Gender Classification of Twitter Data Based on Textual Meta-Attributes Extraction

Chapter · March 2016

DOI: 10.1007/978-3-319-31232-3_97

José Ahirton Batista Lopes Filho
Universidade Presbiteriana Mackenzie

Rodrigo Pasti
AXONDATA / Natural Computing Laboratory -…

Leandro De Castro
Universidade Presbiteriana Mackenzie

Gender Classification of Twitter Data Based on Textual Meta-Attributes Extraction

José Ahirton Batista Lopes Filho[1] , Rodrigo Pasti[1], Leandro Nunes de Castro[1]

[1] Natural Computing Laboratory (LCoN), Graduate Program in Electric and Computer Engineering, Mackenzie Presbyterian University (UPM), Rua da Consolação 930, Higienópolis, São Paulo, SP, Brazil
{ahirtonlopes, rodrigo.pasti}@gmail.com, lnunes@mackenzie.br

Abstract. With the growth of social media in recent years, there has been an increasing interest in the automatic characterization of users based on the informal content they generate. In this context, the labeling of users in demographic categories, such as age, ethnicity, origin and race, among the investigation of other attributes inherent to users, such as political preferences, personality and gender expression, has received a great deal of attention, especially based on Twitter data. The present paper focuses on the task of gender classification by using 60 textual meta-attributes, commonly used on text attribution tasks, for the extraction of gender expression linguistic cues in tweets written in Portuguese. Therefore, taking into account characters, syntax, words, structure and morphology of short length, multi-genre, content free texts posted on Twitter to classify author's gender via three different machine-learning algorithms as well as evaluate the influence of the proposed meta-attributes in this process.

Keywords: machine-learning, classification, gender, social media, Twitter, extraction, meta-attributes, Portuguese language.

1 Introduction

Social Media are a group of services and applications built from the technological foundations of *Web* 2.0, which allow the exchange and creation of collaborative contents, also known as user-generated contents. More specifically, *Web* 2.0 refers to a new way that software developers and final users have found to use the World Wide Web. With it, contents and applications are no longer created or published individually, but in a collaborative manner [1].

Different techniques of image, sound, video, and text processing can be applied to find patterns, tendencies, and provide qualitative and quantitative measures for use in different areas, such as economy [2], politics and government [3], recommendation systems [4], and rumor control [5]. Besides, with the growth of the area in the past years, there is an increasing interest in the automatic characterization of social media users based on the informal content they generate.

Amongst the purposes of this task of automating the characterization of user's profiles in social media, is that of labeling them in demographic categories, such as age

[6], ethnicity, origin, and race [7]. Other attributes inherent to these users, for example, political preferences [8], personality [9], and gender [10], have also been studied, specially based on Twitter data analysis [11] [12] [13] [14].

Such preferences and opinions can be personalized from these data in different contexts, for example, supporting business applications, as in digital marketing; helping in the answer of important social science questions; in the detection and action against embezzlers and falsifiers; as well as in the protection against terrorism, amongst other crimes.

However, these uses still face an essential difficulty, found in the majority of social media: the anonymousness. Knowing that, in cyberspace, users often do not need to provide their real information, such as name, age, gender, and address as well as, in many cases of improper social media usage, authors often hide their addresses using anonymous servers, also their true identities, to avoid being discovered. In that sense, it is necessary to develop and investigate efficient methods to assist the forensic tracking of identity in cyberspace.

The present work provides an experimental methodology based on the use of a set of 60 textual meta-attributes to classify the gender (male/female) of authors of *tweets* in the Portuguese language, by employing three machine learning algorithms. Furthermore, attribute selection via χ^2 (Chi-Square) and Information Gain techniques are performed to analyze which attributes from this set excel in the classification of a *corpus* with the presence of neutral messages, as well as the impact of each set of meta-attributes in the obtained results.

The paper is organized as follows: Section 2 presents the Gender Classification problem; Section 3 presents the proposed methodology; Section 4 provides a brief discussion concerning the Experimental Results obtained with the use of the selected Supervised Classification Model, and, at last, in Section 5, the Conclusions and Future Works are presented.

2 Gender Classification in *Tweets*

A typical authorship attribution problem consists of attributing a text of unknown authorship to a candidate author, given a set of candidate authors. To do that, sample texts of unquestionable authorship are previously available to train the classifiers [15]. Since in gender classification tasks, especially for short and context-free texts (*tweets*), a prior set of candidate authors is not available, the majority of the existing models work with attributes inherent to users.

The gender classification task can be treated as a binary classification problem, i.e., given two classes, male and female, the task is to attribute an anonymous message to one of these classes, without knowing any candidate author [15]. It is important to mention that, when speaking of gender classification, other aspects must be observed, such as *sex* and *gender expression*.

The American Psychology Association conceptualizes *sex*, *gender* and *gender expression* in the following manner [16]: *Sex*, refers to the biological state of a human being and is usually classified as male, female, or *intersex*; *Gender*, refers to attitudes, feelings, and behaviors associated with the biological sex of a person in a given culture;

and *Gender Expression* refers to the way in which a person acts to communicate its gender in a given culture, for example, in terms of clothing, interests, and communication patterns.

The present work is based on the concept of language as a form of *gender expression*, rather than related with sex. For consistency purposes, male and female were used as the two possible genders. Therefore, the studied gender classification can be understood as the task of detecting if a certain social media user is of male or female gender by analyzing the content and behavior demonstrated on their messages, and, furthermore, there is an interest in investigating which of the analyzed meta-attributes, which are addressed in the next section, can be considered meaningful to gender classification.

The first efforts aimed at gender classification in Twitter [10] [11] focused, mainly, on the creation of meta-attributes based on the use of words or specific terms (psycholinguistic characteristics), such as *emoticons*, abbreviations, and affective expressions, information mostly gathered from blogs [10], along with the analysis of other data like user's name, full name, location, URL links, and description, obtained from Twitter itself [11].

Some challenges of this sort of classification are listed in [12], with highlights to the major usage of colloquialism, the simplification of writing, the existence of different stylistic tools created by the users, URLs, and the spread use of acronyms in the English language (LOL, BRB, etc.). As for [13], inferences for non-English languages began to be investigated, together with the exploration of new unique attributes.

Finally, in [14] there is the first investigation of Portuguese language users. However, as in previous works [10], specific words, identified as relevant to the context of the gathered messages, are used as meta-attributes, such as terms with male and female suffix. Nevertheless, this methodology may not be an adequate solution when the objective is to investigate profiles with few *tweets*, or whose language can be considered neutral, like in *corpus* with the presence of journalistic messages [15]. Besides, any approach that considers the user's name inspection [11] [13] is subject to failure, for, while in some cases this information can reveal gender, many users choose names, nicknames and aliases that do not convey this information.

3 GENEC: A Gender Expression Classification Methodology

The proposed system to classify gender expression in tweets is based on the study presented in [15], where the gender classification problem was divided into four main steps: the gathering of an adequate text message *corpus* (*tweets*) that will compose the database; the automatic extraction of the characteristic values (meta-attributes) from each *tweet* or set of *tweets*; the development of a classification model to identify the author's gender; and the selection of the most significant meta-attributes in gender identification (cf. Figure 1).

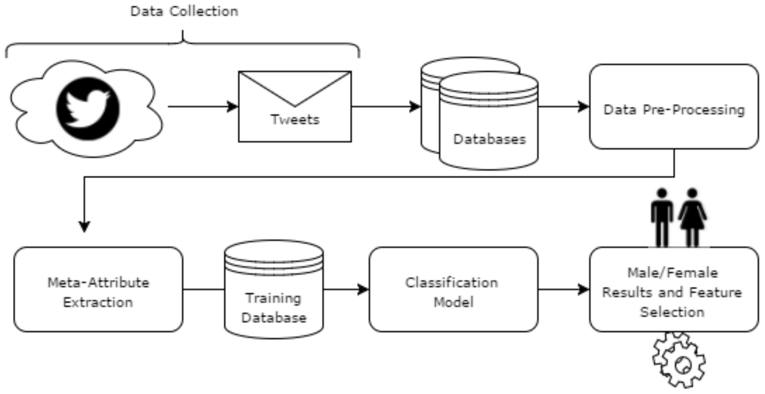

Fig. 1. GENEC: The Gender Expression Classification process.

3.1 Data Collection

A method using the *tweepy* module (https://github.com/tweepy/tweepy) was created in Python to access the Twitter API's as well as to create files containing the last 3,240 *tweets* of a set of 60 Brazilian journalist's profiles (30 male and 30 female). Those profiles were chosen because they are Twitter heavy users, with more than 2.000 published *tweets*, and whose *tweets* have both journalistic nature (neutral) as well as family affairs, day-to-day messages and personal opinions.

3.2 Pre-Processing

The pre-processing step in this work consists of, besides removing graphic accents and non-ASCII characters, using the processes of tokenization and *part-of-speech tagging* (*POS tagging*). The Mac-Morpho *corpus*, the largest POS *corpus* labeled in modern Portuguese, was used to analyze the relation between adjacent and related words in a sentence or paragraph, for the proper identification and classification of these words into categories such as adjectives, adverbs, articles, interjections, conjunctions, prepositions, verbs, etc.

3.3 Meta-Attributes Extraction

For the identification of gender expression in Portuguese *tweets*, the present work proposes the use of a set of 60 textual meta-attributes divided in the following groups:

1. Those based on *Characters and Syntax* (Table 1), stylistic characteristics that have already been used to solve authorship attribution problems;
2. Those based on *Words* (Table 2), which include statistic characteristics, known as vocabulary richness measures, such as Yule's K measure, Simpson's D measure, and Entropy;

3. Those based on *Textual Structures* (Table 3) that represent the way in which the author organizes the structure of a message; and
4. Those based on *Morphology* (Table 4), words that are used to express grammatical relations with other words within a sentence, or to specify the attitude or mood of the text or message author.

Table 1. Characters and Syntax based Meta-Attributes.

Name	Description
C_1	Total number of characters
C_2	Ratio between the total number of low case letters (a-z) and the total number of characters
C_3	Ratio between the total number of upper case letters (A-Z) and the total number of characters
C_4	Ratio between the number of digits and the total number of characters
C_5	Ratio between the number of white spaces and the total number of characters
C_6	Ratio between the number of tab spaces and the total number of characters
C_7	Ratio between the quotation marks (“) and the total number of characters
C_8	Ratio between the number of commas (,) and the total number of characters
C_9	Ratio between the number of colons (:) and the total number of characters
C_{10}	Ratio between the number of semicolons (;) and the total number of characters
C_{11}	Ratio between the number of simple question marks (?) and the total number of characters
C_{12}	Ratio between the number of multiple question marks (???) and the total number of characters
C_{13}	Ratio between the number of simple exclamation points (!) and the total number of characters
C_{14}	Ratio between the number of multiple exclamation points (!!!) and the total number of characters
C_{15}	Ratio between the number of simple periods (.) and the total number of characters
C_{16}	Ratio between the number of suspension points (...) and the total number of characters

Table 2. Word based Meta-Attributes.

Name	Description
W_1	Total number of words
W_2	Average number of characters per word
W_3	Ratio between the number of different words and the total number of words
W_4	Ratio between the number of words with more than 6 characters and the total number of words
W_5	Ratio between the number of words with 1 to 3 characters (short words) and the total number of words
W_6	Ratio between *hapax legomena* (word that appears only once in a whole text) and the total number of words
W_7	Ratio between *hapax dislegomena* (word that appears only twice in a whole text) and the total number of words
W_8	Yule’s K Measure
W_9	Simpson’s D Measure
W_{10}	Sichel’s S Measure
W_{11}	Honore’s R Measure

W$_{12}$	Entropy Measure
W$_{13\_28}$	Ratio between the distribution of words size frequency and the total number of words

Table 3. Textual Structure based Meta-Attributes.

Name	Description
TS$_{1}$	Total number of sentences
TS$_{2}$	Total number of paragraphs
TS$_{3}$	Average sentences per paragraph
TS$_{4}$	Average words per paragraph
TS$_{5}$	Average characters per paragraph
TS$_{6}$	Average words per sentence
TS$_{7}$	Ratio between the number of sentences starting with low case letters (a-z) and the total number of sentences
TS$_{8}$	Ratio between the number of sentences starting with upper case letters (A-Z) and the total number of sentences
TS$_{9}$	Ratio between blank lines and the total number of paragraphs
TS$_{10}$	Average number of characters in non-blank lines

Table 4. Textual Morphology based Meta-Attributes.

Name	Description
TM$_{1}$	Ratio between the number of articles and the total number of words
TM$_{2}$	Ratio between the number of pronouns and the total number of words
TM$_{3}$	Ratio between the number of auxiliary-verbs and the total number of words
TM$_{4}$	Ratio between the number of conjunctions and the total number of words
TM$_{5}$	Ratio between the number of interjections and the total number of words
TM$_{6}$	Ratio between the number of prepositions and the total number of words

After the pre-processing step, the meta-attributes extraction module must then produce a characteristics vector for each *tweet*, or set of *tweets*, in order to represent the values of the different meta-attributes. For example, for the following *tweet* of the @realwbonner profile: "*Como dizem meus sobrinhos de Twitter, rimos litros. Embora eu não entenda como se possa medir o volume espacial de risos. Mas tudo bem*"; the first five output characteristic values, which will take part in the *tweet*'s representative vector, referring to the first five meta-attributes based on characters and syntax are: [134, 0.7761, 0.0298, 0.0, 0.1716].

3.4 Supervised Classification Model

From what was previously presented, it is clear that, after the development of a set of characteristics that can remain relatively constant for a large number of messages written by authors of the same gender, the gender classification problem may be addressed as a binary classification task [15], i.e., given a *tweet t*, or a set of concatenated *t tweets*, represented by a dimensional vector **v**, where *v* is the total number of meta-attributes, assign *t* to $Class_1$ if the author is male, or to $Class_2$ if he/she is female. That is, given a set of formerly known pre-classified messages, classifying-algorithms must be used to categorize such *tweets*.

Mathematically, from a classification model in the form of $y_i = f(\mathbf{v}_i)$, with $\{(\mathbf{v}_i, d_i)\}_{i=1}^{N}$ being the set of problem instances, and $\mathbf{v}_i = (v_{i1}, v_{i2}, \dots, v_{iv}), \forall i$ as the set of meta-attributes representing the *tweet* with d_i the desired output (male or female class). With $y_i = (y_{i1}, y_{i2}, \dots, y_{iN})$ outputs model, where $y_i \in \{+1, -1\}$ are class labels, so that $Class_1$, +1, is for messages with amle gender expression, or $Class_2$, −1, for messages with female gender expression; and *N* refers to the number of *tweets* in the *corpus*. Thus, each *tweet* in the corpus, or all concatenated *tweets* in a single *tweet*, are converted into a multidimensional characteristics vector *v*, with each characteristic contributing to classify the author of the *tweet* under the corresponding gender category.

Furthermore, to ensure that all meta-attributes are treated equally in the classification process, all of them are normalized over the [0,1] interval. Starting with this supervised classification model, this study aims to analyze the performance of three different classifyers (*Best First Tree* - BFTree, *Multinomial Näive Bayes*- MNB, and *Support Vector Machines* - SVM) for the suggested set of 60 meta-attributes. We also performed an investigation about the relevance of each meta-attribute by selecting some of them based on the χ^2 (Chi-Square) and Information Gain techniques.

The following experimental results were obtained by using classifiers implemented in WEKA [17] and the databases were acquired by filtering the group of *tweets* collected for each journalist's profile, excluding *tweets* which were not in Portuguese, *retweets* (RT's), *tweets* that were only mentions to other profiles (@), only link *tweets*, and *tweets* that were composed entirely of non-ASCII characters.

4 Performance Evaluation

To assess the performance of the proposed methodology (GENEC) approximately 187,950 *tweets* were collected, roughly 91,836 male and 96,114 female *tweets*, all from well-known Brazilian journalists. Those *tweets* were arranged in a way so that each file could correspond to a given profile, and that, after filtered, each row from each file represented a *tweet* written in Portuguese. Then, the most recent 100 *tweets* for each of the 60 profiles (30 male, and 30 female), in a total of 6.000 *tweets,* were arranged in two files: one composed of all *tweets* of each profile, which after pre-processing and meta-attributes extraction, turned into characteristic vectors *tweet* by *tweet*; and another composed of all *tweets* of each profile displayed in a concatenated way (like a full text), so that a single characteristic vector is created for each profile.

Table 5 presents the cross-validation (10 × 10-*folds*) results, when each of the methods were used for the three different algorithms, taking into account the following evaluation measures: Total Average Accuracy, that is, the number of instances correctly classified; and Precision, referring to the relative number of instances that truly belong to a class divided by the total classified instances as belonging to that specific class.

Table 5. Obtained results for gender classification.

Method	**Database Size**	**Measures**	**BFTree**	**MNB**	**SVM**
		Accuracy	63.5%	61.96%	**68.08%**

Tweet* by *Tweet	3000 (M) and 3000 (F)	Precision	0.599 (M) e 0.714 (F)	0.597 (M) e 0.656 (F)	0.674 (M) e 0.719 (F)
Concatenated *Tweets*	30 (M) and 30 (F)	Accuracy	**81.66%**	70%	68.33%
		Precision	0.794 (M) e 0.885 (F)	0.667 (M) e 0.75 (F)	0.69 (M) e 0.677 (F)

The data in Table 5 indicates that the *BFTree* algorithm has a good performance when compared to the other classifiers, especially for the concatenated tweets. However, SVM's performance is stable for both databases, showing results equivalent to those published in [14] [15] for a different problem, even when using more neutral language *corpus*. To assess the impact of the meta-attributes used in the classification task, attribute selection was made using the χ^2 (Chi-Square) and Information Gain techniques in conjunction with ranking, through cross-validation (10 × 10-*folds*), to the *tweet by tweet* and *concatenated tweets* databases. The meta-attributes significance ranking, in descending order, is shown in Table 6 (captions for each meta-attribute are better described in Tables 1, 2, 3 and 4).

Table 6. Meta-Attributes significance ranking.

Method	Techniques	Most Significant Meta-Attributes
Tweet by Tweet	χ^2 (Chi-Square)	W7, W9, W3, W10, W8, W13_1, W12, W6, C1, C7, TS5, TS10, W5, C13, C3, C15, C5, W2, W11, TS4, W1, TS6, C9, W13_2, C4, W13_4, C11, W13_7, TM1, TM6, W13_10, W13_5, W13_3, TM4, TM2, W13_6, W13_13, W13_8, W13_9, W13_15, C8, TS8, W13_12, W4, W13_11, TS7, C10, C2, TM5, TM3, C6, TS1, W13_14, W13_16, TS9, TS3, C12, C14, TS2, C16
	Information Gain	W9, W3, W7, W10, W8, W13_1, W12, W6, C1, C7, TS5, TS10, W5, C3, C13, C15, C5, W2, W11, TS4, W1, TS6, C9, W13_2, C4, W13_4, C11, W13_7, TM1, TM6, W13_10, W13_5, W13_3, TM4, TM2, W13_6, W13_13, W13_8, W13_15, W13_9, C8, TS8, W13_12, W4, W13_11, C10, TS7, TM5, C6, TM3, C2, TS1, W13_14, TS9, W13_16, TS3, C12, C14, TS2, C16
Concatenated *Tweets*	χ^2 (Chi-Square)	TS8, W3, W4, W5, W1, W2, W6, W10, C15, W9, W12, W11, W7, C16, W8, TM6, C5, C4, C3, C2, C13, W13_1, C14, C7, C6, C8, C12, C9, C11, C10, TS6, W13_2, TS5, TM5, TS4, TS10, W13_15, TS3, TS7, TS9, W13_13, TS1, TM1, TM4, TM3, TM2, TS2, W13_3, W13_16, W13_8, W13_14, W13_6, W13_5, W13_7, W13_4, C1, W13_9, W13_10, W13_11, W13_12
	Information Gain	TS8, W3, W2, W1, W4, W5, W6, W10, C15, W7 W9, W12, W11, C16, W8, TM6, C5, C4, C3, C2 C13, W13_1, C14, C7, C6, C8, C12, C9, C11, C10, TS6, W13_2, TS5, TM5, TS4, TS10, W13_15, TS3, TS7, TS9, W13_3, TS1, TM1, TM4, TM3, TM2, W13_3, W13_8, W13_16, W13_14, W13_6, W13_5,

		W13_7, W13_4, C1, W13_9, W13_10, W13_11, W13_12

To assess the importance of the different meta-attribute sets (characters and syntax, words, text structure and text morphology) for the gender expression classification problem, each of the four meta-attribute set was investigated separately and compared with that of all meta-attributes used in conjunction, using *BFTree* algorithm, as presented previously. The results presented in Table 7 shows that all meta-attribute sets contribute to the gender classification process and also, as seen in Table 6, the meta-attributes based on *words*, *characters* and *syntax* as well as *textual structure* tend to be important gender discriminators. Thus, the experimental results indicate that there are rather significant differences between users with male and female gender expressions also when using Twitter, even in the presence of neutral language.

Table 7. Obtained results for gender classification using the different meta-attributes sets.

Meta-Attribute Set	***Tweet by Tweet***	***Concatenated Tweets***
Characters and Syntax	62.45%	53.33%
Words	61.40%	51.66%
Textual Structure	63.15%	66.66%
Textual Morphology	57.38%	66.66%
All Meta-Attributes (Accuracy)	63.50%	81.66%

5 Conclusions and Future Works

This study aimed to analyze different algorithms and meta-attributes for the gender classification problem in social media, specifically on Twitter. To date, few studies were dedicated to such problem for tweets written in Portuguese. Moreover, it proposes a methodology and a set of meta-attributes, both based on the already structured area of authorship attribution, to deal with the gender expression classification problem.

As by-products, in addition to the GENEC methodology, a method for extracting tweets directly from Twitter's APIs and the subsequent automatic creation of *tweets* database for each user profile was developed, as well as a textual meta-attributes extractor, which builds, based on the previously created bases, feature vectors for each message or set of messages, taking advantage of already consolidated techniques such as tokenization and POS tagging.

From the use of such tools, together with classifier algorithms, one can see that the use of textual meta-attributes can be very important for the detection of writing style, and gender expression characteristics, for both genders. The experimental results also show that the use of *BFTree* can achieve good results, even in the presence of neutral messages. As future works, we intend to explore other Decision Trees and SVMs, using databases with more user profiles and *tweets*, as well as develop new meta-attributes based on psycholinguistics (such as dictionaries of negative, positive and neutral words), which can complement and improve the results shown here.

References

1. Kaplan, A. M., Haenlein, M.: Users of the World, Unite! The Challenges and Opportunities of Social Media. Business Horizons, 53, pp. 59 -- 68 (2010)
2. Bollen, J., Mao, H., Zeng, X.: Twitter Mood Predicts the Stock Market. Journal of Computational Science, 2(1), pp. 1--8 (2011)
3. Tumasjan, A., Sprenger, T. O., Sandner, P. G., Welpe, I. M.: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM, 10, pp. 178--185 (2010)
4. Pankong, N., Prakancharoen, S.: Combining Algorithms for Recommendation System on Twitter. Advanced Materials Research, 403, pp. 3688--3692 (2012)
5. Tripathy, R. M., Bagchi A., Mehta S.: A Study of Rumor Control Strategies on Social Networks. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10). ACM, pp. 1817--1820 (2010)
6. Nguyen, D., Gravel, R., Trieschnigg D., Meder T.: How Old do You Think I Am? A Study of Language and Age in Twitter. In: Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM), pp. 439--448 (2013)
7. Bergsma, S., Dredze M., Van Durme B., Wilson T., Yarowsky D.: Broadly Improving User Classification Via Communication-Based Name and Location Clustering on Twitter". In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, pp. 1010--1019 (2013)
8. Golbeck J., Hansen D.: Computing Political Preference among Twitter Followers. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1105--1108 (2011)
9. Lima, A. C. E., De Castro, L. N. A Multi-Label, Semi-Supervised Classification Approach Applied to Personality Prediction in Social Media. Neural Networks, 58, pp. 122--130 (2014)
10. Rao D., Yarowsky D., Shreevats A., Gupta. M.: Classifying Latent User Attributes in Twitter. In: Proceedings of the 2nd. International Workshop on Search and Mining User-generated Contents (SMUC), pp. 37--44 (2010)
11. Burger J. D., Henderson J., Kim G., Zarrella G.: Discriminating Gender on Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1301--1309 (2011)
12. Deitrick W., Miller Z., Valyou B., Dickinson B., Munson T., Hu W.: Gender Identification on Twitter Using the Modified Balanced Winnow. Communications and Network, Vol. 4 No. 3, pp. 189--195 (2012)
13. Ciot M., Sonderegger M., Ruths D.: Gender Inference of Twitter Users in Non-English Contexts. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1136--1145 (2013)
14. Filho R. M., Carvalho A. I. R., Pappa G. L.: Inferência de Sexo e Idade de Usuários no Twitter. In: Proceedings of the III Brazilian Workshop on Social Networks Analysis and Mining (BraSNAM), pp. 200--211 (2014)
15. Cheng N., Chandramouli R., Subbalakshmi K. P.: Author Gender Identification from Text. Digital Investigation 8, 1, July, pp. 78--88 (2011)
16. American Psychology Association: The Guidelines for Psychological Practice with Lesbian, Gay, and Bisexual Clients, Adopted by the APA Council of Representatives. February 18-20 (2011)
17. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1 (2009)