Online Social Networks and Media 6 (2018) 41–57

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.elsevier.com/locate/osnem

Who framed Roger Reindeer? De-censorship of Facebook posts by snippet classification

Fabio Del Vigna[a], Marinella Petrocchi[a,*], Alessandro Tommasi[b], Cesare Zavattari[b], Maurizio Tesconi[a]

[a] Institute of Informatics and Telematics (IIT-CNR), via G. Moruzzi 1, Pisa, Italy
[b] LUCENSE SCaRL, Traversa prima di Via della Chiesa di Sorbano del Giudice n. 231 Lucca, Italy

ARTICLE INFO

Article history:
Received 20 November 2017
Revised 6 March 2018
Accepted 9 April 2018

Keywords:
News censorship
Identities censorship
Text mining
Data correlation
Privacy in social networks
Social media mining
Facebook
Named entity recognition
Candidateentity recognition
Transparency

ABSTRACT

This paper considers online news censorship and it concentrates on censorship of identities. Obfuscating identities may occur for disparate reasons, from military to judiciary ones. In the majority of cases, this happens to protect individuals from being identified and persecuted by hostile people. However, being the collaborative web characterised by a redundancy of information, it is not unusual that the same fact is reported by multiple sources, which may not apply the same restriction policies in terms of censorship. Also, the proven aptitude of social network users to disclose personal information leads to the phenomenon that comments to news can reveal the data withheld in the news itself. This gives us a mean to figure out who the subject of the censored news is. We propose an adaptation of a text analysis approach to unveil censored identities. The approach is tested on a synthesised scenario, which however resembles a real use case. Leveraging a text analysis based on a context classifier trained over snippets from posts and comments of Facebook pages, we achieve promising results. Despite the quite constrained settings in which we operate – such as considering only snippets of very short length – our system successfully detects the censored name, choosing among 10 different candidate names, in more than 50% of the investigated cases. This outperforms the results of two reference baselines. The findings reported in this paper, other than being supported by a thorough experimental methodology and interesting on their own, also pave the way for further investigation on the insidious issues of censorship on the web.

1. Introduction

With billions of users, social media probably represent the most privileged channel for publishing, sharing, and commenting information. In particular, social networks are often adopted to spread news content [1]. According to a Pew Research study, Americans often gets news online, with a double share of them preferring social media rather than print magazines[1]. As a matter of fact, popular newspapers have an official account on social platforms. Through their pages, news stories - or their previews - are published - often under the form of a short post, with a further link to the complete text. The readers' community can like, share, and re-post news stories. Users can also comment and discuss issues in the news themselves. Still a Pew Research survey highlights that a share of 37% of social media news consumers comment on news stories, while the 31% "discuss news on the news in the site".[2] Undeniably, users' comments and discussions may help to increase the awareness and value of the published information, thanks to the addition of informative details. Examples of support are, for example, the depiction of the context in which the news facts took place, or to track down mistakes, draw rectifications, and even unveil fake information.

In this paper, we focus on online news articles and, in particular, on those news portions that organisations choose not to make public. News censorship may occur for different reasons. Organisations, be them military, commercial, governmental, or judicial, may decide to veil part of the information to protect sensitive data from, e.g., competitors, customers, or hostile entities.

* Corresponding author.
E-mail addresses: f.delvigna@iit.cnr.it (F. Del Vigna), m.petrocchi@iit.cnr.it (M. Petrocchi), alessandro.tommasi@lucense.it (A. Tommasi), cesare.zavattari@lucense.it (C. Zavattari), m.tesconi@iit.cnr.it (M. Tesconi).

[1] The Modern News Consumers, a Pew Research study: http://www.journalism.org/2016/07/07/pathways-to-news/ , July, 7, 2016 – ; All URLs in this paper have been lastly accessed on February, 20, 2018.

[2] 10 facts about the changing of the digital landscape: http://www.pewresearch.org/fact-tank/2016/09/14/facts-about-the-changing-digital-news-landscape/ , September 14, 2016.

https://doi.org/10.1016/j.osnem.2018.04.002
2468-6964/

Standard examples of censored data are identities: from a business point of view, a press agency may veil the identity of the buyer of a huge amount of fighters. Also, the names of the victims of particularly hateful offences, like rapes and abuses on minors, are typically obfuscated, as for regulations dictated by law. Finally, a peculiar practice when publishing Israeli military-related news on social media is the veiling of the identities of public officers (e.g., Corporal S., rather than the explicit identity of such officer, see, e.g., [2]). However, as highlighted by recent literature [3], given the essential nature of social networking, the "non identification alone is ineffective in protecting sensitive information". This is due to the fact that, featuring a *commented post* structure of the published news, a specific information, withheld in the news, is compromised through the effects of users' comments, where specific content may reveal, either *explicitly* or *implicitly*, that information.

This work places itself amongst a few ones, like, e.g., [2,4] that investigate to which extent the connections among news articles, comments, and social media influence the effectiveness of identities censorship procedures. In particular, we present a novel approach to unveil a censored identity in a news post, by exploiting the fact that, on the social Web, it is not unusual to find the same content, or a very similar one, published elsewhere, e.g., by another publisher with different censorship policies. Also and noticeably, as discussed above, the amount of user generated content on social networks may lead to the very unexpected phenomenon according to which the hidden information may emerge in the users' comments.

Differently from prior work in the area, which exploits the friendship network of the commenters to some censored news, here we inherit from the field of text analysis. In particular, we exploit techniques often used to address co-reference resolution [5], based on recognising the context in which certain names tend to appear, to successfully address the task of unveiling censored names. To the best of our knowledge, this is the first attempt that addresses the task by means of a semi-supervised approach, which only makes use of texts, without relying on metadata about the commenters, and trying to reconstruct missing information exploiting similar contexts. For running and validating our analysis, we make use of Facebook data, which we *purposely* censored for the sake of the experiments. Even if we rely on an experimental setting that is built *ad hoc*, our synthesised scenario is easily connectable to real use cases, as described in subsequent sections.

Our extensive experimental campaign explores a dataset of almost 40,000 posts published on the Facebook pages of the top 25 US newspapers (by weekday circulation). By exploiting an algorithm based on context categorisation, we train a classifier on the posts, and related comments, in the dataset, to demonstrate the capability to reveal the censored term. The system performances are benchmarked against two baselines, obtaining a more than significant improvement.

Summarising, the paper contributes along the following dimensions:

- the design and development of a methodology based on text-analysis, here applied for the first time to spot identities that have been censored in social media content;
- the proposal of an approach that is solely based on very loosely structured data, in contrast to other proposed techniques that leverage the social network structure. The latter have the issues that 1. the association between names and social network nodes needs to be addressed, and 2. the structure of the social network constitutes significant a-priori knowledge. Instead, we simply use raw data, by only assuming a "commented post" structure of the data;
- starting from revealing censored popular identities, our results constitute the prelude to the detection of other kind of censored terms, such as, e.g., brands and even identities of common people, whose veiling is a usual practice often applied by publishers for privacy issues, be them driven by legal, military, or business motivations.

In the next section, we first introduce real identity censorship procedures, discussing the role of comments - and commenters - in bypassing their effectiveness. Section 3 presents the data corpus for our analyses, also highlighting similarities of such corpus with the real scenarios presented in Section 2. Section 4 presents the methodology, and Section 5 describes the experiments and comments the results. In Section 6, we discuss related work in the area of investigation. Finally, Section 7 concludes the paper.

2. Identities censorship in online news and its circumvention

Moving from motivations for identities censorship in news, this section discusses the effectiveness of such censorship when news are published online [3].

Traditionally, the censorship of identities in news occurs for three main reasons: (1) business, since, e.g., it may be not advantageous to disclose the real identity of a participant in a commercial transaction, such as a large quantities of weapons; (2) legal (e.g., do not become known minors abused); and (3) military, to protect the individuals, and their relatives, from being identified by adversaries. As an example, in Israeli " policy dictates many situations in which the identity of officers must not be released to the public [3]. In the above circumstances, the censorship usually takes place either by putting an initial or using a fancy name.

With the advent of the social media era, the publication of news on social networks sites became a usual pratice. Thus, when the news is published on social networks sites, such as on the Facebook pages of newspapers, the identities are still blurred as written above, directly from the organisation that chooses not to publish that information (therefore, either the government, or the news agency that publishes the news, or some other military or commercial stakeholder).

However, when the news post is on the social network, and a "commented post" structure of the data is followed, the comments are freely posted by users other than the news publisher. Also, comments are generally not moderated by the platform administrators, unless they are reported as offensive content, inciting hate campaigns, or some judicial authority required the cancellation of specific comments.

The fact that there are a number of uncensored comments leads to the phenomenon that, although in the post the information is withheld, that information is compromised by one, or more, comments. In fact, it has been proven that, although the organisation that posted the news has censored an identity in the news itself, and so published it, those people who know the censored name and who make comments tend to talk about the name, indirectly or even directly. This is the case featured, e.g., by the Facebook dataset analysed in [3], where 325,527 press items from 37 Facebook news organisation pages were collected. A total of 48 censored articles were identified by a pattern matching algorithm first, and then manually checked. On the whole amount of comments tied to those articles, the 19% of them were classified as comments presenting an explicit identification of the name or the use of a pseudonym. A de-censorship analysis based on the social graph of the commenters has been carried out to recognise the censored names [2]. In the rest of the paper, we will propose a methodology based instead of recognising the textual context in which certain terms tend to appear.

To test the methodology, we rely on a Facebook dataset that we intentionally censored, by however resembling the real scenarios depicted above. We deliberately censored identities in a Face-