Do Violent People Smile? Social Media Analysis of their Profile Pictures

Mauro Coletto
Ca' Foscari University
Venice, Italy
mauro.coletto@unive.it

Claudio Lucchese
Ca' Foscari University
Venice, Italy
claudio.lucchese@unive.it

Salvatore Orlando
Ca' Foscari University
Venice, Italy
orlando@unive.it

ABSTRACT

The popularity of online social platforms has also determined the emergence of violent and abusive behaviors reflecting real life issues into the digital arena. Cyberbullying, Internet banging, pedopornography, sexting are examples of these behaviors, as witnessed in the social media environments.

Several studies have shown how to approximately detect those behaviors by analyzing the social interactions and in particular the content of the exchanged messages. The features considered in the models basically include detection of offensive language through NLP techniques and vocabularies, social network structural measures and, if available, user context information.

Our goal is to investigate those users who adopt offensive language and hate speech in Twitter by analyzing their profile pictures. Results show that violent people smile less and they are dominating by anger, fear and sadness.

KEYWORDS

cyberbullying, violence, social media, offensive language, smile, profile pictures, face++, Twitter, emotion

ACM Reference Format:
Mauro Coletto, Claudio Lucchese, and Salvatore Orlando. 2018. Do Violent People Smile? Social Media Analysis of their Profile Pictures. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3184558.3191594

1 INTRODUCTION

The use of online social networks and micro-blogging platforms as a source to detect and quantify social phenomena is a recurrent task, in particular in the field of Social Network Analysis and Computational Social Science. Several studies can be found in different contexts using social media to study collective phenomena: from pandemics detection [12] to political elections prediction [6], from misinformation spreading [3] to migration analysis [5].

A relevant context where social media have a strong role are abusive behaviors [1, 4, 14] related to offensive language adopted by users in their virtual interactions. Many studies focused on the detection of cyberbullying, violence, internet banging through the analysis of conversations and messages in social media by mean of text mining and machine learning models [10, 13]. While additional features related to the structure of the social network and the user context can be used to improve model accuracy [7, 11, 17], the automatic detection of offensive messages is a critical issue which involves the use of lexical resources (profanity dictionaries), sentiment analysis techniques, NLP processing methods and meta information.

In [8] the authors propose an interesting machine learning approach based on different features (uni-grams, bi-grams, tri-grams, POS tags, Flesch-Kincaid Grade Level and Flesch Reading Ease scores, sentiment and other metrics) to detect offensive language and hate speech.

We adopt this method to detect users utilizing offensive language and hate speech in order to analyze their profile pictures to describe visual features of the typical violent users.

To study the visual traits of the profile picture we use a computer vision approach [21] used in many recent works. In the context of Social Media this method has recently been used to study the categories in which selfies appear on Instagram [9], cultural trends in Facebook photographs [19] or different profile pictures characteristics in different Social Media [20].

2 DATA

We explore two different Twitter collections to be less dataset dependent. The two datasets were collected in different periods with a large temporal gap, thus allowing us to also evaluate possible evolution through time. We also use a dataset of news, and a vocabulary of bad words.

TwitterA. We use the Twitter dataset released by CAW2.0 (Content Analysis in Web 2.0) workshop in WWW 2009 conference which has been widely used in the context of cyberbullying detection [11, 16]. The corpus contains $\approx 977k$ tweets written in English by $\approx 27k$ unique users from Dec 2008 to Jan 2009.

TwitterB. We use a second Twitter dataset, containing $\approx 1M$ tweets written in English by $\approx 643k$ unique users collected through the Twitter APIs in December 2015.

NewsC - Additional dataset. We use a dataset containing $1M$ news, released by Signal Media (NewsIR'16 workshop). The purpose of the dataset in only to reinforce the validity of the machine learning model used. The news, mainly in English, were originally collected from a variety of news sources and blogs for a period of 1 month (1-30 September 2015).

Profanity vocabulary. To create a comprehensive dictionary of bad words, we use the following online sources:

- list of offensive expressions from hatebase.org (1000+)
- list of swear words from bannedwordlist.com (70+ terms)
- list of English terms that could be found offensive by Luis von Ahn (1300+ terms)

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191594

Table 1: Classification of the filtered tweets

dataset	class	percentage	tweets
TwitterA	0	2%	3.5k
	1	24%	34k
	2	74%	106k
TwitterB	0	5%	8.9k
	1	30%	50k
	2	65%	109k

- list of bad words to filter social media content from FrontGate (700+ terms)

By merging them, we obtain a list of 2704 single expressions/terms which can be used to filter social media content and news in order to find offensive content.

3 METHODOLOGY

Hate speech language has been widely studied in the NLP community [15, 18], despite the existence of a persistent critical point in using dictionaries and lexical sources containing terms that are related to hate speech in specific contexts only, but that can have neutral or positive meanings in other contexts. For this reason we adopted both a dictionary approach combined with the machine learning (ML) method proposed by [8] to detect offensive tweets. Specifically, we applied a filter to the collected tweets based on the profanity vocabulary described above. By filtering the datasets we selected: for TwitterA 15% of the dataset (143k documents), for TwitterB 17% (168k documents). After this filtering phase, we applied to the resulting tweets the ML-based method to detect hate speech and offensive language. The ML model is based on a logistic regression with L2 regularization. The features considered in that work are: bi-grams, uni-grams, and tri-grams features (weighted by TF-IDF), Penn Part-of-Speech (POS) tags, Flesch-Kincaid Grade Level and Flesch Reading Ease scores, sentiment lexicon and other general indicators (characters, words, syllables, hashtags, mentions, replies, retweets). The model discriminates among hate speech tweets (**class 0**), offensive tweets (**class 1**), and neutral tweets (**class 2**). We refer to the tweets classified as class 0 or class 1 as *violent* ones, and the tweets of class 2 as *non violent* ones. We trained the ML model on the dataset proposed in the original paper with cross validation. The model is very accurate: 94% in both precision and recall. We considered the same features described in the original papers, among which 10k most frequent terms (if-idf) and 5k pos tags. By applying the learned ML model to our filtered datasets, we classified tweets as described in Table 1. The high percentage of neutral tweets indicates the limitation of the vocabulary approach in identifying properly violent content. To reinforce the validity of the process we applied both the filtering and the ML method to the NewsC dataset, by splitting the news in sentences with a size comparable to a tweet. The filtering excluded 88% of the sentences. Applying to the remaining 12% the ML method, we obtained that 96% of the remaining corpus still resulted neutral, highlighting differences in the violent tones of the content from official newspapers and from Social Media, that are not easily detected by a vocabulary based method. This evidence reinforces the validity of the ML model, which looks not only at terms frequency, in detecting offensive content.

4 PROFILE PICTURE ANALYSIS

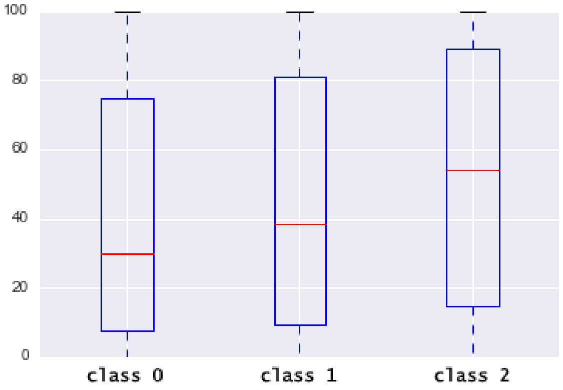


Figure 1: Smile index distribution on TwitterA

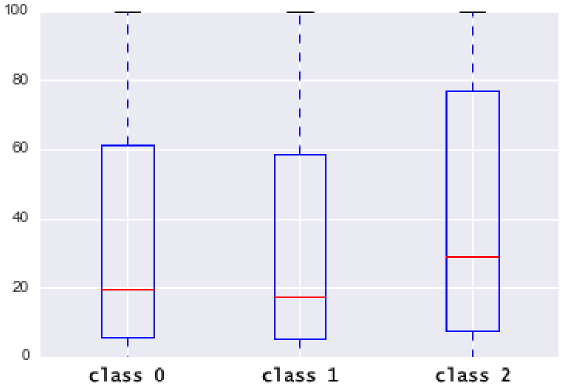


Figure 2: Smile index distribution on TwitterB

We used the publicly available API developed by Face++, a cloud-based facial recognition system. Face++ is a service providing highly accurate user information inferred by their profile pictures. Given a picture containing a face, the Face++ algorithm extracts information concerning demographics of the individual, as well as the emotions of the detected face. Face++ reports a 99.5% accuracy on an established facial recognition benchmark [21]; this accuracy is further supported by the results in [2], which reports a 97% +/- 5% accuracy on similar photos. Through Face++, we thus collected demographic information (gender, age), ethnicity, smile intensity and emotions for 13k users (for TwitterA), 18k users (for TwitterB). Face++ is able to identify emotions including confidence scores for anger, disgust, fear, happiness, neutral, sadness, and surprise.

Table 2: Demographic Analysis

dataset	class	ratio M/F	avg. age
TwitterA	0	1.12	44
	1	1.24	45
	2	1.26	46
TwitterB	0	1.04	40
	1	0.75	37
	2	0.78	40

Table 3: Ethnicity

dataset	class	white	black	asian
TwitterA	0	0.71	0.12	0.17
	1	0.71	0.11	0.18
	2	0.76	0.09	0.16
TwitterB	0	0.54	0.24	0.22
	1	0.53	0.12	0.23
	2	0.64	0.13	0.23

4.1 Demographic Analysis

Table 2 shows the results of the demographic analysis. For lack of space we do not report the full age distributions by gender, but we highlight the most relevant evidences. In both datasets violent people are on average younger, this reinforcing the possible presence of so-called cyberbullies, usually teenagers, who determine a decrease of the average age of the group. The class 1 is characterized by a higher percentage of female users compared to the general population, indicating that using violent language is not predominantly a male behavior. Less evidences can be underlined on hate speech, since the clusters are smaller and less statistically significant.

4.2 Ethnicity

Table 3 reports the results of the ethnicity analysis. No significant differences in the ethnicity of the users in class 0 and 1, but if we look to the differences between violent users and neutral we see that there is an increase of black people compared to white ones. We measured the statistical significance through the t-test and the evidences are confirmed with a predominance of black people being statistically significant among violent users (p-value:0.0015). Differences in the Asian portion are instead not statistically significant. We explored additional features, for instance the presence of glasses and sun glasses on the faces, but there are no statistically significant evidences of a different behavior among people who wear them or not.

4.3 Analysis of Emotions

By analyzing the profile images of the users we can compute a smile index which indicates the intensity of the smile in the picture (scale 0-100: 0 concave curve, 100 extreme smile). In Figure 1 and in Figure 2 we report the boxplot of the smile index for the considered profile pictures. The means in the boxplot are different for the various classes, and the results are statistically significant (t-test with a p values less than 10^6). They show an interesting trend: violent users have a low smile index compared to neutral users whose smile index is higher on average. Through Face++, we also collected information about the emotions of the users depicted in the analyzed profile pictures. Figure 3 reports the emotions analyzed for TwitterA and TwitterB datasets. The results for the two datasets are consistent. Violent people are characterized on average by statistically significantly higher values in anger, fear and sadness. Also the feeling of surprise is higher for violent users, in particular in the first dataset, showing excitement. On the other hand, the feeling of happiness is much more present among users not detected as violent or offensive compared to users belonging to class 2. The difference in happiness between violent and nonviolent users is particularly relevant. The two datasets are different for collection, period, users, but still the results between violent people and neutral ones are consistent, thus indicating the validity of the evidences. We conclude by remarking that there is a correlation between the information extracted from profile pictures and the user violent behavior. Further social and psychological analysis is required to understand if the features highlighted in violent users profiles are caused by an explicit will of exhibiting an aggressive attitude, or whether this is an implicit outcome of a specific group of individuals, or a life conduct.

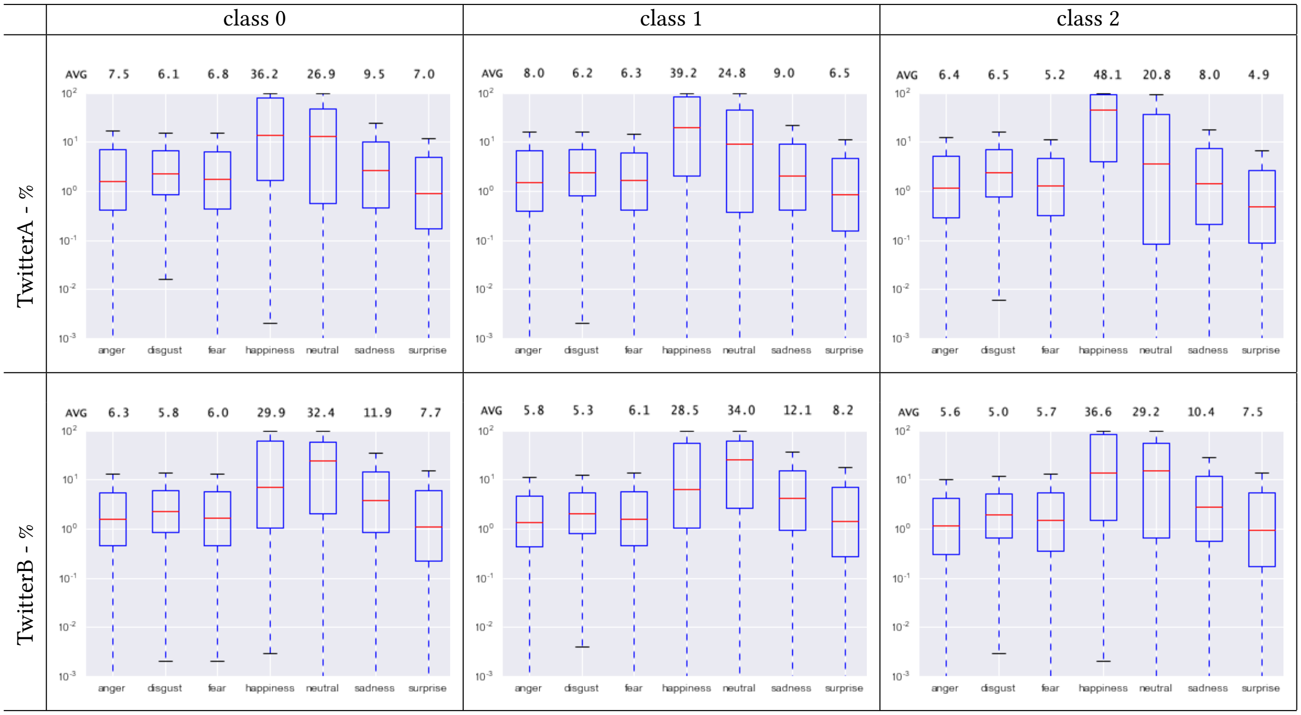
5 CONCLUSION

To the best of our knowledge, this is the first study on profile pictures regarding violent people in Social Media. We adopted a double approach by using an ad hoc vocabulary of profanities, built by merging different sources, and a recent ML model able to detect both hate speech and offensive communication. We applied the two methods in sequence, in order to detect violent users to be further analyzed through their profile pictures. To validate the generality of the method, we considered two datesets collected in different periods (2009 and 2015), containing about 1 M tweets each. The analysis of the profile picture was based on 13k users for the first dataset, and 18k users for the second. Results show that violent users are younger, with a higher percentage of female users adopting offensive language. As regards ethnicity, there is a higher concentration of black people among violent users. One reason might be that especially in US the language spoken by low class society is rich of offensive slang expressions. Moreover aggressive users smile less, and they appear not happy in their profile pictures, dominated by fear, sadness and anger. These feelings are both consequence of their aggressiveness, and probably also of an unconscious desire to appear more violent.

REFERENCES

[1] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. 2016. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior* 63

Figure 3: Analysis of Emotions



(2016), 433–443.

[2] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. 2014. Faces engage us: Photos with faces attract more likes and comments on instagram. In *ACM Human factors in computing systems.* 965–974.

[3] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one* 10, 2 (2015), e0118093.

[4] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Detecting Aggressors and Bullies on Twitter. In *Proceedings of the 26th International Conference on World Wide Web Companion.* International World Wide Web Conferences Steering Committee, 767–768.

[5] Mauro Coletto, Claudio Lucchese, Cristina Ioana Muntean, Franco Maria Nardini, Andrea Esuli, Chiara Renso, and Raffaele Perego. 2016. Sentiment-enhanced Multidimensional Analysis of Online Social Networks: Perception of the Mediterranean Refugees Crisis. In *ASONAM 2016.*

[6] M Coletto, C Lucchese, S Orlando, and R Perego. 2015. Electoral Predictions with Twitter: a Machine-Learning approach. In *IIR 2015, Cagliari, Italy.*

[7] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval.* Springer, 693–696.

[8] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *ICWSM 2017* (2017).

[9] Julia Deeb-Swihart, Christopher Polack, Eric Gilbert, and Irfan A Essa. 2017. Selfie-Presentation in Everyday Life: A Large-Scale Characterization of Selfie Contexts on Instagram.. In *ICWSM.* 42–51.

[10] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying. *The Social Mobile Web* 11, 02 (2011).

[11] Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber Bullying Detection Using Social and Textual Analysis. In *Intl. Workshop on Socially-Aware Multimedia, ACM SAM 2014.* 3–6.

[12] Vasileios Lampos, Tijl De Bie, and Nello Cristianini. 2010. Flu detector-tracking epidemics on Twitter. In *Machine Learning and Knowledge Discovery in Databases.* Springer, 599–602.

[13] Parma Nand, Rivindu Perera, and Abhijeet Kasture. [n. d.]. âĂIJHow Bullying is this Message?âĂİ: A Psychometric Thermometer for Bullying. ([n. d.]).

[14] Desmond Upton Patton, Robert D Eschmann, and Dirk A Butler. 2013. Internet banging: New trends in social media, gang violence, masculinity and hip hop. *Computers in Human Behavior* 29, 5 (2013), A54–A59.

[15] Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. *SocialNLP 2017* (2017).

[16] Vivek K Singh, Qianjia Huang, and Pradeep K Atrey. [n. d.]. Cyberbullying detection using probabilistic socio-textual information fusion. In *IEEE/ACM ASOMAN 2016.* 884–887.

[17] A Squicciarini, S Rajtmajer, Y Liu, and Christopher Griffin. 2015. Identification and characterization of cyberbullying dynamics in an online social network. In *IEEE/ACM ASOMAN 2015.* 280–285.

[18] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2014. Cursing in english on twitter. In *ACM conf. on Computer supported cooperative work & social computing.* ACM, 415–425.

[19] Quanzeng You, Darío García-García, Mahohar Paluri, Jiebo Luo, and Jungseock Joo. 2017. Cultural Diffusion and Trends in Facebook Photographs. *ICWSM 2017* (2017).

[20] Changtao Zhong, Hau-wen Chan, Dmytro Karamshu, Dongwon Lee, and Nishanth Sastry. 2017. Wearing Many (Social) Hats: How Different are Your Different Social Network Personae? *ICWSM 2017* (2017).

[21] Erjin Zhou, Zhimin Cao, and Qi Yin. 2015. Naive-deep face recognition: Touching the limit of LFW benchmark or not? *arXiv preprint arXiv:1501.04690* (2015).