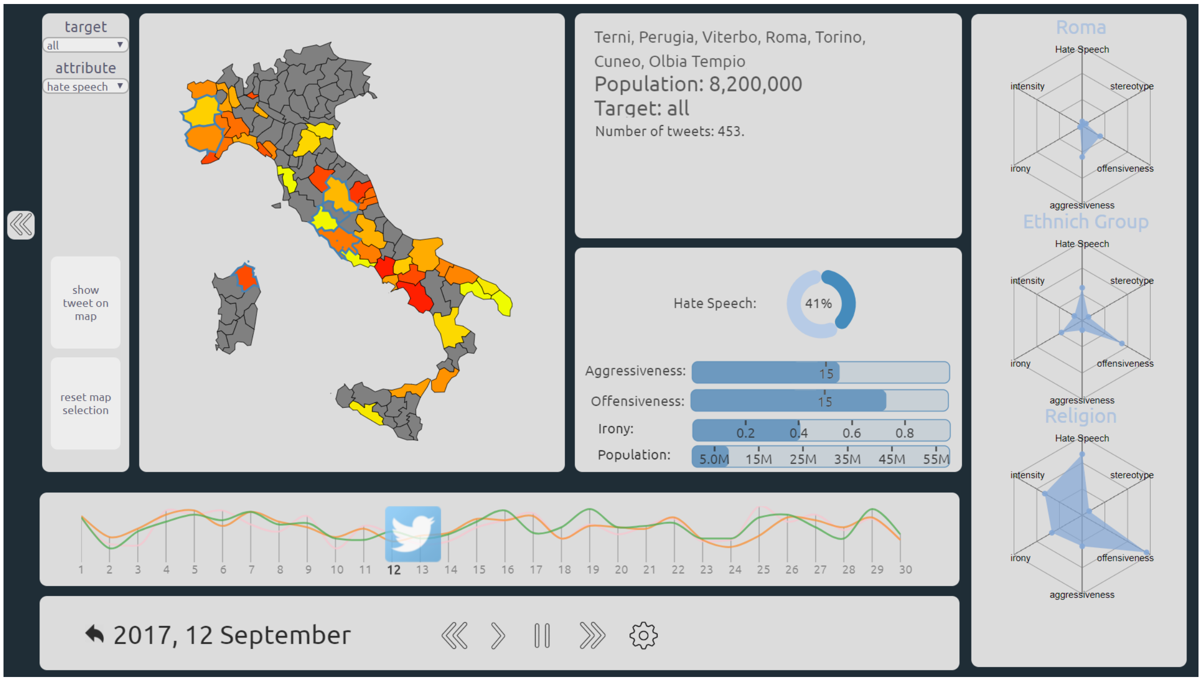
Offensiveness: 0.9314

Aggressiveness: 1.0882

Irony: 0.4902

2017, 28 September

Figure 4: Main page. All targets are represented as circles



target: all

attribute: hate speech

Terni, Perugia, Viterbo, Roma, Torino, Cuneo, Olbia Tempio

Population: 8,200,000

Target: all

Number of tweets: 453.

2017, 12 September

Figure 5: Map page

temporal analysis of the total number of tweets and the average value for each attribute.

On the other hand, the map page allows us a combined temporal and geographic analysis taking into consideration only a subset of the dataset, that of geo-referenced tweets.

All the examples of the dashboard shown in this paper are created using a random generated dataset based on the previously presented data model.

Overview page. At the bottom of the main screen (Fig. 3), there is a time slider through which the user can browse the dataset by selecting a day, a month or a year. In the central part, we have a visualization containing an overview of the aggregated data that will be updated according to the temporal selection done by the user. The temporal slider, moreover, shows, in the form of a line chart, the total number of tweets of the selected month (in relation to all the aggregate targets) giving context without losing focus.

The bubble (or circle, from now on) diagram representing the overall corpus is placed in a (x, y) space. Each axis can be associated to one of the attributes used to characterize each tweets. This means that position, area, and arc of the circle are used to represent different attributes' values.

The *hate speech, irony,* and *stereotype* attributes have boolean values, so we grouped them together (group A). Attributes *offensiveness* and *aggressiveness* have the same domain values (i.e., *no, weak, strong*) and as a consequence it makes sense if they are selectable from a different set (group B). Attributes in group A can be selected and then displayed along the x axis as well as attributes in group B are representable on the y axis (Fig. 3). Obviously each axis represents only one attribute at a time, but the user will be able to customize the viz by changing the attribute to be displayed for each axis. Hence, (x, y) position can help to compare the value of one attribute in group A with the value of another attribute in group B.

The last attribute used to characterize tweets in our corpora is *intensity*, and the length of the arc is calculated in function of the values of the tweets along this dimension.

This viz is designed to give a general view about hate speech and the total number of tweets to allow us to look for possible correlations between different attributes, or even significan outliers. For example, the position of the circle inside the Cartesian axes is calculated considering the average values of hate speech and aggressiveness occurring in the day which has been selected in the time slider. These values, in Fig. 3, refer to all our dataset targets aggregated in a single circle. Aggregating the targets in a single circle let us to evaluate the average attributes in a given period to address all the targets of the dataset. The circle is also surrounded by a donut chart which shows the average intensity level in the selected day (the length the arc mentioned above).

A radar plot on the right of the main page can help to compare the values of all the attributes in one shot (see Fig. 6).

The area of the circle is instead calculated in function of the number of tweets considered in the selected time period.

Double-clicking on the circle, we can also expand the aggregate representation to analyze the individual targets (example in Fig. 4). The line chart get expanded as well and it shows the total number of tweets for each target.

Fisheye effect applied to the axis values allows us to reduce the circles over plotting in the Cartesian plane, as we can see in figure4.

By clicking on a circle, a contextual menu appears and it lets us to open a sidebar on the right with the details about all the attributes of the selected circle. In the right sidebar in Fig. 4 there are three radar charts that show a more general view on all the attributes for each target.

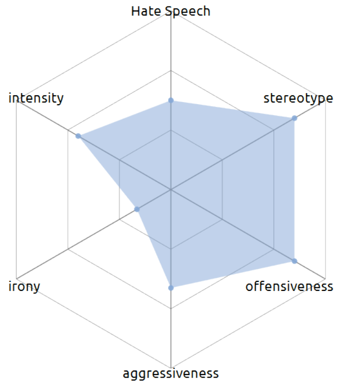
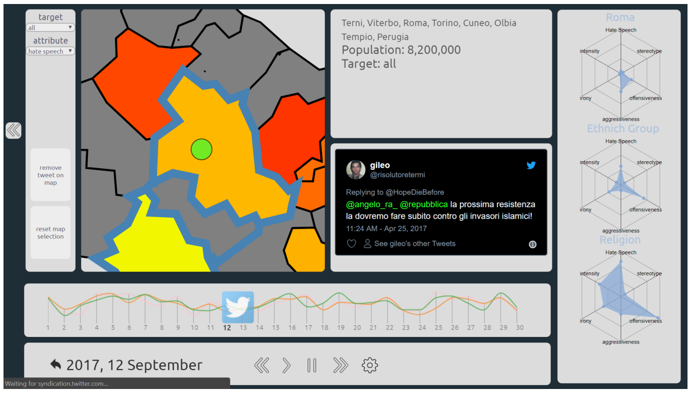


Figure 6: It is possible to compare all the attributes' values of the tweets referring to all targets considered in the corpus.

The Map View. We can switch to the map view to allow the users to further aggregate the data; in fact, in addition to the date of creation of the tweets, we can focus on messages that have been also geo-referenced. In the choropleth map in the Fig. 5, the user can display one of the attributes of the dataset, keeping a view on the average values of the other attributes relating to the selected provinces (on the right side of the map). As in the main page, it is possible to view all the targets as a single aggregate target or to analyze them individually.

Since the attributes are displayed as a percentage, we decided to visualize the total number of tweets classified (together with the number of inhabitants) for the selected provinces.



Terni, Viterbo, Roma, Torino, Cuneo, Olbia Tempio, Perugia
Population: 8,200,000
Target: all

gileo
@risolutoretermi
Replying to @HopeDieBefore
@angelo_ra_ @repubblica la prossima resistenza la dovremo fare subito contro gli invasori islamici!
11:24 AM - Apr 25, 2017
See gileo's other Tweets

2017, 12 September

Waiting for syndication.twitter.com...

Figure 7: We can visualize the contents of some geo-referenced tweets.

It is also possible to display in the map, for a small portion of the dataset, some geo-referenced tweets and show the contents as shown in Fig. 7.

At a finer granularity, we can also select a single tweet and analyze it individually. For example, radar plots comparing visually attributes values of tweets shown in Fig. 1 and 2 are displayed in Fig. 8 and 9.

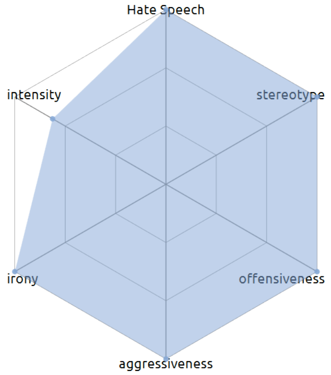


Figure 8: Radar plot representing tweet displayed in Fig. 1

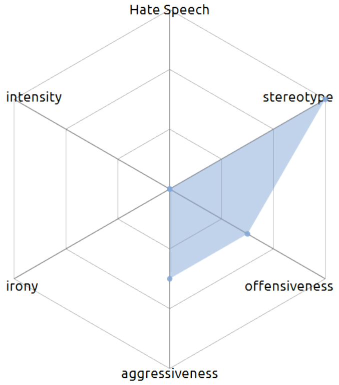


Figure 9: Radar plot representing tweet displayed in Fig. 2

CONCLUSIONS

In this paper we presented a data viz platform to support the research of a data analyst as well as the work of a data journalist whose objective is to study, analyze, and understand the hate speech phenomenon. The dashboard is highly interactive, and we solved the problem of dealing with unstructured data, such as textual ones, just dealing with attributes used to annotate the corpus. These attributes were converted to quantities; as a consequences, bubble charts, cloropleth maps, line and radar plots, and other traditional graphical tools were used to let the user explore complex corpora containing evidence of hate speech.

ACKNOWLEDGMENTS

A. Capozzi, C. Bosco, V. Patti and G. Ruffo were partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media, S1618_L2_BOSC_01*), and by Italian Ministry of Labor (*Contro l'odio: tecnologie informatiche, percorsi formativi e storytelling partecipativo per combattere l'intolleranza*, avviso n.1/2017 per il finanziamento di iniziative e progetti di rilevanza nazionale ai sensi dell'art. 72 del decreto legislativo 3 luglio 2017, n. 117 - anno 2017). The work of C. Bosco, V. Patti and G. Ruffo was also partially funded by Fondazione CRT (*Hate Speech and Social Media*, 2016.0688).

REFERENCES

[1] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1 – 8. https://doi.org/10.1016/j.jocs.2010.12.007

[2] Cristina Bosco and Viviana Patti. 2017. Social Media Analysis for Monitoring Political Sentiment. In *Encyclopedia of Social Network Analysis and Mining*. Springer, New York, NY, USA, 1–13. https://doi.org/10.1007/978-1-4614-7163-9_110172-1

[3] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3: Data-Driven Documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2011). http://vis.stanford.edu/papers/d3

[4] Jessica Chung and Eni Mustafaraj. 2011. Can Collective Sentiment Expressed on Twitter Predict Political Elections?

[5] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. https://www.microsoft.com/en-us/research/publication/predicting-depression-via-social-media/

[6] Isabel Meirelles. 2013. *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations.* Rockport Publishers.

[7] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.

[8] Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate Speech Annotation: Analysis of an Italian Twitter Corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017. (CEUR Workshop Proceedings)*, Vol. 2006. CEUR-WS.org.

[9] Erik Tjong Kim Sang and Johan Bos. 2012. Predicting the 2011 Dutch Senate Election Results with Twitter.

[10] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the 11th Conference on Language Resources and Evaluation (LREC2018), May 2018, Miyazaki, Japan.* 2798–2895.

[11] Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, 1–10. http://aclweb.org/anthology/W17-1101

[12] Ben Shneiderman. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages (VL '96)*. IEEE Computer Society, Washington, DC, USA, 336–. http://dl.acm.org/citation.cfm?id=832277.834354

[13] Frank van Ham, Martin Wattenberg, and Fernanda B. Viégas. 2009. Mapping Text with Phrase Nets. *IEEE Trans. Vis. Comput. Graph.* 15, 6 (2009), 1169–1176. https://doi.org/10.1109/TVCG.2009.165

[14] Fernanda B. Viégas, Scott Golder, and Judith Donath. 2006. Visualizing Email Content: Portraying Relationships from Conversational Histories. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 979–988. https://doi.org/10.1145/1124772.1124919

[15] Fernanda B. Viegas, Martin Wattenberg, Frank van Ham, Jesse Kriss, and Matt McKeon. 2007. ManyEyes: A Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (Nov. 2007), 1121–1128. https://doi.org/10.1109/TVCG.2007.70577

[16] Martin Wattenberg and Fernanda B. Viégas. 2008. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (Nov. 2008), 1221–1228. https://doi.org/10.1109/TVCG.2008.172