American Journal of Engineering and Applied Sciences

Original Research Paper

Evaluation and Sociolinguistic Analysis of Text Features for Gender and Age Identification

[1,2]Vasiliki Simaki, [3]Iosif Mporas and [4]Vasileios Megalooikonomou

[1]*Department of Computer Science, Linnaeus University, Växjö, Sweden*
[2]*Centre for Language and Literature, Lund University, Lund, Sweden*
[3]*School of Engineering and Technology, University of Hertfordshire, Hatfield, UK*
[4]*Multidimensional Data Analysis and Knowledge Management Laboratory,*
Dept. of Computer Engineering and Informatics, University of Patras, 26500- Rion, Greece

Article history
Received: 03-08-2016
Revised: 21-09-2016
Accepted: 22-09-2016

Corresponding Author:
Vasiliki Simaki
Centre for Languages and Literature, Lund University, Lund, Sweden
Email: vasiliki.simaki@englund.lu.se

Abstract: The paper presents an interdisciplinary study in the field of automatic gender and age identification, under the scope of sociolinguistic knowledge on gendered and age linguistic choices that social media users make. The authors investigated and gathered standard and novel text features used in text mining approaches on the author's demographic information and profiling and they examined their efficacy in gender and age detection tasks on a corpus consisted of social media texts. An analysis of the most informative features is attempted according to the nature of each feature and the information derived after the characteristics' score of importance is discussed.

Keywords: Sociolinguistics, Text Mining, Feature Ranking, ReliefF Algorithm, Gender Detection, Age Identification

Introduction

People of different gender and age tend to have also different linguistic attitudes. They make different choices in language, strongly related to the influence their social identity has in the use of language. These differentiated choices have been the subject of sociolinguistics, an active field of theoretical and empirical linguistics. Sociolinguistics study how the social identity of a speaker affects his linguistic attitude and markers of these choices indicating the specific social group (e.g. a woman, a teenager, etc.) are proposed. The information about a social group's linguistic attitude can be of great importance, especially nowadays, with the expansion of social media: except research purposes (linguistics, sociology and anthropology), it can be a powerful tool for marketing, advertising, forensics, e-government services and applications.

In the present study, the authors are interested on the demographic information of users (gender and age) and how this can be derived by linguistic clues only. Characteristics used in a wide set of text mining tasks are investigated and collected and their efficacy for the age and gender identification task is evaluated. Studies in author's identification are overviewed carefully and each possible statistical or other feature has been collected in order to be tested for the age and gender detection. The most significant clues have been associated to existing sociolinguistic markers and conclusions about the nature of important features of gendered and age differentiated linguistic choices have been derived. This social dimension of the language used in social media has been of a great interest recently and a new research field combining the sociolinguistic theory and the text mining techniques has arisen, the Computational Sociolinguistics (Nguyen *et al*., 2015).

Besides the evaluation of feature informativity for age and gender identification, another issue has been examined: does the knowledge of a social variable (gender or age) is helpful through an identification process of the other variable? In other words, could the gender be calculated as a classification feature in an age identification task and vice versa, is age a differential characteristic of gender identification? As mentioned above, the social factors interact with the user's linguistic attitude and more specifically men and women, as well as teens, adults, elders, perform differently in oral and spoken discourse. In each life stage people adopt diverse linguistic choices affected also by their gender. Consequently, gender and age are variables that, for each case study, are fixed and constant. It becomes accordingly, the subject of investigation if among these two variables exists a dependence relationship. From the search of a possible association among these two variables, further and more specific information about the users' social identity and their linguistic attitude may be extracted.

The remainder of this paper is as follows. In Section 2, an overview of the most important studies on automatic gender and age identification is presented. Section 3 describes the present study's methodology and in Section 4 the dataset and the features used for the experiments are described. Section 5 presents the results of the feature ranking experiments and an analysis of these results is attempted. Finally, Section 6 concludes the present paper by summarizing the most important findings of the study.

Related Work

In this chapter, studies exploring the gender and age identification are described, implemented with various machine learning techniques. The research works at this field are divided into three major categories: studies exploring the author's gender only, approaches around age only and studies exploring multiple social factors, including gender and age.

Author's Gender Identification

The automatic identification of the author's gender has been typically perceived as a text classification task (Cheng *et al*., 2011; Soler and Wanner, 2014). It consists of computational methods focusing in machine learning techniques and algorithms, for the most accurate possible performance for the attribution of the gender demographic information to an anonymous author.

Newman *et al*. (2008) analyzed a collection of 14,000 texts from different sources of written and spoken discourse, in their effort to highlight the most important differences in the language system usage between women and men. To the question posed if men use the language differently than women and if context plays any role in that, the answer is complex, due to the social and psychological- depended factors, which are not related to the linguistic system. Their research on function and content words ended to the conclusion that small but systematic differences are traced between women's and men's language and their basic finding is that women use more words that are related to psychological and social processes, while men refer to more objective and impersonal topics. Argamon *et al*. (2006) conducted a similar research in a subset of British National Corpus, (British National Corpus (BNC): http://www.natcorp.ox.ac.uk/) a collection of literary texts. The authors used a wide set of syntactic and other features, in order to detect significant gendered linguistic differentiations, observing that men choose more frequently the use of articles and particles, while women tend to use more personal pronouns, modal verbs and conjunctions.

Koppel *et al*. (2003) proposed text classification methods in order to extract after formal texts the authors' gender, using n-grams and function words as classification features, which are quite standard clues in authorship attribution. This study combined stylometric and classification techniques in order to achieve accuracy around 80% in author's gender identification. In the case though, that the text genre is defined, the accuracy of the results is up to 98%.

Sarawgi *et al*. (2011) explored the author's gender in both scientific and web blog texts. They used statistical and machine learning methods, in both text types, without topic and genre bias. After several approaches, the comparative results led to the most reliable method, which is based mostly in character language models trained on morphological patterns than in token language models trained on lexico-syntactic patterns. Their study results the detection of gendered linguistic choices, without taking into account the topic and the genre of the text, either the literary or the web blogs texts. Corney *et al*. (2002), in one of the first attempts in gender identification, investigated in a collection of e-mail texts from different authors how stylistic markers, structural features and gender preferential linguistic characteristics identify correctly the author's gender, achieving 70% classification accuracy. An important observation is that the gender preferential linguistic characteristics in this method did not improve much the classification accuracy.

Many studies in gender identification focused their interest in texts derived from web blogs. Kobayashi *et al*. (2007) used a wide set of SVM algorithms and word weight metrics, estimated the bloggers' gender from their own posts by deriving words which use is related to male or female choices. In 92% of the bloggers, posts are categorized with 85% of accuracy, when in 83% of the total number of posts, they achieve 90% of accuracy. Zhang and Zhang (2010) attempted the same problematic using diverse features as words, POS tags, etc. and they combined them to different algorithms and approaches. Their best results are achieved with the SVM linear kernel algorithm with feature selection, surpassing the 72% classification accuracy.

Mukherjee and Liu (2010) created a corpus of web blog texts and after tracing sets of words, word categories, POS tags, n-gram features, they performed feature selection and classification experiments achieving more than 88% accuracy. The identification of the author's gender has been investigated in data collections constituted by users' posts on Facebook. In a corpus of 170,000 posts, Keeshin *et al*. (2010) calculated statistical features based on word, structural and frequencies clues, using diverse machine learning techniques. Similarly, Holgrem and Shyu (2013), used machine learning methods with feature vectors from word metrics, in a dataset consisting Facebook users' posts.

Burger *et al.* (2011) studied the Twitter users' gender, after his tweets. They combined the tweet content to the user's name and to any other information related to the user and they achieve the same accuracy of the automated process as if human evaluators performed the gender attribution manually. Miller *et al.* (2012) predicted, from a Twitter dataset, the user's gender using Stream Algorithms by calculating the most important (by their grade of informativity) character n-gram features. Bamman *et al.* (2014) attempted the gender recognition of Twitter users and explored the linguistic variation highlighted through the data used. In their study, authors are grouped into clusters, the prediction results are analyzed and lexical frequencies features into a wide set of stylistic descriptors indicating gender preferential choices that attribute clues of participations or information.

Most of the studies on automatic gender identification focused on statistical features of gendered linguistic differentiation, on character, word, POS tag level. Sociolinguistic research on the other hand, has indicated the significance of qualitative linguistic markers that are important in gender linguistic variation. In a preliminary study, Simaki *et al.* (2015a) proved that the use of sociolinguistic-based features additionally to a wide set of statistical features can improve the gender classification accuracy.

Age Identification

The studies focusing on the author's age identification are less numerous than the previous category. This can be explained in terms of difficulties either in accepting universal age classes and life stages, or in the continuous values range of the authors' exact age. Burger and Henderson (2006) investigated the evolution of the web blog posts' form through time and they attempted the prediction of the author's age (based on the birth date). The researchers observed that the documents' size (the total number of words per post) is a distinctive characteristic and they calculated the occurrence rate of punctuation, capitalized letters and spaces. Simaki *et al.* (2015b) performed age class identification experiments using regression algorithms, in the "Blog Authorship Corpus", from a wide set of text mining features.

Tam and Martell (2009) implemented text classification experiments in terms of age using Bayesian and SVM classifiers. They extracted character n-grams and word meta-data as features, in order to classify the "NPS Chat Corpus" (NPS Chat Corpus: http://faculty.nps.edu/cmartell/NPSChat.htm) into five age classes. Other studies on age prediction (Rosenthal and McKeown, 2011; Nguyen *et al*., 2011) proved that the content and the stylistic features are clues of great importance and when the author's online activity is added, the classification accuracy rises to 80%. Nguyen *et al.* (2013), subsequently, studied the linguistic use among different age categories of Twitter users. Their analysis highlighted differences in style, the references, the conversations and the notifications, clues that are dependent not only from the estimation of the age class, but also from the author's life stage and his exact age.

Gender and Age Identification

The investigation of more than one demographic clues of authors in their texts, has been an active research field the last few years. Schler *et al.* (2006) created the "Blog Authorship Corpus", a text collection annotated with the author's gender and age and in some cases with more information about the blogger's identity. They used stylistic and content-based features, in order to trace the author's gender and age. They observed that specific forms and unigrams are more frequently used by young bloggers and that the writing style differs significantly among the age groups according to their classification in 10's, 20's, 30's. Argamon *et al.* (2007) used the same corpus, in order to refine in the gender and age identification task. They used stylistic and content-based features, highlighting gendered and age lexical choices in order to prove the linguistic variation among different genders and ages. They associated the gender and age findings, which are close (316 statistically significant common words out of 1,000 words with informational weight for the gender and 1,000 for the age) and they assumed that deeper differences in language usage and communication are underlined (introvert and extrovert among the correlated categories).

Goswami *et al.* (2009) performed a stylometric analysis in terms of gender and age, using out-of-dictionary forms and the sentence length as features. Slang, emoticons, out-of-dictionary forms, abbreviations in online conversations and the sentence length proved to be significantly differentiated among different age classes and genders. Peersman *et al.* (2011) performed a gendered and age classification in small texts, using features based on characters and words frequently used in online chat and they achieved accuracy higher than 88%.

The identification of demographic information and generally of the author's profile has been the research subject of PAN 2013 (PAN 2013: http://www.uni-weimar.de/medien/webis/events/pan-13/pan13-web/index.html), where studies using different sets of features were presented. These features may be finally grouped as follows: stylistics, content-based, n-grams, IR, collocations. Among other researchers, Flekova and Gurevych (2013) focused on gender and age identification using shallow linguistic features (located at the morphology-lexicon level of linguistic analysis), syntactic, punctuation, readability, semantic, context, lexical and stop-words features. They observed that the gendered and age profile are independent issues, but they are determined by the same characteristics. Rangel and Rosso (2013) used the "PAN-AP-13" dataset in order to implement gender and age classification experiments, using cognitive features after neurological studies. This approach proved to be more effective for the age estimation than the gender identification, highlighting the age linguistic differentiations in English and Spanish datasets.

Schwartz *et al.* (2013) conducted an integrated study on personality profile, gender and age of Facebook users. Standard techniques were implemented and a particular method for the data linguistic analysis and the evaluation was proposed with convincing results and important feedback to future interdisciplinary studies on the field. The identification of the author's demographic profile has been the research subject of multilingual efforts: Amasyali and Diri (2006) performed a gendered and age classification in Turkish text data and Verhoeven and Daelemans (2014) investigated diverse variables, among others the gender and the age, in a Dutch corpus.

After the presentation of the existing knowledge in the field of automatic identification of the author's gender and age, the methodology followed at the present study is consequently described.

Methodology

In this study, the researchers examined the efficacy of linguistic features, used in diverse text mining investigations about the author's profile (authorship attribution, gender classification, age identification). They attempted an interdisciplinary study, combining the sociolinguistic knowledge on gendered and age linguistic variation to the existing text mining techniques for the author's profile exploitation.

A large set of standard text mining features is created and sociolinguistic-inspired and content-based features are also calculated in an annotated corpus consisting of texts derived from social media. These features are extracted and ranked according to their contribution in author's gender and age detection. For the feature ranking process the ReliefF algorithm (Kononenko, 1994) was selected. Consecutively, the results are presented, analyzed and discussed under the scope of sociolinguistic theory on gendered and age variation. It has been also examined if the highest ranked features can be paired to existing sociolinguistic markers. The labels of gender and age are also used as features in the corresponding searches and their efficacy is attempted to be explained. The researchers tried to answer finally the following question: In a text, could the a priori knowledge of a variable (gender) assist the detection process of the other variable (age) as a classification feature and vice versa?

In the following section, we describe the dataset used in the present study and the set of the different features which have been evaluated in the given task.

Dataset and Features

Dataset Description

In the present study the "Blog Authorship Corpus" (Schler *et al.*, 2006) was used, which is publicly available. It consists of a post collection in web blogs by 19,320 bloggers. These posts were extracted from blogger.com on August 2004. The corpus' size is 681,288 posts containing more than 140 million words, which corresponds to 35 posts and 7,250 words per blogger. The authors are grouped into three age classes: 10's, 20's and 30's. The first category (10's) contains the posts of 8,240 web blogs, written by authors between 13 and 17 years old. The second category (20's) contains the posts of 8,096 web blogs, written by authors between 23 and 27 years old. Finally, the third category (30's) contains the posts of 2,994 web blogs, written by authors between 33 and 47 years old.

Each blog is structured in a separate file, containing the blogger's posts. It is annotated with the blogger's number id, his/her gender, his/her exact age (apart the age class the blogger belongs to) and, in cases it was possible, any further anonymous personal information could be extracted. The "Blog Authorship Corpus" is primarily selected for the present study due to its annotation with both gender and age information. A second reason that attracted the researchers' interest was the text type of the posts: informal and spontaneous text samples produced by social media users. This can provide to the study more information about the author's profile, due to the clues of oral discourse that "slip" into the bloggers' posts.

Feature Extraction

A wide set of different features has been selected for this study and these clues can be grouped as follows:

- The statistical features, forming a feature vector F^{STAT} equal to 30, as presents below in Table 1
- The POS-tags features, forming a feature vector F^{POS} equal to 9, as presents below in Table 2
- The content-based features, forming a feature vector F^{CB} equal to 3: the normalized number of future tenses, the normalized number of self-references and the normalized number of hyperlink uses
- The gender feature, forming a feature vector F^{GEND} equal to 1
- The age feature, forming a feature vector F^{AGE} equal to 1

The last two features (gender and age) came from the given labels after the corpus annotation and they are alternately examined accordingly to the variable investigation: for the gender identification, the age feature is calculated and for the age identification, the gender feature is taken into account. The union of all previous vectors creates the super vector *F* of total size 30+9+3+1+1 = 44.

Table 1. The statistical features used in the study

Statistical features
average # of characters per sentence
average # of words per sentence
normalized # of different words
of words that appear once in the document (hapax legomena)
of words that appear twice in the document (hapax dislegomena)
average # of sentences per paragraph
of function words
of punctuation symbols (".", ",", "!", "?", ":", ";", """, """)
average # of characters per paragraph
normalized # of words that start with a capital letter
normalized # of emoticons
normalized # of words whose letters are all capital
STD of the word length
maximum word length
minimum word length
of characters per web post
normalized # of characters in capital
normalized # of alphabetic characters
normalized # of space characters
of words that consist of less than 4 characters (short words)
of occurrence of each alphabetic character
normalized # of digit characters
normalized # of occurrence of special characters ("@", "#", "$", "%", "&", "*", "~", "^", "-", "=", "+", ">", "<", "[", "]", "{", "}", "|", "\", "/")
normalized # of tab ("\t") characters normalized
normalized # of characters per word
total # of words
average word length
of sentences
of paragraphs
of lines

Table 2. The POS-tags features used in the study

POS tags features
of nouns
of proper nouns
of adjectives
of prepositions
of verbs
of pronouns
of interjections
of adverbs
of articles

In the section below, the feature ranking experiments are presented and the results on the features' informativity are discussed.

Evaluation of Different Features on Gender and Age Identification

Feature Ranking

After the feature extraction process, the competence of each feature was investigated, in order to highlight the most efficient and informative features and/or feature types for the gender and age identification tasks. A Relief feature selection algorithm (Kira and Rendell, 1992) was used, which is heuristics-independent, noise-tolerant, robust to feature interactions and it runs in low-order polynomial time. For the present case the updated ReliefF algorithm proposed by Koronenko (1994) was used, which improves the reliability of the probability approximation, it is robust to incomplete data and generalized to multi-class problems. The dataset was processed by the ReliefF algorithm, implemented using the WEKA (WEKA: http://www.cs.waikato.ac.nz/ml/weka/) machine learning toolkit and feature ranking scores were estimated. The feature ranking results for the age identification task are tabulated in Table 3 and 4 for the gender identification task.

The first observation is about the number of the significant and informative features for the age and the gender identification. After the investigation of the informativity grade in the feature set for the age estimation, 18 out 43 features proved to be significant, when the rest 25 appear to be useless in a task around age identification. Among the 18 most informative features, the gender characteristics is detected and ranked at the 14th place. Accordingly, in the investigation of the features' importance when the gender detection task is tried, it appears that 37 features are informative enough and only 6 from the initial set of 43 features proved to be statistically non-significant. In that case the age feature appears at the 4th position.

A first observation can be formulated as follows: the feature set used proved to be more effective in the gender identification task, highlighting more informative features, than during the age identification task. This may be easily explained due to the fact that many of the statistical and POS features have been arisen through gender classification studies, where they are mostly used. It is rational though, to be proven as statistically significant when they are evaluated towards their informativity through a gender detection task than through an age detection task. Another clue that could explain the small number of important features for the age identification is related to the to the experiments' data set: the data used in the present study do not correspond to a continuous age range, given that age gaps exist among the pre-defined age classes, which makes the sample having many missing values. That fact could possibly mispresent the informativity of the evaluated features.

After the feature ranking experiments, a study of the evaluated features is tried in the section bellow and an analysis in terms of sociolinguistic principles of the most informative ones is attempted, in order to derive any correlation among gender and age linguistic variation and text mining techniques for gender and age automatic detection.

Table 3. The feature ranking results for the age identification task

Ranking	ReliefF Score	Feature description
1	0.000345547	chars_paragraph
2	0.00032066	acronyms
3	0.000256627	capitalized
4	0.000217562	avg_word_length
5	0.000182655	sents_paragraph
6	0.000174966	std_word_len
7	0.000098023	num_different_words
8	0.000061215	articles
9	0.000053754	short_words
10	0.000050792	emoticons
11	0.00004542	tabs
12	0.000035037	future_tense
13	0.000032787	prepositions
14	0.000028831	gender
15	0.000023995	adverbs
16	0.000018702	pronouns
17	0.000000426	punctuation
18	0	min_word_len
19	-0.000001558	links
20	-0.000002004	letter_frequency
21	-0.000004697	digits
22	-0.000005964	Characters
23	-0.000007713	function_words
24	-0.000014724	verbs
25	-0.000017899	spaces
26	-0.000020011	chars_in_words
27	-0.000021881	words
28	-0.000024073	adjectives
29	-0.00002818	alphab_chars
30	-0.000037205	max_word_len
31	-0.000037788	nouns
32	-0.000042111	self_references
33	-0.000055621	sentences
34	-0.000057555	hapax_dis
35	-0.000082602	lines
36	-0.000082602	paragraphs
37	-0.000084395	proper_nouns
38	-0.000088401	avg_word_sentence
39	-0.000091581	avg_char_sentence
40	-0.000091734	upper_case_chars
41	-0.000094306	special_chars
42	-0.00013412	hapax_leg
43	-0.000186145	interjects

Table 4. The feature ranking results for the gender identification task

Ranking	ReliefF Score	Feature description
1	0.00047653	pronouns
2	0.00023575	articles
3	0.00022174	adverbs
4	0.00021752	age
5	0.00021488	avg_word_length
6	0.0001852	self_references
7	0.00018123	short_words
8	0.0001741	verbs
9	0.00014442	hapax_dis
10	0.0001407	sentences
11	0.00012399	std_word_len
12	0.00011971	hapax_leg
13	0.00010492	sents_paragraph
14	0.00009957	num_different_words
15	0.00008918	interjects
16	0.00008609	chars_paragraph
17	0.00008184	words
18	0.0000779	proper_nouns
19	0.00007701	prepositions
20	0.00007117	punctuation
21	0.00006736	Characters
22	0.00006566	nouns
23	0.00006379	letter_frequency
24	0.00005759	function_words
25	0.00005159	adjectives
26	0.00004718	special_chars
27	0.00004149	chars_in_words
28	0.0000317	avg_char_sentence
29	0.00002085	avg_word_sentence
30	0.00002018	paragraphs
31	0.00002018	lines
32	0.00001912	max_word_len
33	0.00001857	spaces
34	0.0000161	acronyms
35	0.00001134	tabs
36	0.00000154	links
37	0	min_word_len
38	-0.00000479	digits
39	-0.00000704	upper_case_chars
40	-0.00001195	emoticons
41	-0.0000202	future_tense
42	-0.00004414	capitalized
43	-0.00004933	alphab_chars

Feature Analysis

In this part of the study, an analysis of the most informative features is attempted. As described above, for the age detection task only 18 over a set of 43 features appeared to be important enough. In Table 5 the 18 higher ranked features are tabulated for both the age and the gender identification tasks, despite the fact that in the gender identification task the informative features are more than 18.

The first observation, according to Table 5, is that 9 out of the 18 most important features are common among the two tasks. They are ranked in different places and achieved a different grade of informativity, but they are crucial in both gender and age detection. It is clear that they perform differently in each task and their values are not the same according to the variable investigated in each case and the predefined classes.

A more extensive study of the important features that are common in both tasks, may lead to a grouping of these characteristics in terms of the linguistic level of analysis they are located: morphological level, lexical level, syntactic level and context level. The characteristics “avg_word_length”, “std_word_len” and “short_words”, are located at the morphological level of linguistic analysis and demonstrates the importance of the length of words that each age or gendered class chooses to form and proved to be a differential clue in both identification tasks. The "num_different_words" characteristic is located at the lexicon level of linguistic analysis and the "articles", "adverbs", "pronouns" features are syntactic clues of differentiated linguistic choices. Finally, the "char_paragraph" and the "sents_paragraph" are features carrying semantic information, important to the information, context and length of each paragraph.

The morphological features need to be more investigated and analyzed in a set of different tasks for the gender and age identification, in order to enable the outcome of more specific conclusions. Furthermore, it has to be compared the values of these features given a different variable (gender/age). On the other hand, for the syntactic features, we could have an idea about their informatively, based on research in the fields of sociolinguistics and the automatic gender and age identification: the use of pronouns and adverbs are strongly related to the gender and age linguistic variation as age and gendered preferential choices. The new clues though, of the use of articles and prepositions should be in a future work be more investigated and analyzed.

According to the feature sets used for the ranking process, the statistical features appear to be more numerous than the other features. Concerning the age detection task, statistical features based in character calculations (e.g. "chars_paragraph", "avg_word_length", "short_words") proved to be very important and can be associated to existing sociolinguistic knowledge on age linguistic variation. To be more specific, the word length or the short words have been observed as markers distinguishing the linguistic use that teens and adults make (Androutsopoulos and Georgakopoulou, 2003). The features that are based in word counts (e.g. "acronyms", "capitalized forms", "num_diff_words") confirm also the sociolinguistic findings about the different lexical choices that people of different life stages make, provide information about clues that need to be further investigated and identify the exact value of each characteristic depending the corresponding age category. Concerning the POS features, some of them appear to be a useful tool in age identification and one content-based feature observed within the informative clues (the use of future tenses). Finally, the gender feature is among the informative clues and proves that the knowledge of the gender improves the correct age identification of the user's text in social media.

Concerning the informative features about the gender identification task, it appears that the POS features are critical in distinguishing the gendered linguistic choices of a user: the use of adverbs could also be associated to the sociolinguistic marker of "empty" forms carrying a sense of admiration/acceptance (Lakkoff, 1975).

Table 5. The 18-top ranked features for age and gender detection. The features that are common in both tasks are highlighted in bold

Informative features for age detection	Informative features for gender detection
chars_paragraph	pronouns
acronyms	articles
capitalized	adverbs
avg_word_length	*age*
sents_paragraph	avg_word_length
std_word_len	self_references
num_different_words	short_words
articles	verbs
short_words	hapax_dis
emoticons	sentences
tabs	std_word_len
future_tense	hapax_leg
prepositions	sents_paragraph
gender	num_different_words
adverbs	interjects
pronouns	chars_paragraph
punctuation	words
min_word_len	proper_nouns

Table 6. The 18-top ranked features for age and gender detection grouping according to the level of linguistic analysis in which they provide information

Morphological level	**Lexical level**
avg_word_length	num_different_words
std_word_len	capitalized^{AGE}
short_words	acronyms^{AGE}
min_word_len^{AGE}	hapax_dis^{GENDER}
	hapax_leg^{GENDER}
Syntactic level	**Semantic-context level**
articles	chars_paragraph
pronouns	sents_paragraph
adverbs	emoticons^{AGE}
prepositions^{AGE}	tabs^{AGE}
punctuation^{AGE}	future_tense^{AGE}
verbs^{GENDER}	interjects^{GENDER}
words^{GENDER}	self_references ^{GENDER}
proper_nouns^{GENDER}	sentences^{GENDER}
	gender, age

Additionally, the verbs and the words characteristic could be a strong indication of the syntactic complexity and the different syntactic structures that people of different gender make. The age characteristic is highly informative and the content-based self-references characteristics is also among the most important features. As regards the character- and word-based features, they are also numerous as in the corresponding age task and they provide information which is located in morphological, lexical and semantic-context linguistic level. In Table 6, it is tabulated our attempt to group the informative features according to the level of linguistic analysis they could be located and in which they provide information.

Conclusion

In the present paper, an evaluation and sociolinguistic analysis of text features for the tasks of gender and age identification was attempted. A large feature set of standard and novel text features was made and feature ranking experiments were performed. After the results of the experiments, an analysis of the most informative clues for both investigations has been made. One conclusion is that 9 over 18 most informative clues are common in both tasks and the knowledge of gender and age is of great importance as feature for the corresponding investigations. It should be emphasized though, that a more extensive investigation of the values of the informative features during each task should be made, in order to confirm the sociolinguistic indications concluded by the present study to standard theories. Another important conclusion is the fact that each feature carrying a given grade of informativity, provides useful information to a certain level of linguistic analysis and depending the age or gender identification task contributes to the existing sociolinguistic knowledge.

Author's Contributions

All authors contributed equally to the implementation of the present paper.

Ethics

This original research has been conducted with respect to ethics.

References

Amasyali, M.F. and B. Diri, 2006. Automatic Turkish text categorization in terms of author, genre and gender. Proceedings of the 1 1th International Conference on Applications of Natural Language to Information Systems, May 31-Jun. 2, NLDB, Klagenfurt, Austria, pp: 221-226. DOI: 10.1007/11765448_22

Androutsopoulos, J.K. and A. Georgakopoulou, 2003. Discourse Constructions of Youth Identities. 1st Edn., John Benjamins Publishing, Amsterdam, ISBN-10: 1588113558, pp: 338.

Argamon, S., M. Koppel, J. Fine and A.R. Shimoni, 2006. Gender, genre and writing style in formal written texts. Interdisciplinary J. Study Discourse, 23: 321-346. DOI: 10.1515/text.2003.014

Argamon, S., M. Koppel, W. Pennebaker and J. Schler, 2007. Mining the Blogosphere: Age, gender and the varieties of self-expression. First Monday, 12: 1-9. DOI: 10.5210/fm.v12i9.2003

Bamman, D., J. Eisenstein and T. Schnoebelen, 2014. Gender identity and lexical variation in social media. J. Sociolinguist., 18: 135-160. DOI: 10.1111/josl.12080

Burger, J., J. Henderson, G. Kim and G. Zarrella, 2011. Discriminating gender on Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). Association for Computational Linguistics, Stroudsburg, PA, USA, pp: 1301-1309.

Burger, J.D. and J.C. Henderson, 2006. An exploration of observable features related to blogger age. Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp: 15-20.

Cheng, N., R. Chandramouli and K.P. Subbalakshmi, 2011. Author gender identification from text. Int. J. Digital Forensics Incident Response, 8: 78-88. DOI: 10.1016/j.diin.2011.04.002

Corney, M., O. De Vel, A. Anderson and G. Mohay, 2002. Gender-preferential text mining of e-mail discourse. Proceedings of the 18th Annual Computer Security Applications Conference, (CSAC' 02), IEEE Xplore press, pp: 282-289. DOI: 10.1109/csac.2002.1176299

Flekova, L. and I. Gurevych, 2013. Can we hide in the web? Large scale simultaneous age and gender author profiling in social media. In CLEF 2012 Labs and Workshop, Notebook Papers.

Goswami, S., S. Sarkar and M. Rustagi, 2009. Stylometric analysis of bloggers' age and gender. Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media.

Holgrem, J. and E. Shyu, 2013. Gender Classification of Facebook Posts.

Keeshin, J., Z. Galant and D. Kravitz, 2010. Machine Learning and Feature Based Approaches to Gender Classification of Facebook Statuses.

Kira, K. and L.A. Rendell, 1992. The feature selection problem: Traditional methods and a new algorithm. *AAAI* , 2: 129-134.

Kobayashi, D., N. Matsumura and M. Ishizuka, 2007. Automatic estimation of bloggers' Gender. Proceedings of International Conference on Weblogs and Social Media. Boulder: Omnipress.

Kononenko, I., 1994. Estimating attributes: analysis and extensions of RELIEF. Proceedings of the European Conference on Machine Learning Catania, April 6-, Springer Berlin Heidelberg. Italy, pp: 171-182. DOI: 10.1007/3-540-57868-4_57

Koppel, M., S. Argamon and A.R. Shimoni, 2003. Automatically categorizing written texts by author gender. Literary Linguistic Comp., 17: 401-12. DOI: 10.1093/llc/17.4.401

Lakkoff, R., 1975. Language and Woman's Place. Harper & Row, New York.

Miller, Z., B. Dickinson and W. Hu, 2012. Gender prediction on Twitter using stream algorithms with N-gram character features.

Mukherjee, A. and B. Liu, 2010. Improving gender classification of blog authors. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10).

Newman, M.L., C.J. Groom, L.D. Handelman and J.W. Pennebaker, 2008. Gender differences in language use: An analysis of 14,000 text samples. Discourse Processes, 45: 211-236.
DOI: 10.1080/01638530802073712

Nguyen, D., A.S. Doğruöz, C.P. Rosé and F. de Jong, 2015. Computational Sociolinguistics: A Survey. *arXiv preprint arXiv:1508.07544*.

Nguyen, D., N.A. Smith and C.P. Rosé, 2011. Author age prediction from text using linear regression. Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities, Association for Computational Linguistics, pp: 115-123.

Nguyen, D., R. Gravel, D. Trieschnigg and T. Meder, 2013. How old do you think i am? A study of language and age in Twitter. Proceedings of the 7th International AAAI Conference on Weblogs and Social Media. AAAI Press.

Peersman, C., W. Daelemans and L. Van Vaerenbergh, 2011. Predicting age and gender in online social networks. Proceedings of the 3rd International Workshop on Search and Mining user-Generated Contents (MCC' 11), ACM, pp: 37-44.
DOI: 10.1145/2065023.2065035

Rangel, F. and P. Rosso, 2013. Use of language and author profiling: Identification of gender and age. Proceedings of the 10th International Workshop on Natural Language Processing and Cognitive Science (PCS' 13), Marseille, France.

Rosenthal, S. and K. McKeown, 2011. Age prediction in blogs: A study of style, content and online behavior in pre-and post-social media generations. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT' 11), Association for Computational Linguistics, pp: 763-772.

Sarawgi, R., K. Gajulapalli and Y. Choi, 2011. Gender attribution: Tracing stylometric evidence beyond topic and genre. Proceedings of the 15th Conference on Computational Natural Language Learning, 24 June, ACLS, Portland, USA, pp: 78-86.

Schler, J., M. Koppel, S. Argamon and J.W. Pennebaker, 2006. Effects of age and gender on blogging. AAAI Spring Symposium: Computational Approaches Analyzing Weblogs, 6: 199-205.

Schwartz, H.A., J.C. Eichstaedt, M.L. Kern, L. Dziurzynski and S.M. Ramones *et al*., 2013. Personality, gender and age in the language of social media: The open-vocabulary approach. PloS one, 8: e73791. DOI: 10.1371/journal.pone.0073791

Simaki, V., C. Aravantinou, I. Mporas and V. Megalooikonomou, 2015a. Using sociolinguistic inspired features for gender classification of web authors. Proceedings of the 18th International Conference, TSD, Sep. 14-17, Springer International Publishing, Pilsen, Czech Republic, pp: 587-594. DOI: 10.1007/978-3-319-24033-6_66

Simaki, V., C. Aravantinou, I. Mporas and V. Megalooikonomou, 2015b. Automatic estimation of web bloggers' age using regression models. Proceedings of the 17th International Conference Speech and Computer, Sep. 20-24, Springer International Publishing, Athens, Greece, pp: 113-120. DOI: 10.1007/978-3-319-23132-7_14

Soler, J. and L. Wanner, 2014. How to use less features and reach better performance in author gender identification. Universitat Pompeu Fabra, Spain.

Tam, J. and C.H. Martell, 2009. Age detection in chat. Proceedings of the IEEE International Conference on Semantic Computing, (ICSC' 09), IEEE Xplore Press, pp: 33-39.

Verhoeven, B. and W. Daelemans, 2014. CLiPS Stylometry Investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. University of Antwerp, Belgium.

Zhang, C. and P. Zhang, 2010. Predicting gender from blog posts. Technical Report, University of Massachusetts Amherst, USA.