Mining Offensive Language on Social Media

Alessandro Maisto, Serena Pelosi, Simonetta Vietri, Pierluigi Vitale
University of Salerno
Department of Political, Social and Communication Science
Via Giovanni Paolo II, 132
{amaisto,spelosi,vietri,pvitale}@unisa.it

Abstract

English. The present research deals with the automatic annotation and classification of vulgar ad offensive speech on social media. In this paper we will test the effectiveness of the computational treatment of the taboo contents shared on the web, the output is a corpus of 31,749 Facebook comments which has been automatically annotated through a lexicon-based method for the automatic identification and classification of taboo expressions.

Italiano. *La presente ricerca affronta il tema dell'annotazione e della classificazione automatica dei contenuti volgari e offensivi espressi nei social media. Lo scopo del nostro lavoro consiste nel testare l'efficacia del trattamento computazionale dei contenuti tabù condivisi in rete. L'output che forniamo un corpus di 31,749 commenti generati dagli utenti di Facebook e annotato automaticamente attraverso un metodo basato sul lessico per l'identificazione e la classificazione delle espressioni tabù.*

1 Introduction

Flaming, *trolling*, *harassment*, *cyberbullying*, *cyberstalking*, *cyberthreats* are all terms used for referring to vulgar and offensive contents shared on the web. The shapes can be different and the focus can be on various topics, such as physical appearance, ethnicity, sexuality, social acceptance and so forth.

Although taboo language is generally considered to be the strongest clue of harassment in the web, it must be clarified that the presence of bad words in posts does not necessarily indicate the presence of offensive behaviors. The words collected in vulgar lexicons, in some cases, are neutral or even positive. Moreover, profanity can be used with comical or satirical purposes, and bad words are often just the expression of strong emotions (Yin et al., 2009).

In this paper, we propose a system for the automatic treatment of vulgar and offensive utterances in Italian. The strength of our method is that lexical items are not considered in isolation. Instead, we recognize the power of the local context of the words, which can modulate the meaning of words, phrases and sentences.

Section 2 briefly illustrates the state of the art contributions on offensive language modeling. Next, Section 3 describes the Italian lexical and grammatical resources for the automatic detection of taboo language in Italian. Then, Section 4 explains how we tested our method and resources on a Facebook corpus and describes the results of the taboo expressions automatic annotation. Finally, Section 5 reports the future works that will enhance our research.

2 State of the Art on the Computational Treatment of Offensive Language

As it is anticipated, taboo words are basically considered a strong clue of online hate speech (Chen et al., 2012; Reynolds et al., 2011; Xu and Zhu, 2010; Yin et al., 2009; Mahmud et al., 2008). Nevertheless, the methods that simply match offensive words stored in blacklists, are clearly not meant to reach high levels of accuracy. Consistent with this idea, in the recent years many studies on offensive cyberbullying and flame detection integrated the bad words context in their methods and tools. Chen et al. (2012) exploited a Lexical Syntactic Feature (LSF) architecture to detect offensive content and identify potential offensive users in social media. Xu and Zhu (2010) proposed a sentence-level semantic filtering approach that combined grammatical relations with offensive words. Insulting phrases and derogatory comparisons of human beings with insulting items or animals were clues used by Mahmud et al. (2008). Razavi et al. (2010) proposed an automatic flame detection method based on the variety of statistical models and the rule-based patterns. Among the flame topics that they identified, there are attacks and abuses that embarrass the readers. Xiang et al. (2012) learned topic models from a dataset of tweets through Latent Dirichlet Allocation (LDA) algorithm. Waseem and Hovy (2016) and Kwok and Wang (2013) focused on racist and sexist slurs on Twitter; Waseem and Hovy (2016) made reference to hate speech expressed without any derogatory term, and Kwok and Wang (2013) focused on the relation between the tweet content and the identity of the user, on the base of which a post is considered to be racist or not. Badjatiya et al. (2017) also used Twitter in order to investigate the application of deep neural network architectures.

3 Lexical and Grammatical Resources

In this paragraph we will describe the Italian lexical database and the grammatical rules which have been used as indicators for the automatic identification of the taboo language.

The items of the lexicon are labeled through the use of the following three main categories:

- *Trait*, that specifies if the taboo expression is addressed to other users, to events or to things;
- *Type*, that verifies if an expression is *offensive*, if it represents a *threat* or if it is just *rudeness*;
- *Semantic Field* which specifies the taboo domain (namely *sex, sexism, aesthetics, behavior, homophobia, racism, scatology*).

Such tags have been collected and classified by a team of four annotators (one linguist and three Italian native speakers), which annotated the linguistic resources through an agreement of 92%.

Taboo words, which were impossible to classify through a defined semantic field, have been annotated with the residual category "N.C.". Our taboo lexicon is composed of

- *Simple Words*, which include nouns, adjectives, verbs and adverbs collected from the Sentiment Lexicon *SentIta* (Pelosi, 2015) and manually evaluated with reference to the categories described above;
- *Multiword Expressions* (MWE), that are nouns automatically annotated through the integrated use of the simple words list and *ad hoc* regular expressions (e.g. see section 3.2);
- *Idiomatic Structures*, which are verbs + frozen complement collected from Vietri (2014) and manually annotated on the grounds of the hate speech tags.

This choice is due to the fact that in colloquial and informal situations, a taboo expression can work simply as intensifier, also for positive sentences (e.g. *it's fucking nice*!). This is why the words' semantic orientation must be, case by case, modulated when occurring into the context of (semi)frozen structures. Concrete examples are idiomatic structures that involve concrete nouns indicating body part (with a vulgar meaning) as fixed constituents (e.g. *essere culo e camicia*, "to be thick as thieves").

3.1 Simple Vulgar Words

Our project is grounded on a collection of 342 taboo simple words that include the following grammatical categories: nouns, verbs, adjectives, adverbs and exclamations. Nouns count 242 entries, among which 216 are simple words (e.g. *cozza*, "mussel", addressed to ugly women) and 26 are monorematic compounds (e.g *rompiballe* "pain in the ass"). Verbs count 72 entries, among which 27 are verbs indicating bodily predicates that involve acts of violence, e.g. *violentare*, "to rape", and 21 are pro-complementary and pronominal verbs (e.g. *incazzarsi* "to get mad"). Adjectives count 16 entries (e.g. *cazzuto* "die-hard"), adverbs 4 entries (e.g. *incazzosamente* "grumpily") and exclamations 8 entries (e.g. *vaffanculo* "fuck off").

3.2 Taboo Multiword Structures

The simple words listed in our database, especially the ones with an uncertain semantic orientation (see "N.C. in Figure 1"), can be part of frozen or semi-frozen expressions that can make clear, for each occurrence, the actual meaning of the words in context.

Idioms are particularly interesting in a work on online harassment, because they are open to wordplays and trolls. Indeed, it must be reported a higher than expected presence of idiomatic structures in our corpus. Nevertheless, their syntactic flexibility and the lexical variations make them very difficult to automatically locate, if compared with other multiword expressions. A very typical Italian example is *cazzo* "dick", with its, more or less vulgar, stilistic and regional variants (e.g. *minchia*, *pirla*, *cavolo* "cabbage", *cacchio* "dang", *mazza* "stick" , *tubo* "pipe", *corno* "horn", etc...). The context systematically gives the word under examination a clear connotation. Examples are (negative) adverbial and adjectival expressions (e.g. *a cazzo*, "fucked up"); (emphatic) exclamations and interrogative forms (e.g. *che cazzo* "what the hell"); intensification of negations (e.g. *non* V *un cazzo*, "don't *V* shit").

Multiwords Expressions. With Multiwords Expressions, we mean sequences of simple words separated by blanks, characterized by semantic atomicity, restriction of distribution, shared and established use and lack of ambiguity. In this research, we automatically located and annotated MWEs through the combined use of the taboo simple words that trigger the recognition and a set of regular expressions (based on part of speech patterns) that locate the MWEs (e.g. *culo rotto*, "lucky" from the simple noun *culo* and the pattern *NA*). Other MWEs are those ones related to idioms (see next paragraph, e.g. *rottura di palle* "nuisance"). The regular expressions used to identify the taboo MWEs are summarized below:

- *Taboo Noun + Preposition + Noun* (NPN)
- *Noun + Preposition + Taboo Noun* (NPN)
- *Taboo Noun + Adjective* (NA)
- *Noun + Taboo Adjective* (NA)
- *Adjective + Taboo Noun* (AN)
- *Taboo Adjective + Noun* (AN)

Idiomatic Expressions. Among the possible idiomatic structures, the present research focuses on those idioms (verb and at least one frozen complement) which have vulgar nouns of body part as frozen complement.

The lexical resources used in this research are composed of 52 items that include 28 ordinary verb structures (e.g. *girare le palle* "to bust the balls") and 23 support verb idioms (*avere culo* "to be lucky"). The classes to whom they belong (Vietri, 2014) are various and can be in systematic correlation, as it happens with *girare le palle/avere le palle girate*, "to bust the balls/to have the balls busted".

The idioms under examination can be also related to some derived nominals in *-tore,-trice,-ura,-ata* (e.g. *rottura di palle* "pain in the arse") and/or with VC compounds (verb + fixed constituent e.g. *rompipalle* "ball-buster"). These compounds occur in the corpus as both simple words and multiword units.

The automatic recognition of taboo idioms, similar to MWEs, start from the nouns indicating vulgar body parts, and proceed with another lexical anchor that is associated to the idiom in the lexical resources (e.g. *girare le palle* "to piss off" is annotated in the corpus when the tool locates at the same time *palle* e *girare* with a maximum distance of three word forms). This procedure streamlines the automatic recognition of the idioms, guaranteeing high levels of recall in spite of the large variety of syntactic transformations that the frozen structure can go through (causative constructions, infinitive forms preceded by *da*, dislocation, modification, among others (Vietri, 2014)).

4 Experiment and Evaluation

The linguistic resources described so far have been tested on a large corpus of User-Generated-Contents scraped by Facebook. We chose an Italian Facebook page called *Sesso Droga e Pastorizia*, which became popular for its explicit and offensive contents. The page has been shut down the 10/03/2017 for the social network policy violation; therefore, the page's administrators created a set of connected pages in order to continue the activity in case of temporal or definitive closing. For our experimentation, posts and comments have been extracted from three pages correspondent to the following indices: *sessodrogapastorizia1*, *sessodrogapastorizia3* and *sessodrogapastoriziariserva*. The corpus includes 31,749 comments published between 28 March 2017 and 13 April 2017 by over 20 thounsand users, replying to 122 status. We extracted 2,797 taboo expressions with a Recall of 97% and a Precision of 83% by applying dictionaries and grammars to the generated corpus[1].

Figure 1 represents a bubble chart which illus-

[1] The Recall has been evaluated on the entire corpus of over 31,000 comments, while the Precision has been calculated on the extracted 2,700+ sentences.

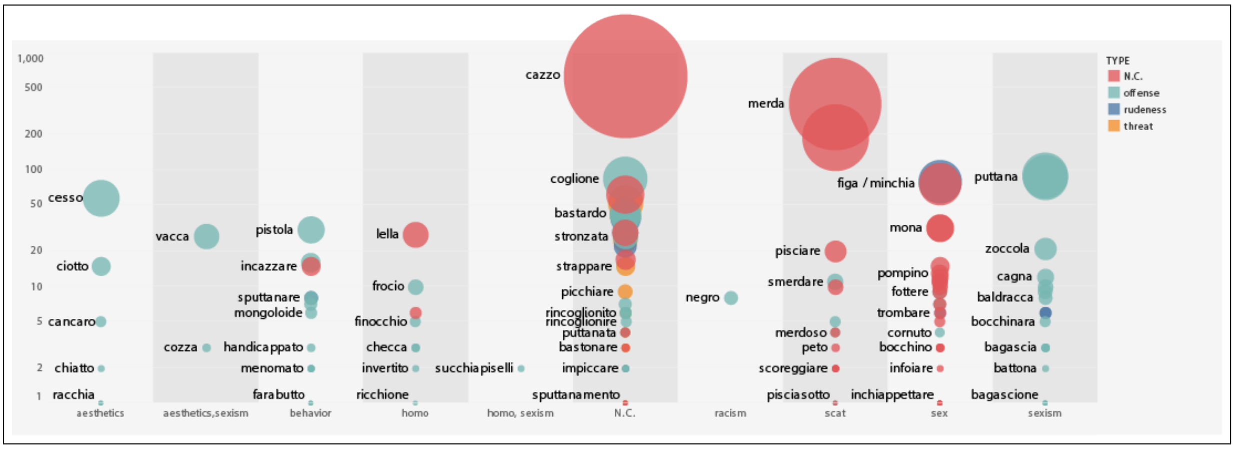

Figure 1: Extracted words occurrence, types and fields

trates the distribution of *Semantic Fields* and *Types* of the extracted words. The fields are listed in the horizontal axis. The vertical axis and the size of the circles describe together the frequency of the extracted items. Finally, the colors of the circles represent the words' types.

As far as MWEs are concerned, we extracted 134 idioms; moreover 597 MWEs have been annotated as NPN structures, 213 as NA and 175 as AN. Among the most frequent MWEs we mention some items which were already listed into the dictionaries (e.g. 9 occurrences of *testa di cazzo*, "dickhead"; 7 occurrences of *pezzo di merda*, "piece of shit").

Also new vulgar structures belonging to various fields have been automatically located through our strategy (e.g. 51 occurrences of *cazzo duro*, "hard-on" from the field *sex*; 3 occurrences of *gran troia* "total slut" from the *sexism* field; 2 occurrences of *busta di piscio* "box of piss" from the *scat* field).

The extracted patterns underline the relevance of the local context in the disambiguation of some words which have classified N.C. as simple words, because of their ambiguity out of the context. An example is *cazzo* which, alone, did not receive any field or type label, but as a MWE clearly belongs to defined categories. *Cazzo duro* belongs to the *sex* field. *Cazzo di + Noun* "this fucking + N" is a generic offense (e.g. *cazzo di pagina* "this fucking page") and *cazzo di + Taboo Noun* represents an intensification of the expressed offensive term (e.g. *cazzo di zingaro* "this fucking gypsy").

5 Conclusion

In this paper we described an experiment on the detection and classification of offenses, threats and insults shared through User Generated Contents.

As a matter of fact, in May 2016, the European Commission, together with companies like Facebook, Twitter, YouTube and Microsoft, underlined the relevance of these topics by presenting a code of conduct[2] which aimed to constrain the virality of illegal online violence and hate speech, with a special focus on utterances fomenting racism, xenophobia and terrorist contents. The negative impact of such practices is not limited to individuals, but strongly affects the freedom of expression and the democratic discourse on the Web.

Our research focused on a particular Facebook page, which became famous in Italy for the number of times it has been shut down due to its disturbing content. More than 31,000 users' comments downloaded from this page have been automatically annotated according to a dataset of taboo expressions, in the form of simple words and multiword expressions. This operation has led to a hate speech annotated corpus which distinguishes eight harassment *semantic fields*, four *types* of insult and four hate targets (*traits*). The evaluation of the experiment performances confirmed the hypothesis that the local context of words represents an essential feature for an effective hate speech mining on the web.

In future works we will test the interaction of the taboo item located in the corpus with some Italian Contextual Valence Shifters (Maisto and Pelosi, 2014) in order to verify if the sentence context of the insult indicators affects the semantic orientation of the items into an Opinion Mining

[2] `http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf`

view.

Furthermore, it would be interesting to verify the efficacy of our resources and our method on different domains, Political Communication, among others.

In the end, just because the automatic extraction has been done in this paper on a very polarized corpus, future analyses will focus on testing the reliability of this research on more neutral collections of texts.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80. IEEE.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *AAAI*.

Altaf Mahmud, Kazi Zubair Ahmed, and Mumit Khan. 2008. Detecting flames and insults in text. In *Proceedings of the Sixth International Conference on Natural Language Processing*. BRAC University.

Alessandro Maisto and Serena Pelosi. 2014. A lexicon-based approach to sentiment analysis. the italian module for nooj. In *Proceedings of the International Nooj 2014 Conference, University of Sassari, Italy*. Cambridge Scholar Publishing.

Serena Pelosi. 2015. Sentita and doxa: Italian databases and tools for sentiment analysis purposes. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 226–231. Accademia University Press.

Amir Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. *Advances in Artificial Intelligence*, pages 16–27.

Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 241–244. IEEE.

Simonetta Vietri. 2014. *Idiomatic Constructions in Italian: A Lexicon-grammar Approach*, volume 31. John Benjamins Publishing Company.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW@ HLT-NAACL*, pages 88–93.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984. ACM.

Zhi Xu and Sencun Zhu. 2010. Filtering offensive language in online communities using grammatical relations. In *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. In *Proceedings of the Content Analysis in the WEB*, volume 2, pages 1–7.