Computational Stylometry and Machine Learning for Gender and Age Detection in Cyberbullying Texts

Antonio Pascucci
UNIOR NLP Research Group
L'Orientale University of Naples
Naples, Italy
apascucci@unior.it

Vincenzo Masucci
Expert System Corp.
Naples, Italy
vmasucci@expertsystem.com

Johanna Monti
UNIOR NLP Research Group
L'Orientale University of Naples
Naples, Italy
jmonti@unior.it

***Abstract*—The aim of this paper is to show the importance of Computational Stylometry (CS) and Machine Learning (ML) support in author's gender and age detection in cyberbullying texts. We developed a cyberbullying detection platform and we show the results of performances in terms of Precision, Recall and F-Measure for gender and age detection in cyberbullying texts we collected.**

***Index Terms*—Computational Stylometry, Gender Detection, Age Detection, Machine Learning, Cyberbullying Detection**

I. Introduction

In this paper, we show the results of a research carried out in November 2018, during the 32nd edition of *Futuro Remoto*, the oldest European event of scientific dissemination, held in Naples by *Città della Scienza* since 1987. This research has been developed in the framework of an Innovative Industrial PhD project in CS by the "L'Orientale" University in co-operation with Expert System Corp. This research has two objectives: i) to increase the number of cyberbullying texts in our corpus for future work and ii) to demonstrate the efficiency of CS and ML in text meaning understanding.

To this aim, a cyberbullying detection web platform was developed which we tested during *Futuro Remoto*. During this event, we asked users to write texts with possible cyberbullying content, and we used the platform to detect gender and age of the writer.

This paper is organized as follows. Section II presents related work while Section III describes CS and ML. The phenomenon of cyberbullying and all the cyberbullying categories are described in Section IV. In Section V we describe our cyberbullying detection platform and the data we collected during *Futuro Remoto*. Results and conclusions are in Section VI and Future Work is discussed in Section VII.

II. Related work

As already shown in other researches, text analysis and Authorship Attribution (AA) techniques with ML algorithms support are a valid tool to oppose cyberbullying phenomena. For example, ACTS (Frommholz, et al., 2016) is a framework for automatic detection of cyberstalking texts that has a specific module based on ML algorithms trained to detect and classify cyberstalking messages.

(Yin et al., 2009) detect harassment thanks to a supervised ML approach, that takes into account the content of the harassment texts only and leaves out author's characteristics. In their study N-grams, TF-IDF score weighting and foul words frequency are used as a baseline. The results show an improvement of the baseline.

Important research has been carried out by (Bogdanova, Rosso and Solorio, 2014) and shows that understanding the behaviour of paedophiles could help to detect and prevent children sexual abuse in social media. The authors highlight that paedophiles try to be nice with a victim and make compliments, at least in the beginning of a conversation but afterwards they tend to be emotionally unstable and prone to loose temper easily and use negative words. The research is based on the following features: percentage of positive (*cute*, *pretty*) and negative (*dangerous*, *annoying*) words, percentage of sadness (*bored*, *sad*) and anger(*furious*) markers, percentage of fear (*scared*, *panic*) markers and percentage of relationship words (i.e. *boyfriend*, *date*) among others. The SVM classification based on combinations of highlevel features achieves 97% accuracy in identifying conversations with cyberpedophiles from cybersex chat logs.

(Rangel and Rosso, 2016) hypothesize that the way people write about topics expressing emotions may help to identify their demographics, their age and gender. Carrying out their experiment on Spanish texts belonging to *PAN-AP-13* corpus, they achieve an accuracy of 64% for gender detection and an accuracy of 66% for age detection. Their research is grounded in an innovative approach, where both style-based (frequency of punctuation marks, capital letters, quotations, POS tags and dictionary-based words) and emotion-labelled graphs features (that not only capture the syntactic structure but also its location in the text) are combined. This research highlights that young people tend to write more about physics, linguistics, literature, law, medicine and chemistry, maybe due to the fact that this is the stage of life when young people mostly speak about their homework. Females seem to write

978-1-7281-3891-6/19/$31.00

more about chemistry or gastronomy, and males about physics or law. In conclusion, the scholars state that the selection of the position of nouns, verbs or adjectives, which mainly give the meaning of the sentence, is the best discriminating features for gender identification, whereas the selection of connectors such as prepositions, punctuation marks or interjections are the best discriminating features for age identification.

III. Computational Stylometry (CS) and Machine Learning (ML)

Our research on AA in cyberbullying texts is based on the assumption that CS and ML have a significant role in Semantic Analysis and Text Meaning Recognition and represent indispensable approaches to the analysis of cyberbullying. CS is the study of linguistic features that represent unintentional linguistic choices by the writer of a text. Right now, we detected one-hundred twenty-five stylometric features for the analysis of Italian texts and one-hundred eighteen stylometric features for the analysis of English texts. ML is the computer ability to learn from data. ML allows the system to preserve in its knowledge base each feature characteristic learned during the training process. In the following list, we report some stylometric features:

- number of sentences in a text;
- number of words in a sentence;
- length of the text;
- lexical richness (namely being able to use different words to develop the same concept);
- use of repeated concepts;
- presence of abstract concepts;
- use of fillers or intensifiers;
- use of slang;
- use of highbrow language;
- use of anaphoras.

Thanks to the statistical analysis of these and other unintentional linguistic choices, we can find out about the psychological and sociological traits of writers. Each author has his/her own writing style (unique as the fingerprint), that takes origin from psychological (i.e. personality, mental health, being a native speaker or not) and sociological (i.e. age, gender, education level) characteristics (Daelemans, 2013).
That's the reason why we talk about *Authorial DNA*.

IV. Analysis of the Cyberbullying Phenomenon

A. Digital Natives

The emergence of the digital age changed people's way of thinking and acting. Many aspects have acquired increased significance compared to previous decades. In 2001 Mark Prensky introduced the concept of *digital natives* (Prensky, 2001), namely people born in the last decades and grown up with the constant presence of digital technologies in their lives. Cyberbullying can occur to any person online and can cause profound psychological outcomes, including depression, isolation and suicide (O'Keeffe and Clarke-Pearson, 2011). In this respect, the term *bullycide* is used: it is a neologism coined by (Marr and Field, 2001) referred both to the suicide of the bullied person or the murder of the cyberbully at the hands of the victim after physical or psychological harassment. Cyberbullying is no less serious compared to the classic forms of bullying, because the so-called *online disinhibition effect* is recognized in cyberbullying cases (Suler, 2004).

B. Cyberbullying in Italy

Over the years, the age of the first approach to digital technologies decreases, and in many aspects it is anything but positive. On the one hand, it is inevitable that games are the first virtual ground used by children, but on the other hand, these virtual grounds represent a place where cyberbullying is spreading very quickly. The Italian regions where the majority of cyberbullying cases occurred in 2017 are Lombardia, Calabria, Campania and Sicilia. An *ISTAT* (Italian National Institute of Statistics) report from 2014 highlights that more than half of the interviewed teenagers claimed to have been victims of cyberbullying during the previous twelve months. A research carried out by (Lazzari, 2015) shows, inter alia, a worrying situation regarding misinformation: nearly 4% of the interviewed children declared not to know the meaning of "cyberbullying". Fig.1 shows the reasons why victims were targeted in Italy in 2017: in 22% of the cases, victims were targeted because of their appearance and in 21% because of their shyness. Another worrying situation is represented by 10% of the cases due to disability.

C. Cyberbullying Categories

Cyberbullying is a complex phenomenon which includes different types of attack (*Flaming*, *Sexual Harassment*, *Mas-*

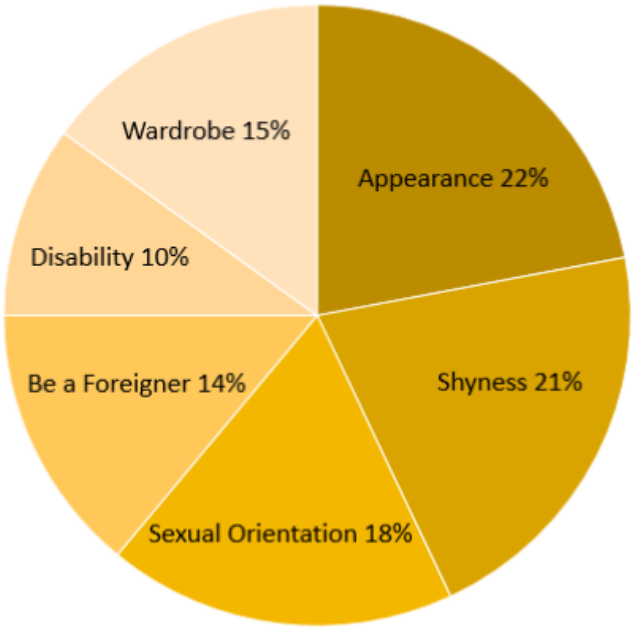

Fig. 1. Reasons why the victim were targeted (cases in Italy in 2017). Source: *studenti.it*

querade, *Cyberstalking*, *Racism*, *Trickery*, *Denigration*, *Homophobia*, *Pull a pig* and *Grooming*). Apart from the aforementioned categories, we identified other two cyberbullying categories, namely (*Cyberbashing* and *Banning*), but we focused mainly on the cyberbullying categories based on written texts:

- **Flaming**: this term means, above all, sending aggressive messages to a victim in chat groups of online games, but *Flaming* cases can be found in social media as well, where hate-speech is taking roots. Usual targets of hate-speech on social media are celebrities (such as politicians, athletes and singers) that almost always don't react. Instead, their fans react sticking up for their idol and thus starting a so-called *Flame war*, where the person who insulted their idol is threatened. Sometimes these offensive messages go on and on. Fig.2 shows a case of *Flaming* detected on twitter: on 24th April 2018, Giorgio Napolitano (second-to-last President of Italian Republic) had heart surgery.
  The tweet in Fig.2 says: *doctors, put your hand on your heart, holy God. Go on strike!*;
- **Sexual Harassment** means sending abusive messages in rapid fire. Persistence is identifiable here, as opposite to *Flaming*. Sex is almost always the topic of messages of this category, with the addition of threats if the victim doesn't do what the cyberbully asks;
- **Trickery**: at the beginning, the cyberbully stands as an equal of the victim, who shares confidences with the cyberbully, feeling familiar with him. In short time, the cyberbully starts threatening the victim to post online confidences if he/she doesn't do favours (e.g. send nude pictures or videos to the cyberbully);
- **Cyberstalking** is the evolution of *Sexual Harassment*. Victims worry about their life, because stalkers force them to send compromising pictures or information and threaten to kill them if they do not obey. *Cyberstalking* also occurs when ex-wives intimidate their ex-husbands or vice-versa. It is worth noting that in 2017 65% of *Cyberstalking* victims in Italy were men;
- **Masquerade**, also known as *Impersonation*, consists in violating of the social accounts of the victims, in order to take over their identity, to provide their web audience with fake news about the victims;
- **Denigration**: the purpose is to publish untrue rumors about the victim. In this way victims are socially isolated because of their damaged reputation;
- **Racism**: it refers to the discrimination because of a different race, ethnicity or nationality. Starting in 2018 *Racist* cases in Italy are dramatically increasing;
- **Homophobia**: it refers to a hateful attitude towards gay, lesbian or bisexual people and to all those actions put in place to denigrate people with a different sexual orientation;
- **Pull a Pig**: victims of *Pull a Pig* cases are girls considered ugly or fat. Cyberbullies send flattering messages in order to convince the victims that they are trying to seduce them. The hapless victims fall for it and then cyberbullies post their conversations on social media with the purpose of mocking. In a very short time *Pull a Pig* cases have been rising significantly, reaching 12% of cases in Italy in 2017. The term derives from the formulaic statement used by the cyberbully *"You have been pigged"* to end the conversation with the victim and unveil the cruel joke. It was used by the cyberbully of the first *Pull a Pig* case, which happened in Barcelona in October 2017;
- **Grooming** represents the ultimate practice to groom children online. Fig.3 shows that *Grooming* was the most frequent type of cyberbullying in Italy in 2017. At the beginning cyberbullies approach the victims using a friendly or fatherly conduct, aimed at gaining victim's trust (usually posing like a policeman, a teacher or a doctor, namely as someone that would never harm children). Once trust is earned, victims are pulled in

Fig. 2. A case of *Flaming*

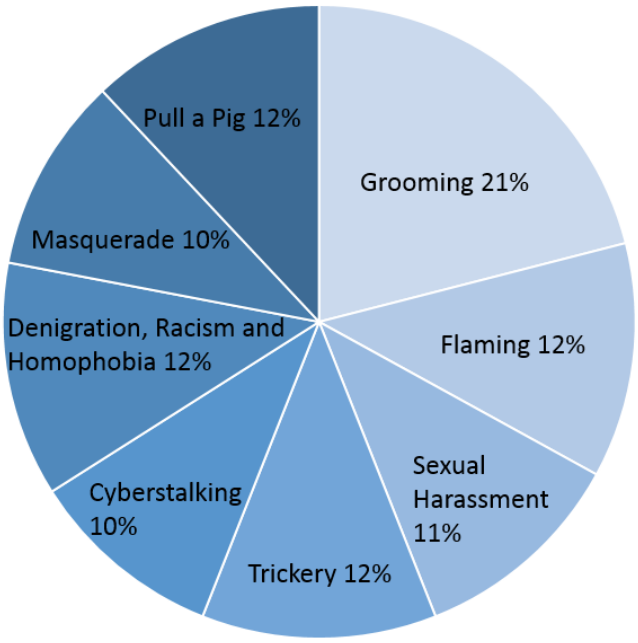

Fig. 3. Cyberbullying cases in Italy in 2017. Source: ANSA.it

sexual activities using for instance webcams.

D. COGITO® support: a rule-based approach

Our standard approach to features extraction consists in the following workflow:

- *Linguistic Definition of Stylometric Features*: since each author operates grammatical choices when writing a text, we organize all the grammatical characteristic of the texts under study in a taxonomy in order to detect the authorial fingerprint based on grammatical choices;
- *Semantic Engine Development*: we train Expert System's semantic engine in order to extract the features from the analyzed texts. The semantic engine is implemented thanks to COGITO®'s semantic network (Sensigrafo) able to operate word-sense disambiguation;
- *Features Extraction*: texts are analyzed and all features (based on the grammatical characteristics) are extracted;
- *Machine Learning*: the features extracted are used to train the model (supervised ML process) in order to detect these features in the untagged texts.

In order to train the two classifiers (*gender detection* and *age detection*), we built a corpus composed of about five thousand texts characterising both gender classes and the nine different age classes. About three thousand cyberbullying texts were found on online blogs run by psychologists and students. The remaining two thousand texts are hotel reviews, Instagram posts, YouTube comments, journalism texts and terrorism texts. These cyberbullying texts have been then splitted in *high cyberbullying content texts* and *low cyberbullying content texts*. The following list provides the statistics about the six hundred-thirty eight cyberbullying texts used for the stylometric analysis and the development of the linguistic rules:

- **Flaming**: 63 texts;
- **Sexual Harassment**: 303 texts;
- **Trickery**: 3 texts;
- **Cyberstalking**: 21 texts;
- **Masquerade**: 2 texts;
- **Denigration**, **Racism** and **Homophobia**: 154 texts;
- **Pull a Pig**: 5 texts;
- **Grooming**: 87 texts.

Thanks to our knowledge in CS, we identified all stylistic and linguistic features characterizing texts belonging to each one of the aforementioned categorie. These features provide the knowledge base to COGITO®. Thanks to COGITO® it is possible to write linguistic rules in order to process the texts and organize a taxonomy in which each limb represents a single category. Once the taxonomy is organized, inside the single limb it is possible to write linguistic rules that allow to recognize the characteristics of the texts and to ascribe them to the corresponding cyberbullying category.
Corpus analysis can be summarized in four steps:

- *Taxonomy Editing*: We edit a taxonomy composed of all cyberbullying categories;
- *Linguistic Rules Writing*: All the linguistic rules are the result of the analysis of the collected corpus. For example, if a text we tagged as a *Flaming* case shows specific linguistic characteristics, we develop a linguistic rule based on these characteristics. Several rules are based on the use of keywords (rules to detect *Racism* and *Homophobia*), the use of positive words in the first part (*Trickery*), or intensifiers (*Masquerade*): i.e. *sono un grande cretino!* (*I'm a big idiot!*);
- *Texts Tagging and Categorization*: We analyze the whole corpus and we tag the texts according to the cyberbullying category it belongs to (if present) in order to assign each text to the corresponding cyberbullying category. The texts of the corpus were manually tagged by two annotators;
- *Model Training*: After this analysis phase, we train our model in order to learn all stylistic features characterising the texts belonging to the ten cyberbullying categories we mention in Section IV.B.

V. Cyberbullying Detection Platform at *Futuro Remoto*

The platform we developed for *Futuro Remoto* combines our knowledge in CS, ML algorithms and a linguistic rules-based approach. Here we describe how the platform works. Texts are categorized according to the stylometric analysis and by means of all the features detected by the semantic engine. The texts are submitted to two different classifiers:

- *gender detection classifier*, trained with Random Forest algorithm support. It classifies the texts according to one of the two classes: male or female;
- *age detection classifier*, trained with Sequential Minimal Optimization algorithm support. It classifies the texts according to one of the nine classes: 10s (from 10 to 19 years old), 20s (from 20 to 29 years old) and so on to 90s (90 to 99 years old).

The choice to train our two classifiers with Random Forest algorithm for gender detection and Sequential Minimal Optimization algorithm for age detection is based on their better performances compared to other algorithms used during the testing phase, such as Simple Logistic and Tree J48.
During *Futuro Remoto* we asked users i) to write texts with possible cyberbullying content, ii) to select the cyberbullying category they believed the text belonged to, iii) to enter nickname and age and lastly to select their own gender.
The platform is developed in Angular 2+, the framework is in JavaScript and uses web services to interact with the back-end. Fig. 4 shows a screenshot of the platform. In this example, we simulate to be an eighteen years old male (on the top right of the screen) with the nickname *ciao* (on the top center) and to write a *Racism* text (on the top left) and to select the *Racism* category. After text processing, gender and age predictions appear on the lower part of the screen. As we can see, the platform predicts (in this case, correctly) author's gender and age class (on the bottom right) and the cyberbullying category present (if present) in the text (bottom left and bottom center).

Fig. 4. A screenshot of the platform

A. Linguistic Analysis of Data Collected at Futuro Remoto

In four days we collected five-hundred twenty-seven cyberbullying texts written by men and women of all ages. We got immediately struck by the writing style standardisation. Especially in texts written by teenagers (10s age class), we noticed that their vocabulary is down to a few words and expressions, determining an ever-increasing standardization of language.

In the *Flaming* cases we collected, for instance, teenagers proved to be familiar with new slang words from online games. An example is the Italian expression *Nabbo*, calque of *Noob*, used by cyberbullies to assert their primacy or to abuse new gamers. During the trial, almost all teenagers declared to use this expression daily during their game sessions to offend other people, although they did not know it was a cyberbullying expression and, above all, without knowing its real meaning.

Several *Racism* cases we detected contain the text "devi morire, negro" ("you must die, nigger"). About half of the teenagers declared to use these words and expressions belonging to *Racism* (or *Homophobia* too) towards people they wanted simply to mock. We have also noticed that in texts belonging to these two categories, the offense is almost always followed by the incitement to commit suicide: "ammazzati, frocio!" ("kill yourself, fag!"). Here again, most of the texts (especially *Homophobia* texts) are from teenagers.

VI. Results and Conclusions

Before the beginning of *Futuro Remoto* we provided the platform with a knowledge base of one-hundred sixty-nine cyberbullying texts characterising the texts written by men and women and one-hundred thirty cyberbullying texts characterising the nine different age classes. We show the number of cyberbullying texts we collected for Gender (Tab. I) and Age (Tab. II) detection and the results in terms of Precision, Recall and F-Measure concerning the first, second, third and fourth day of *Futuro Remoto*.

Results in **bold** in Tab. I and Tab. II represent the best performance achieved for the corrisponding evaluation measure.

Our gender detection knowledge is based on previous work carried out on AA. The number of texts in the tables is the incremental number of texts collected (i.e. two hundred-seventy texts for gender detection in the first day means one hundred-sixty-nine knowledge base texts plus one hundred-one texts written in that day by the users).

TABLE I
Gender Detection

Data for Gender	Performances				
	Base	*1st day*	*2nd day*	*3rd day*	*4th day*
Number of Texts	169	270	421	532	696
Precision	0.612	0.643	0.634	0.674	**0.686**
Recall	0.621	0.652	0.646	0.677	**0.688**
F-Measure	0.608	0.633	0.627	0.669	**0.679**

TABLE II
Age Detection

Data for Age	Performances				
	Base	*1st day*	*2nd day*	*3rd day*	*4th day*
Number of Texts	130	231	382	493	657
Precision	0.272	0.380	0.462	**0.500**	0.475
Recall	0.280	0.420	0.542	0.546	**0.549**
F-Measure	0.273	0.391	**0.459**	0.451	0.438

Precision, Recall and F-Measure performances in Tab. I and Tab. II are always to be considered on the analysis of the total number of texts we collected at *Futuro Remoto*.

Several writing style differences between males and females already detected during previous experiments are confirmed in the cyberbullying texts we collected during *Futuro Remoto*. These stylistic differences are:

- *semantics*: males prefer causative verbs and females prefer i) chromatic adjectives describing an object and ii) verbs belonging to state and opinion classes;
- *grammar*: men use almost always the first person, and, as confirmed by (Argamon, Koppel, Fine and Shimoni, 2003), female writers use more pronouns (*we*, *you*, *she*). Finally male writers use more determiners (*a*, *the*, *that*) besides making more spelling errors;
- *pragmatics*: females prefer to use vague words (*something like*, *especially if*, *you know what I mean?*), fillers and the so-called Tag-questions;
- *lexicon*: women elaborate their questions and prefer to use more euphemisms (*advanced in years* instead of *old*) and males, however, use short and direct sentences and prefer using the basic form of words.

As shown in the two tables above, performances increase day by day, expecially for age detection, with a Recall increase from 0.280 (on the analysis of knowledge base texts) to 0.546 (on the analysis of the whole dataset). Results for gender detection are encoraging too. Gender detection Recall performance is 0.688 on the total of texts. It would appear as a bad result, considering that gender detection is a binary classification, but we have to consider too that certain texts analyzed were really short, which does not allow to detect stylistic features belonging to female or male writers.

VII. Future Work

All the data collected at *Futuro Remoto* will set up a new training set for our platform. The training set will also be

enriched with new cyberbullying texts found on blogs. The main idea is to use the platform to warn students and their parents by means of an awareness campaign in schools.
Considering that right now we identified one hundred-twenty-five stylometric features for Italian and one hundred-eighteen for English, our aim is also to increase the number of stylometric features both for Italian and for English. Finally, we aim to develop an app for cyberbullying detection.

ACKNOWLEDGMENT

We are grateful to Expert System Corp. for providing COGITO® for research.
This research has been fully supported by the PON Ricerca e Innovazione 2014/20 fund.

REFERENCES

[1] Argamon, S., Koppel, M., Fine, J., Shimoni, A.R. (2003). Gender, genre, and writing style in formal written texts. Text - Interdisciplinary Journal for the Study of Discourse. 23, 321-346
[2] Bogdanova, D., Rosso, P., Solorio, T. (2014). Exploring high-level features for detecting cyberpedophilia. Computer speech language, 28(1), 108-120.
[3] Daelemans, W. (2013, March). Explanation in computational stylometry. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 451-462). Springer, Berlin, Heidelberg.
[4] Frommholz, I., Al-Khateeb, H. M., Potthast, M., Ghasem, Z., Shukla, M., Short, E. (2016). On textual analysis and machine learning for cyberstalking detection. Datenbank-Spektrum, 16(2), 127-135.
[5] https://www.expertsystem.com/
[6] http://www.cittadellascienza.it/futuroremoto/2018/11/
[7] Lazzari, M. (2015). Spazi ibridi tra la Rete e la Piazza: l'evoluzione della comunicazione degli adolescenti ai tempi dello smartphone. In Marco Lazzari e Marcella Jacono Quarantino (a cura di), Virtuale e/è reale. Adolescenti e reti sociali nell'era del mobile (pp. 45-80) Bergamo: Bergamo University Press, ISBN: 9788866422211
[8] Marr, N., Field, T. (2001). Bullycide: Death at playtime. Success Unlimited.
[9] O'Keeffe, G. S., and Clarke-Pearson, K. (2011). Clinical reportthe impact of social media on children, adolescents, and families. Pediatrics, peds-2011.
[10] Prensky, M. (2001), Digital natives, digital immigrants part 1. On the horizon, 9(5), 1-6.
[11] Rangel F., Rosso P. (2016). On the Impact of Emotions on Author Profiling. In: Information Processing Management, vol. 52, issue 1, pp. 73-92 DOI: 10.1016/j.ipm.2015.06.003
[12] Suler, J. (2004). The online disinhibition effect. Cyberpsychology behavior, 7(3), 321-326.
[13] Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L. (2009). Detection of harassment on Web 2.0. Proceedings of the Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW2009, Madrid, Spain.