Age Identification of Twitter Users: Classification Methods and Sociolinguistic Analysis

Vasiliki Simaki[1], Iosif Mporas[2], Vasileios Megalooikonomou[1]

[1]Multidimensional Data Analysis and Knowledge Management Laboratory,
Dept. of Computer Engineering and Informatics, University of Patras, 26500- Rion, Greece
[2]School of Engineering and Technology, University of Hertfordshire, Hatfield, UK

[1]{simaki, vasilis}@ceid.upatras.gr,[2]i.mporas@herts.ac.uk

Abstract. In this article, we address the problem of age identification of Twitter users, after their online text. We used a set of text mining, sociolinguistic-based and content-related text features, and we evaluated a number of well-known and widely used machine learning algorithms for classification, in order to examine their appropriateness on this task. The experimental results showed that Random Forest algorithm offered superior performance achieving accuracy equal to 61%. We ranked the classification features after their informativity, using the ReliefF algorithm, and we analyzed the results in terms of the sociolinguistic principles on age linguistic variation.

Keywords: text mining, age identification, text classification, computational sociolinguistics, sociolinguistics.

1 Introduction

The extensive expansion of the web and the plethora of options that social media provide, have resulted in the increase of the web users population. The daily use of the social media as a tool of communication and socialization with other users is a dominating reality, especially in the most developed countries. Twitter is one of the most popular media among men and women of different ages, with more than 255 million of active users per month, and 500 million Tweets sent per day. The automatic extraction of information from the everyday enormously growing volumes of data related not only to the twitter message itself but also to the gender, age and other demographic characteristics of the user are essential for e-government, security and e-commerce applications.

The age is the second most important social factor, after the author's gender, that can be easily extracted among the social media users, in the cases where the user declares his/her age explicitly (e.g. "Sharon, 33"), or when the user provides implicitly his/her age in his/her posts (e.g. "I was born in 1990"). The user's age is in strong correlation with the user's language, as several sociolinguistic theories have proven [6]. Depending to the person's life stage, different linguistic attitudes and choices are observed, resulting the age linguistic variation. A basic principle that delimitates the language of adults from the "teens' language", is that adults use more standard types and normative structures than adolescents, who prefer neologisms (morphological, semantic, etc.), non-standard types and generally more unconventional language structures.

The age of social media users is an essential clue among the demographic information provided, which can be identified for several scientific, commercial and security purposes. Teens use Twitter often without being permitted to or supervised by adults, which could lead to unpleasant situations. It is thus important to identify the age category of social media users, in order to prevent harassment incidents.

Apart the minors' security, the detection of the user's age can be informative about the different trends, opinions, political and social views of each age group. This can enable social scientists to derive important clues about the anthropography among social media users, and how different age groups behave online. Market analysts and advertisers are also interested in such a study, in order to better promote a product or a service to an age group accordingly to their expressed interests and opinions.

In this article we attempt a multifaceted study including computational and theoretical aspects, bounded in the emerging field of Computational Sociolinguistics [11]. More specifically, we investigate the ability of several and dissimilar classification algorithms for the classification of Twitter users with respect to their age class. In order to evaluate them we relied on a large number of text-based features that have been reported in several publications related to age identification from text input. We perform then a feature ranking method, using the ReliefF algorithm, in order to detect the most efficient features during the classification process, and we try at the end, a sociolinguistic- driven interpretation of the most informative features in terms of the basic theories describing the age linguistic variation.

The rest of the paper is organized as follows: In section 2 we describe the background work in the field of automatic age identification. In section 3 we describe the evaluation methodology. Section 4 presents the experimental results. In Section 5 the most informative features are presented and discussed in analyzed according to existing sociolinguistic theories, and finally, in Section 6 we conclude this work.

2 Related Work

Several approaches have been proposed in the literature for the age class identification of documents. In [5] the evolution of the blogs' form during time and aims to predict the age of the users, based on their date of birth was studied. The authors observed that the size of the documents is a discriminative element and they also computed the percent of punctuation appearances, capital letters and spaces. Emoticons, acronyms, slang types, punctuation, capitalization, sentence length and part of speech tags were used, among others, as features in [18], where it was examined whether the online behavior of users can contribute effectively to the prediction of their age category. They observed that there are important and distinguishing changes in the writing style of bloggers before and after the social media expansion.

The stylistic differences in the writing style of bloggers according to their age were investigated in [8], [17], [19]. In [8], [19] they performed a stylometric analysis in terms of gender and age by using non-dictionary forms and the sentence length features. The slang, smileys, out-of-dictionary words, chat abbreviations, on the one hand, and the sentence length on the other, proved to be highly distinctive among different ages and gender, when combined with features proposed by earlier studies. The Naive Bayes classifier was used for the experiments. In [17] the authors focused on the use of slang words. They represented the co-occurrences of slang words that the bloggers use as a graph based model, where the slang words are the nodes and the edges stand for the number of co-occurrences.

Content-based and style-based characteristics were explored in [2], [3], [20]. In [2], one of the age-related conclusions was that the basic difference between older and younger bloggers is the extent to which their communication is outer or inner–directed. In [3], for the style-based features the researchers relied on parts of speech and function words. For the content-based features, they used the words that appear sufficiently enough in the dataset and that have a high value on the information gain measure. In [20] the authors created the “Blog Authorship Corpus”, in order to identify the author’s age and gender. They used style-related features (selected part-of-speech, function words, blog words and hyperlinks) and content-based characteristics in order to detect the gender and the age. They observed that specific forms and unigrams are more frequent in young bloggers, the blogging style and topics are different among 10’s, 20’s and 30’s.In addition, lexical, syntactic and structural features were employed in [14], in order to identify the age and gender of the authors. Among other features, they relied on positive and negative words, stop words frequency, smiley word list and punctuation appearances. A decision tree classifier was used for classifying the author profile.

In [12] a study in language use among different age categories of Twitter users was performed. Their analysis proved that differences in style, references and conversation depended not only on the age category, but also on the life stage of the user.

Except as a categorical variable using age classes, age can be considered as a continuous variable also. In [13] the age identification task was performed using linear regression, using part of speech unigrams and bigrams and features from the LIWC [22], such as the percent of words having more than 6 letters.

3 Age Class Identification Methodology

For the evaluation of classification algorithms on the task of Twitter users' age identification we adopted a standard approach followed in most of the previous related work, i.e. preprocessing, feature extraction and classification structure was utilized, as illustrated in Figure 1.

Specifically, each Twitter post is initially preprocessed. During pre-processing, each post is split into sentences and each sentence is split into words. Afterwards, three feature extraction methodologies are applied in parallel and independently to each other to each post. In detail, text mining, sociolinguistic based and context-based features are extracted constructing vectors V_{TM} , V_{SL} and V_{CT} , respectively. These features are consequently concatenated to a super vector $V = V_{TM} \| V_{SL} \| V_{CT}$. This results to one feature vector, V , per Twitter post, which is processed by a classification algorithm in order to label each post with an age class.

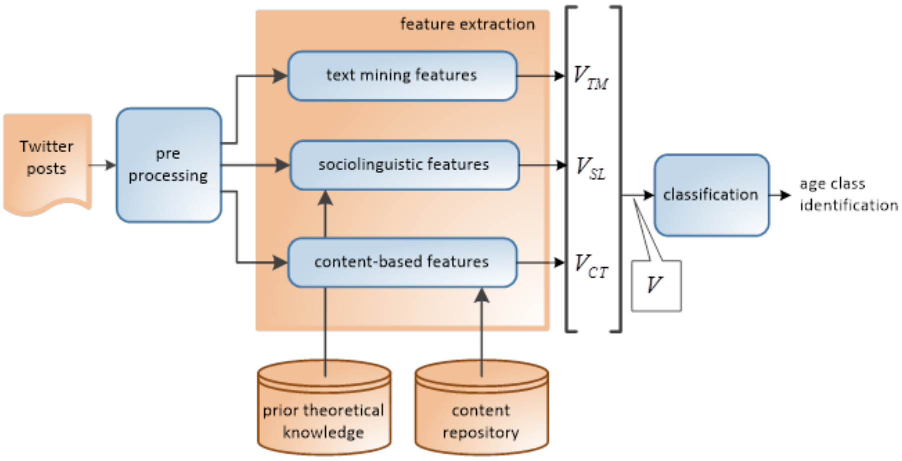

Fig.1. Block diagram of the Twitter's posts age class identification scheme.

3.1 Feature Extraction

Three categories of features were computed for each Twitter post, namely the text mining features, the sociolinguistic-based features and the content-based features.

As considers the text mining features, they consist of statistical values in character and word level, used in several text classification tasks (authorship attribution, gender and age identification, genre classification), as presented in Table 1. These feature set is included in a vector V_{TM} , which has length equal to 40.

Table 1. The text mining classification features.

Text mining features	
# of characters per tweet	standard deviation of the word length
normalized # of alphabetic characters	# of function words
normalized # of upper case characters	average # of sentences per paragraph
normalized # of digit characters	average # of characters per paragraph
normalized # of space characters	minimum word length
normalized # of tab (“\t”) characters	normalized # of emoticons
# of occurrence of each alphabetic character	# of “hapax legomena”

# of adverbs
total # of words
normalized # of words with length less than 4 characters
normalized # of characters per word
average word length
# of sentences
# of paragraphs
# of lines
average # of characters per sentence
average # of words per sentence
normalized # of different words
# of articles
normalized # of words whose letters are all capital
# of punctuation symbols (“.”, “,”, “!”, “?”, “:”, “;”, “’”, “\””)

of “hapax dislegomena”
maximum word length
normalized # of words that start with a capital letter
normalized # of stop words
# of nouns
# of proper nouns
# of adjectives
# of prepositions
# of verbs
# of pronouns
# of interjections
normalized # of occurrence of special characters (“@”, “#”, “$”, “%”, “&”, “*”, “~”, “^”, “-“, “=”, “+”, “>”, “<”, “[“, “]”, “{“, “}”, “|”, “\”, “/”)

Regarding the second feature set, it consists of theoretical sociolinguistic markers of linguistic differentiation, which have been calculated as quantitative classification features, used in previous studies on gender classification [21]. The sociolinguistic features used in this evaluation are presented in Table 2, they are included in V_{SL}, which has length equal to 6.

Table 2. The sociolinguistic classification features.

Sociolinguistic features
syntactic complexity (normalized # of the sentence’s words)
of slang types per tweet
of bad words per tweet
vocabulary richness (normalized # of different words per tweet without the stop words)
sentimental language (normalized # of positive & negative words per tweet according to SentiWordNet [7])

For the third category, we implemented a number of features that are related to the age linguistic variation and could be useful in order to distinguish the writing style among the different age classes of the authors. As has been reported in [4], [15], with increasing age there is a more frequent use of future tense and fewer self-references. Therefore, these content-based features are the following: the *normalized number of future tense uses*, e.g. “will”, “going to”, “gonna”; the *normalized number of self-references* [16], e.g. “I”, “me”, “myself”; the *normalized number of hyperlink uses* [20]. The V_{CT} feature vector contains the above content-based features and has length equal to 3.

The concatenation of the three feature vectors results to V, as described above. Thus, for each Twitter post one final feature vector, V, is constructed, which has length equal to 40+6+3=49.

For the estimation of the above text features we used the NLTK [23] open-source toolkit.

3.2 Classification Algorithms

For the classification stage, we used a number of dissimilar machine learning algorithms, which are well studied and have extensively been used in several text classification tasks. In particular, we used:

- a multilayer perceptron neural network (MLP), using the back-propagation algorithm for training and three layers,
- the support vector machines (SVMs) using the sequential minimal optimization algorithm, which was tested using two different kernels, namely the radial basis kernel (rbf) and polynomial kernel (poly),
- four tree algorithms, namely the: pruned C4.5 decision tree (J48); the random tree (RandTree)constructing a tree that considers K randomly chosen attributes at each node; the random forest (RandForest) constructing a multitude of decision trees and the fast decision tree learner (RepTree) that builds a decision tree using information gain or variance and prunes it using reduced-error pruning with back-fitting.
- from the Bayesian classifiers, we used: the Bayes network learning (BayesNet) which is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph and the naive Bayes multinomial updateable (NBMU), in which feature vectors represent the frequencies with which certain events have been generated by a multinomial.
- two meta-classifiers: the Adaboost.M1, which is a boosting algorithm combined with Decision Stump and a bagging algorithm (Bagging) using a REPTree, aiming to reduce variance.

All classifiers were implemented using the WEKA toolkit [25]. For all algorithms, the free parameters that are not reported were kept in their standard values.

4 Age Class Identification Evaluation

The evaluation methodology for Twitter users' age class identification presented in the previous section was evaluated using the text-based features and the classifiers presented above. In order to avoid overlap between training and test subsets a 10-fold cross validation evaluation protocol was followed.

4.1 Dataset Description

For the present evaluation we collected and annotated posts from Twitter. Our dataset consists of 19,377 Twitter posts, written in English. The size of the corpus is 247,925 words. The number of the characters is equal to 1,486,681. The posts were divided in 6 age classes and each class corresponds to a different age range of the authors. The distribution of the age classes is tabulated in Table 3.The collection of the data includes Twitter posts from 46 different anonymous users.

Table 3. Distribution of the age classes.

Age Class	Age Range	Number of posts
A	14 - 19	1,513
B	20 - 24	2,149
C	25 - 34	1,640
D	35 - 44	6,433
E	45 - 59	4,662
F	> 60	2,980

4.2 Results

The experimental results for the evaluated classification algorithms in terms of correctly classified web posts (i.e. accuracy), the proportion of actual positives which are correctly identified as such (i.e. sensitivity) and the proportion of negatives which are correctly identified as such (i.e. specificity) are tabulated in percentages in Table 4. The best performance for each of the above metrics is indicated in bold.

Table 4. Evaluation results for age class identification.

Classifier	Accuracy	Sensitivity	Specificity
MLP	57.63	57.60	86.40
SVM-poly	52.82	52.80	81.00
SVM-rbf	52.80	52.80	81.00
J48	47.36	47.36	85.70
RandForest	**61.00**	**60.60**	84.40
RandTree	41.87	41.90	83.90
Bayes Net	42.61	42.60	**86.50**
NBMU	41.59	41.60	77.60
REPTree	50.65	50.78	82.90

Bagging with REPTree	56.56	56.60	84.30
AdaBoost.M1	36.26	36.30	71.00

As can be seen in Table 4, the best performance both in terms of accuracy and sensitivity was achieved by the Random Forest algorithm. Specifically RandForest algorithm achieved 61% accuracy and 60.60% sensitivity, followed by the MLP neural network classifier with accuracy equal to 57.63% and the Bagging with accuracy equal to 56.56%. The worse performing algorithms are the NBMU, RandTree and BayesNet with accuracies equal to 41.59%, 41.87% and 42.61% respectively.

As can be observed from Table 4, the specificity performance is quite high for all the classifiers used and reaches up to 86.50% for the Bayes Net algorithm, i.e. there is a low percentage of false positive instances in the identification of the age class.

The discriminative algorithms that were evaluated, namely the multilayer perceptron neural network, the support vector machines using radial basis and polynomial kernel and the boosting algorithm, in average achieved an accuracy and sensitivity percent equal to 49.9% and a specificity percent equal to 79.8%. The tree algorithms, namely the J48, the random forest, the random tree, RepTree and Bagging, in average succeeded 51.5% in terms of accuracy and sensitivity and 84.2% in terms of specificity. The two probabilistic classifiers, Bayes Net and multinomial Naive Bayes, in average achieved accuracy and sensitivity equal to 42.1% and specificity equal to 82%. In average, the discriminative algorithms had the lowest value of specificity, while the tree algorithms had the highest specificity. Also, in average, as regards the accuracy and the sensitivity, the tree algorithms outperformed the rest of the algorithms, while the probabilistic classifiers had the lowest value in these metrics. The superior performance of the tree algorithms can be explained by the fact that they are robust, scalable and can perform well with large datasets, while the probabilistic classifiers are not as powerful when the dimensionality increases.

Finally, for the RandomForest and Bagging algorithms the results were quite good, while the AdaBoost algorithm performed poorly, since it returned the lowest values in all the three metrics that we examined. This can be explained by the fact that, although this algorithm may dramatically improve performance, sometimes over-fits. In contrast to this, over-fitting is not a problem for Random Forests, which are not very sensitive to outliers in training data. Their superiority is also owed to the fact that they do not need pruning trees and are easy to set parameters since accuracy and variable importance are generated automatically [25].

5 Feature Ranking and Sociolinguistic Analysis

After the classification experiments the competence of each feature was investigated, in order to highlight the most efficient and informative features and/or feature types for the age identification task. We used a Relief feature selection algorithm [9], which is heuristics-independent, noise-tolerant, robust to feature interactions and it runs in low-order polynomial time. In our case we used the updated ReliefF algorithm proposed by Koronenko et al. [10], which improves the reliability of the probability approximation, it is robust to incomplete data, and generalized to multi-class problems. Our dataset was processed by the ReliefF algorithm, implemented using the WEKA machine learning toolkit [25], and feature ranking scores were estimated. The feature ranking results are tabulated in Table 5.

Table 5. The first 20-ranked classification features.

Ranking	ReliefF Score	Feature Description
1	0.0265742	normalized # of words that start with a capital letter
2	0.0153342	normalized # of occurrence of special characters
3	0.0131761	normalized # of hyperlink uses
4	0.0098577	# of punctuation symbols
5	0.0093385	# of proper nouns
6	0.0086022	standard deviation of the word length
7	0.007794	normalized # of characters in capital
8	0.007241	hapax legomena
9	0.0071615	normalized # of short words
10	0.0069192	# of adjectives
11	0.0065469	# of verbs
12	0.0063017	maximum word length
13	0.0062667	# of nouns
14	0.0060678	hapax dislegomena
15	0.0059304	# of occurrence of each alphabetic character
16	0.0056742	average word length
17	0.0054966	# of acronyms
18	0.0053539	# of function words
19	0.0053327	# of pronouns
20	0.0052823	# of prepositions

The ranking of the classification features demonstrates the importance and efficacy of the text mining features. Among the 20 first ranked features, 19 of them belong to the text mining category of characteristics. Only one content-based feature is highly ranked, in the third place, the normalized number of hyperlink uses. We observe though that the sociolinguistic features are not informative for the current task of the age identification. We may conclude that the characteristics based on theoretical markers of gender linguistic variation are not that significant for the age linguistic discrimination. Although the linguistic choice of slang types by teens appears to be a common ground among sociolinguists, in our experiments it is not detected as an important feature.

An important parameter to consider concerning the features selected is the text type of the dataset: our data collection consists of texts limited to 140 characters, a fact that influences the linguistic attitude of the user, by making different choices in morphological, lexical and syntactic level during formulating his message, in order to achieve the less semantic loss possible. This fact could explain the importance of features related to the differentiated use of proper nouns, acronyms and capitalized forms by the age classes, instead of other forms as stop words, articles, interjections, etc. that are probably eliminated as easily-meanings. The POS-tags features (verbs, adjectives, nouns, etc) and their discriminative use by the different age categories, demonstrate that people of different ages produce different structures even in tweet level. The variation in lexical density (the contrast between function and content words in a text) proves that this marker is an important clue to be further investigated in age linguistic variation. Finally, the importance of the word length features evinces the theoretical markers discussed in [1], that teens use smaller lexical forms that adults, and short, semantic and other, neologisms.

Several of the most important features, according to their ReliefF scores, may be grouped in terms of their connection and the conclusions derived: the features that are related to the use of capitalized forms (1, 5, 7 according to their ranking), the features related to the sentence's lexical density (10, 11, 13, 17, 18, 19, 20) and the features related to the lexical forms' length (6, 9, 12, 16). There exists also a group of features that are related to the non-linguistic choices the user make: the use of special characters, punctuation, hyperlinks, and other characters, that may be connected and should be further examined. We observe though and differences in the use of hapax and dislegomena, but this might be a challenging task due to the ambiguity in interpreting the features: either it consists of neologisms and slang that the dictionaries used cannot recognize, which lead to a teen marker, or it consists of more complex forms that older people know and use.

6 Conclusion

In the present article we investigated the appropriateness of a number of machine learning algorithms for classification on the task of age class identification of web users. For the evaluation we relied on several text-mining, sociolinguistic and content-based features extracted from Twitter posts. The experimental results showed that Random Forest classification algorithm outperformed the rest of the evaluated algorithms in terms of accuracy and sensitivity, while the BayesNet had the best performance in terms of specificity. Moreover, the evaluation results showed that in general the decision tree algorithms perform well on the task of age class identification from web text. We used three different feature sets, each one inspired by different tasks, and we evaluated their efficacy during the classification experiments, using the ReliefF algorithm for the feature ranking process. The 20 most informative features are derived and analyzed, in terms of their efficiency in age linguistic variation. Subcategories among the most important characteristics have arisen resulting to major differentiations between the different age groups in lexical density, capitalized forms, word length, and non-linguistic clues.

References

1. Androutsopoulos, J. K., & Georgakopoulou, A. (Eds.). (2003). *Discourse constructions of youth identities* (Vol. 110). John Benjamins Publishing.
2. Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*,*12*(9).
3. Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, *52*(2), 119-123.
4. Barbieri, F. (2008). Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics* 12(1), pp 58--88.
5. Burger, J. D., & Henderson, J. C. (2006, March). An Exploration of Observable Features Related to Blogger Age. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (pp. 15-20).
6. Eckert, P. (1997). Age as a sociolinguistic variable. *The handbook of sociolinguistics*, 151-167.
7. Esuli, A., & Sebastiani, F. (2006, May). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC* (Vol. 6, pp. 417-422).
8. Goswami, S., Sarkar, S., & Rustagi, M. (2009, March). Stylometric analysis of bloggers' age and gender. In *Third International AAAI Conference on Weblogs and Social Media*.
9. Kira, K., and Rendell, L. 1992. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*.
10. Kononenko, I. (1994, January). Estimating attributes: analysis and extensions of RELIEF. In *Machine Learning: ECML-94* (pp. 171-182). Springer Berlin Heidelberg.
11. Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2015). Computational Sociolinguistics: A Survey. *arXiv preprint arXiv:1508.07544*.
12. Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). " How Old Do You Think I Am?"; A Study of Language and Age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. AAAI Press.
13. Nguyen, D., Smith, N. A., & Rosé, C. P. (2011, June). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 115-123). Association for Computational Linguistics.
14. Patra, B. G., Banerjee, S., Das, D., Saikh, T., & Bandyopadhyay, S. (2013). Automatic Author Profiling Based on Linguistic and Stylistic Features.*Notebook for PAN at CLEF*.
15. Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: language use over the life span. *Journal of personality and social psychology*, *85*(2), 291.
16. Pfeil, U., Arjan, R., & Zaphiris, P. (2009). Age differences in online social networking–A study of user profiles and the social capital divide among teenagers and older users in MySpace. *Computers in Human Behavior*, *25*(3), 643-654.
17. Prasath, R. R. (2010, June). Learning age and gender using co-occurrence of non-dictionary words from stylistic variations. In *Rough Sets and Current Trends in Computing* (pp. 544-550). Springer Berlin Heidelberg.
18. Rosenthal, S., & McKeown, K. (2011, June). Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 763-772). Association for Computational Linguistics.
19. Rustagi, M., Prasath, R. R., Goswami, S., & Sarkar, S. (2009). Learning age and gender of blogger from stylistic variation. In *Pattern Recognition and Machine Intelligence* (pp. 205-212). Springer Berlin Heidelberg.

20. Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006, March). Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (Vol. 6, pp. 199-205).
21. Simaki, V., Aravantinou, C., Mporas, I., & Megalooikonomou, V. (2015, September). Using Sociolinguistic Inspired Features for Gender Classification of Web Authors. In *Text, Speech, and Dialogue* (pp. 587-594). Springer International Publishing.
22. Linguistic Inquiry and Word Count. http://www.liwc.net/.
23. http://www.nltk.org/.
24. http://www.adweek.com/socialtimes/social-media-statistics-2014/499230
25. http://www.cs.waikato.ac.nz/ml/weka/