Gender Classification with Data Independent Features in Multiple Languages

Tim Isbister*, Lisa Kaati*† and Katie Cohen*
*Swedish Defence Research Agency (FOI)
Stockholm, Sweden
†Uppsala University
Uppsala, Sweden
Email:firstname.lastname@foi.se

***Abstract*—Gender classification is a well-researched problem and state-of-the-art implementations achieve an accuracy of over 85%. However, most previous work has focused on gender classification of texts written on English language and in many cases the results cannot be transferred to different datasets since the features used to train the machine learning models are dependent on the data. In this work we investigate the possibilities to classify the gender of an author on five different languages: English, Swedish, French, Spanish and Russian. We use features from the word counting program Linguistic Inquiry and Word Count (LIWC) with the benefit that these features are independent of the dataset. Our results shows that by using machine learning with features from LIWC we can obtain an accuracy between 79% and 73% depending on the language. We also show some interesting differences between the uses of certain categories among the genders on different languages.**

I. Introduction

The Internet provides us with hundreds of millions of different kinds of user generated content including blog posts, tweets, Facebook likes and Instagram pictures. The large amount of user generated information includes all kinds of information including different forms of illegal communication such as anonymous threats, terrorism propaganda, and the selling and buying of illegal drugs and substances on hidden services. When the Internet is used for illegal purposes, the users commonly try to hide their identity by using different kinds of social media user accounts that offer some degree of anonymity. In most social media services you create your own profile and yourself can determine the information you reveal about your true identity.

In this work we will focus on the use of machine learning for determining the gender of an author of a blog. We will do experiment on five different languages (English, Swedish, French, Spanish and Russian) to get an understanding on how well the techniques works on the different languages. Determining the gender of an author can be seen as a deanonymization technique - a set of techniques that aims at revealing the physical identity behind an anonymous user or a user making use of a pseudonym (alias). There are several different kinds of techniques that can be used for deanonymization purposes; examples are determining the age of a writer, the education level and the origin or location of the writer.

The focus of this work is to get an understanding of how well data independent features works for classifying the gender of an author on different languages. Our features are obtained from the word counting program Linguistic Inquiry and Word count (LIWC) [1]. LIWC was developed by James W. Pennebaker at the University of Texas and has been evaluated and tested in a number of different studies [2], [3]. LIWC sorts words in psychologically meaningful categories by counting the relative frequencies of words in a text and dividing them into different categories. For each text, a profile describing how much the author uses words from the different LIWC categories (in percent) is created. LIWC dictionaries are currently translated to more than twelve different languages - in this work we will use four translations of LIWC as well as the original dictionaries on English. An interesting aspect of this work is that our experiments allow us to reason about the linguistic differences among genders and whether the differences are present in all languages.

A. Ethical aspects of deanonymization

There are many ethical aspects on developing techniques that pose a threat towards the personal integrity and privacy of Internet users. Deamonymization techniques can be used to support law enforcement and intelligence agencies to solve crimes and detect threats towards the security of the society but it can also be used companies generating personal advertisement. The fact that all deamonymization techniques have many different areas of usage leads to questions such as: is it legal to download data from various forms of social media and use deanonymization techniques to learn more about an author? From a research perspective it is indeed interesting to learn more about how well this kind of methods can work but from another perspective one could argue that the development of this kind of techniques is a severe threat towards the personal integrity and privacy of all kind of social media users. There are no universal guidelines about when and how these kinds of techniques can and should be used and more research regarding these difficult issues is indeed needed.

B. Outline

The rest of this paper is structured as follows. In Section II we discuss differences among linguistic differences between the genders that has been identified in previous research. Section III we describe LIWC and the different versions of LIWC that we use in our experiments. Section IV describes some of the work related to ours and in particular some of the features that have been used in gender classification. Our experiments and experimental setup is described in Section V. The experimental results are discussed in Section VI and the paper is concluded and with some directions for future work in Section VII.

II. Gender and Language

Since Lakoff [4], several studies have examined gender differences in spoken and written language. These differences have been observed in various conversational features such as turn-taking, intonation, pitch etc., but also in word use (e.g. [5]). Though subtle, there is no question that gender differences in language do exist. Whether these differences are caused by biological differences or a result of social conditioning falls outside the scope of the present work. Suffice it to say that these differences exist, and that they can be employed for finding a reliable way to identify the gender of an anonymous writer.

The results of various studies on gender differences in language have been somewhat contradictory. One possible explanation for this is that different studies have used language samples from different contexts, influencing the size and direction of the gender differences. As Newman et al. [6] point out, a frustration of studying natural language is that people use words differently in different contexts. Also, studies of natural language often come out with small differences and large standard deviations, which is why small samples with low statistical power usually do not render reliable results [6]. Hence, large samples may be required both to find differences that are valid across different contexts, and to detect small differences between men's and women's language. A third possible reason for conflicting results across different studies is that different coding schemes have been used. Larger studies where the word count program LIWC is used to find gender differences in word use show relatively consistent results [5].

LIWC studies have proven to be especially fruitful when it comes to function words, the small words that are used to make our sentences grammatically correct. Function words are processed differently in the brain than content words (nouns, adjectives etc.) and they can reflect psychological state independent of content. Differences in the use of function words reflect differences in the ways that individuals relate to the world, other people, and themselves [7]. For example, using more pronouns in general (rather than nouns) suggests a shared reality, in that both parties have to understand whom the referent is to make sense of what is being said. A study by Newman et al. [6] using LIWC to analyze over 14,000 text files from 70 separate studies found that women used more pronouns and verbs than men, and men used more numbers, articles, long words and swear words than women. According to Pennebaker [5] several studies show that women also use more "I-words" than men. In different studies, the use of first-person singular has been associated with age, depression, insecurity and self-focus [7].

III. Linguistic Inquiry and Word count (LIWC)

Pennebaker and his colleagues has shown that it is possible to connect word use to psychological constructs such as personality, drives, cognition and emotion (see for example [2] and [7]). The aim of the text analysis tool LIWC is to sort words in psychologically meaningful categories. By counting the relative frequencies of words in a text it is possible to say how much a person uses words from the different LIWC categories and by using this information we can gain a deeper understanding about the person who actually wrote the text. One of the greatest benefits with LIWC is that it can be seen as a way to gain indirect information about subjects who will not directly provide information about themselves.

LIWC is available in several different versions. The most recent version is called LIWC 2015. The official language of LIWC is English and plenty of work has been put into developing the creating the different categories (and dictionaries) on English. Table I shows a subset of the LIWC categories and some sample words from each category. LIWC 2015 contains approximately 4,000 words and word stems that is categorized into grammatical (e.g., articles, numbers, pronouns), psychological (e.g., cognitive, emotions, social), or content (e.g., achievement, death, home) categories [2], [3].

In this work we have used LIWC and translations of the LIWC dictionaries into four other languages. The translations that we have used are based on LIWC 2007 while the English version of LIWC that we use is LIWC 2015. The French version of LIWC that we used in our experiments is described in [8]. The Russian version and the Spanish version where obtained through the LIWC tool. The Swedish version of LIWC is still a work in progress conducted by [9].

TABLE I: A subset of the LIWC 2015 categories and some example words.

Category	Examples
Linguistic Dimensions	
Total function words	it, to, no, very
Total pronouns	I, our, they, your
1st person singular	I, my, me
1st person plural	we, our, us
3rd person plural	they, their, them
Prepositions	as, at, except, after
Article	a, an, the
Psychological processes	
Affective processes	happy, cried
Positive Emotions	happy, pretty, good
Negative Emotions	hate, worthless, enemy, hurt
Anxiety	worried, fearful
Anger	hate, kill, annoyed
Sadness	crying, grief, sad
Biological processes	eat, blood, pain
Body	cheek, hands, spit
Health	clinic, flu, pill
Sexual	horny, love, incest

IV. Related work

Classification of an author?s gender has been done previously in many different ways and on many different kind of datasets. In [10] gender classification on users' comments on social media is done. By using linguistic features such as the number of verbs, pronouns, articles, adjectives, adverbs, preposition and numbers an accuracy of 66.66% is obtained.

In [11] a set of experiments on 566 documents from the British National Corpus is done. Each document has an average of 34320 words. Initially, a total of 1081 different features are used and then the features were reduced. According to [11] an optimal performance are determined to lie between 64 and 128 features. The authors of [11] notes that male indicators are largely noun specifiers (determiners, numbers, modifiers) while the female indicators are mostly negation, pronouns and certain prepositions. The results show that it is possible to determine the gender of a writer of a text from the British National Corpus with an accuracy of 80%.

One of the more recent work on gender classification is described in [12] where gender classification is done on a set of blogs using two different techniques. Firstly, a pattern based feature where the patterns are frequent sequences of part of speech tags that capture the complex stylistic characteristics of male and female authors. Secondly, a feature selection algorithm that uses an ensemble of feature selection criteria and methods is used. Mukherjee and Bin [12] obtains an impressive accuracy of 88.56% using their method on blog data (3100 blogs).

Deep learning has also been applied to gender classification. In [13] deep learning is applied to a set of blogs. Their approach using deep learning model on a set of blogs reports an accuracy of 86%.

Most gender classification has been done on English data and therefore we know very little about how well gender classification performs on different languages. Most likely, the differences among the genders are transferable to other languages but it is still interesting to get an understanding in differences and similarities between the languages. One attempt to do a completely language independent method for classifying the gender of a twitter user is described in [14]. The gender classification is done based only on the colors of the profile and using five different colors an accuracy of 71.4% can be obtained.

V. Experimental Results

In this section we describe the experimental setup and the results of our experiments. Our experiments are conducted on data consisting of blogs written in five different languages using different LIWC versions as features.

A. Dataset

We have collected a set of blogs from [15] where a list of bloggers can be obtained and sorted by land code, interests, etc. In our case, we sorted by different land codes to obtain different languages. The gender of a blog author is in the majority of the cases also accessible. We have collected a set of blogs that are written by an equal amount of men and women. Blogs texts includes several features that can not be found in more formal texts such as books. Examples are irregular punctuation and grammatical errors.

TABLE II: Average words per document.

Language	Females	Males
English	2030.6	2656.4
Swedish	1473	1806.2
French	1550.4	1666
Spanish	1534.5	2396.9
Russian	737.6	1187.2

TABLE III: Average words per sentence.

Language	Females	Males
English	22.3	20.4
Swedish	16.7	18.9
French	22.4	24
Spanish	19.6	22.8
Russian	13.5	13.3

The blogs touches various topics, such as, politics, sports, personal writings, traveling and so forth. In average 18.9 words per sentences where used by women, and 19.88 words per sentence by men. An example post:

"A little posy of joy! Wishing you all a beautifully happy weekend. xoxoAutumn has now officially arrived in the Highlands. Time to wrap up warm! This time of year always feels like a new beginning to me much more then New Year ever has."

All blog posts from a single blogger were merged into a text file. We denote each such text file a document. Only the bloggers who posted text that contains more than 30 characters are included in our dataset. No upper boundary for a document were defined, the largest document consisted of 27669 words. Table II show the average number of words in a document for all languages and Table III shows the average word per sentence. In average, the English blogs contains more text more than twice as much as the Russian blogs. This is something that could affect the results. The Swedish, French and Spanish blogs has almost the same average of words. In all languages, the average number of words for males is larger than the average for females. Males have a higher average number of words in a sentence in all languages except English and Russian (where the average is almost equal).

B. Experimental setup

All experiments have the same setup. The datasets for the different languages were divided into two sets: 75% of the dataset is used as training data, and 25% of the dataset is used as test data. We repeated a 10-fold cross validation three times on the training data. In our classification we have used Support Vector Machines (SVM). The resulting model is tested on the 25% unseen test data. The accuracy is measured on the

test data. The results for the different languages are described as confusion matrices in which we present the number of true positives, false negatives, true negatives, and false positives as illustrated in Table IV.

	Predicted class	
Actual class	**True Neg. (TN)**	**False Pos. (FP)**
	False Neg. (FN)	**True Pos. (TP)**

TABLE IV: Confusion matrix

To evaluate the results we use the measures accuracy, precision and recall that can be derived from the confusion matrix. Accuracy is defined as:

$$\frac{TP + TN}{TP + FP + TN + FN}$$

precision is defined as:

$$\frac{TP}{TP + FP}$$

and recall as:

$$\frac{TP}{TP + FN}$$

C. Results

The results of our experiments can be seen in V. The table shows how many different documents that are used in each experiment (each document represents a blogger), a confusion matrix, how many features that where used in the classification and the accuracy, precision and recall. The results shows that our gender classification has the highest accuracy on English (79.6%) followed by Swedish (77.1%). On third place we have Russian (76.6%) followed by Spanish (74.2%) and French (73.8%). In general, it is harder to recognize females than males. This is particularly true for French where females are misclassified more than twice as often than males. Spanish and Russian also have a percentage of misclassified females than Swedish and English.

VI. Discussion

As can be noted the accuracy when using language independent features are well over 73% for all different languages using data independent features. The results are slightly better on English - one of the reasons for this could be that a newer version of LIWC (LIWC 2015) is used another reason could be that the average number of words per blog where higher for English which gives the classifier more data to train on.

When using machine learning, the selection of features has a high impact on the result. To get an understanding on how much the different features contribute to the classification we have plotted how often a term is occurring among the different classes for some of the features. The features we have focused on are articles, pronouns, first person singular, function words, verbs and affect. The first five features are well known from previous studies to contribute to separating text written by the two different genders.

Fig. 1 shows the differences between how the two genders use articles. Articles in english are the words *a, an* and *the*. As can be seen there are hardly any differences between the genders in Swedish. One reason for this could be that articles are often included in the end of a word on Swedish (and not before as in English). On English, French and Spanish there is a difference in the use of articles, males use more articles something that reflects a more informative writing style.

The use of pronouns is something that has been shown in previous studies to differentiate genders. A pronoun is used to replace a noun and examples of pronouns are *I, she* and *our*. In Fig. 2 the use of pronouns among the genders on the different languages is shown. As can be seen females use more pronouns on all languages. A higher use of pronouns often reflects a more personal writing. How the two groups use the specific part of pronoun called first person plural (e.g me, myself, mine) is illustrated in Fig. 3. Since first person plural is included in the use of pronouns it is not a surprise that the differences remain.

Function words are words that have very little lexical meaning but are important elements that form the structures of sentences. Examples of function words are *it, here, can* and *no*. In Fig. 4 the differences of the use of functions words between the genders are shown. The difference in the use of function words is less significant on Russian than on the other languages.

Fig. 5 shows how the genders use verbs. The difference among the genders is again, less significant on Russian.

Fig. shows how the genders use affect words. Affect words are words for example *pleasing, worthless, smart* and *fun*. The difference in the use of affect words is significant for all languages except Russian.

Table VII shows a statistical significant test on how the features articles, pronouns, first person singular, function words, verbs and affect are used between the genders. The symbols used in the significant test are shown in Table VIII.

VII. Conclusions and Future Work

In this work we have tested a to use data independent features from the text analys program LIWC to classify the gender of a blog author. Our experiments using five different languages shows that this approach is promising and could the results could be useful in a real world setting.

When using machine learning, the selection of features has a high impact on the result. A direction for future work is to investigate more on what features that could be added to improve the results. In previous work, several different classes of features have been for gender classification some of them are might not be fully captured in our approach. Stylistic features in the form of punctuation, commas, and word length as well as syntactic features such as part-of-speech tag and n-grams are features that successfully have been used in previous work. A feature that would be particularly interesting to add is the F-measure. The F-measure feature explores the notion

TABLE V: The experimental results when classifying gender in the different languages.

Language	No. of documents	Female	Male	No. of Features	Accuracy	Precision	Recall
English	2274	223 63	53 230	73 (LIWC2015)	79.61%	80.79%	77.97%
Swedish	1986	189 63	51 194	63 (LIWC2007)	77.06%	79.18%	75.48%
French	1662	140 74	35 167	64 (LIWC2007)	73.8%	82.67%	65.42%
Spanish	2236	219 81	63 196	74 (LIWC2007)	74.24%	75.67%	73.0%
Russian	1178	110 41	28 116	61 (LIWC2007)	76.61%	80.55%	72.85%

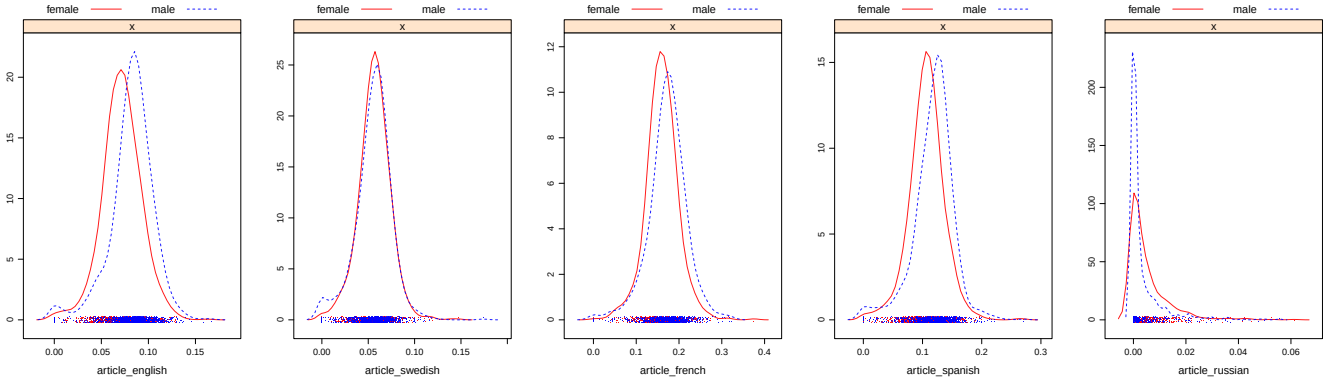

Fig. 1: The differences in the use of articles among female and males in English, Swedish, French, Spanish and Russian.

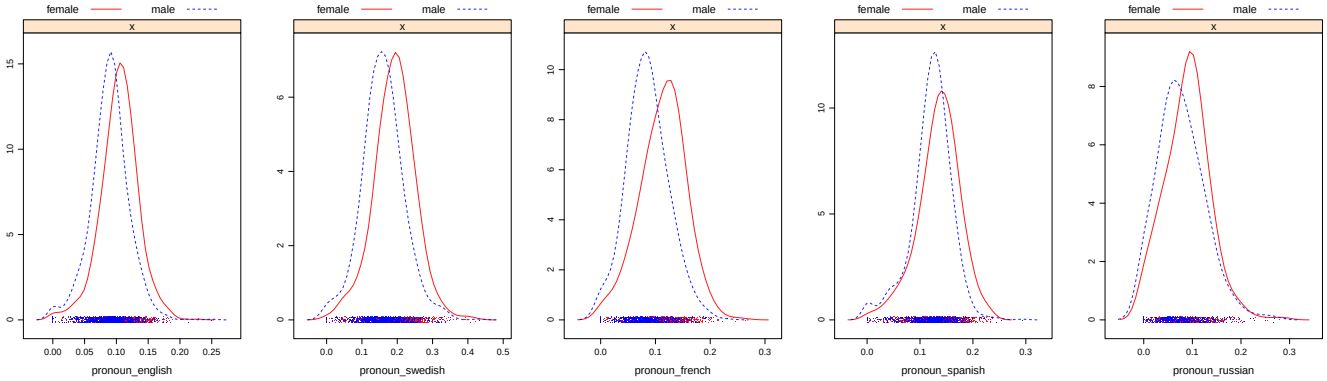

Fig. 2: The differences in the use of pronouns among female and males in English, Swedish, French, Spanish and Russian.

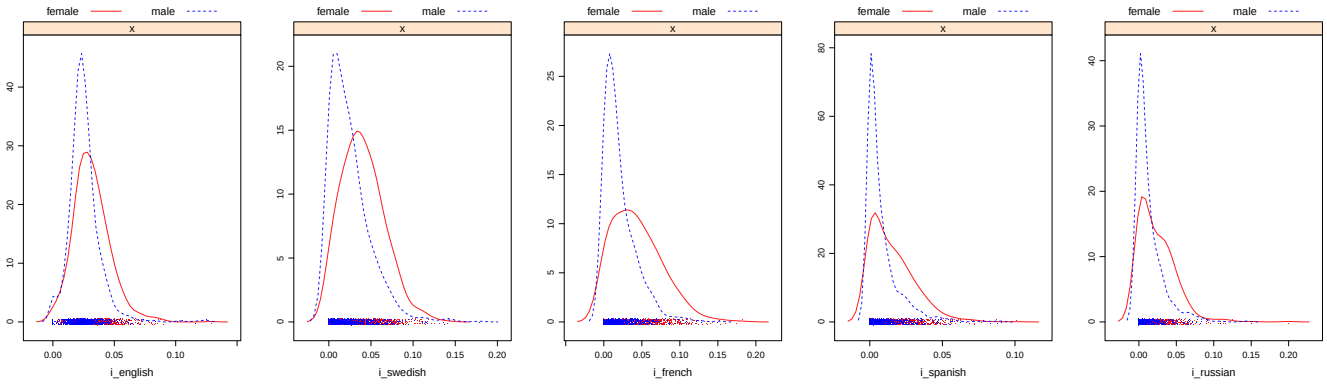

Fig. 3: The differences in the use of first person singular among female and males in English, Swedish, French, Spanish and Russian.

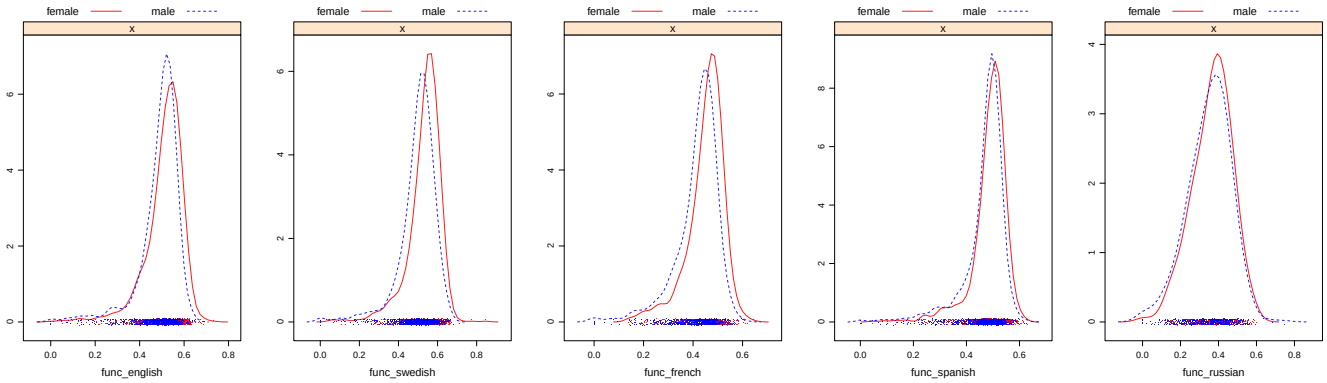

Fig. 4: The differences in the use of function words among female and males in English, Swedish, French, Spanish and Russian.

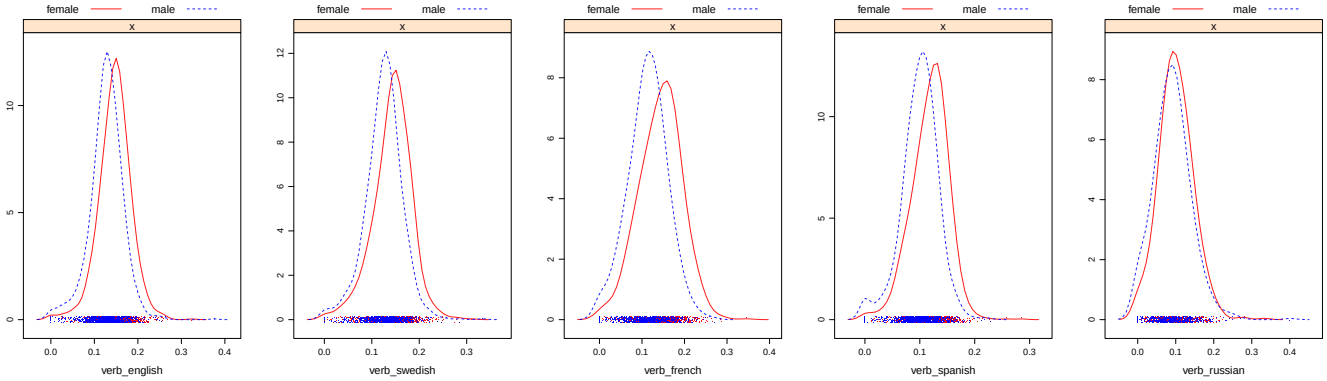

Fig. 5: The differences in the use of verbs among female and males in English, Swedish, French, Spanish and Russian.

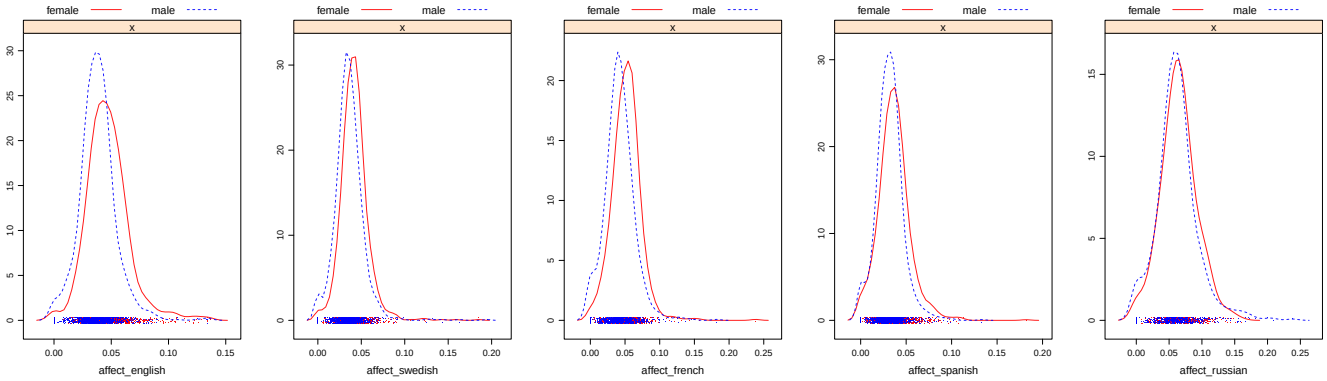

Fig. 6: The differences in the use of affect among female and males in English, Swedish, French, Spanish and Russian.

TABLE VI: The 10 most important features for gender classification in each language.

Order	**English**	**Swedish**	**French**	**Spanish**	**Russian**
1	Positive emotion	Personal pronouns	1st pers plural	1st pers plural	1st pers plural
2	Article	1st pers singular	Personal pronouns	Present focus	See
3	Affect	Pronoun	Positive emotion	Verb	Sexual
4	Verb	Biological processes	Pronoun	Cognitive processes	Personal pronouns
5	Pronoun	Sexual	Present	3rd pers plural	Perceptual processes
6	1st pers singular	Prepositions	Auxiliary verbs	2nd person	Article
7	Biological processes	Verb	Verb	Article	Positive emotion
8	Personal pronouns	Function words	2nd person	Auxiliary verbs	Leisure
9	Power	Positive emotion	Biological processes	Positive emotion	Negative emotions
10	Female references	Adverb	Friends	Prepositions	Anger

TABLE VII: Significance tests for features between genders

Feature	English	Swedish	French	Spanish	Russian
Article	*******	-	******	*******	*****
Pronoun	*******	*******	*******	*******	*****
1st person singular	*******	*******	*******	*******	*******
Function words	*******	*******	*******	*******	*
Verb	*******	*******	*******	*******	***
Affect	*******	*******	*******	*****	-

TABLE VIII: Symbols used in the statistical significance tests

Significance	p-value
-	$p > 0.05$
*	$p \leq 0.05$
**	$p \leq 0.01$
***	$p \leq 0.001$
****	$p \leq 0.0001$
*****	$p \leq 0.00001$
******	$p \leq 0.000001$
*******	$p \leq 0.0000001$

of implicitness of text and is a unitary measure of text's relative contextuality (implicitness), as opposed to its formality (explicitness)[12]. The F-measure is described in [17] and used as a feature in [12].

Another direction for future work would be to work more on the translation of LIWC into the different languages and make sure that all aspects of the language is captured in the translations.

ACKNOWLEDGMENT

This research was financially supported by the Security link project *Deanonymization of webdata*, the EU H2020 project ASGARD: Analysis System for Gathered Raw Data Grant agreement no: 700381 and by the R&D programme of the Swedish Armed Forces.

REFERENCES

[1] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count (liwc): A text analysis program." New York: Erlbaum Publishers, 2001.

[2] Y. Tausczik and J. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, no. 1, March 2010.

[3] J. W. Pennebaker and C. K. Chung, "Computerized text analysis of al-Qaeda transcripts," in *The Content Analysis Reader*, K. Krippendorf and M. A. Bock, Eds. Sage, 2008.

[4] R. Lakoff, in *Language and womans place*. New York: Harper Colophon Books, 1975.

[5] J. W. Pennebaker, *The secret life of pronouns: What our words say about us.* CT New York: Bloomsbury Press., 2011.

[6] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker, "Gender differences in language use: An analysis of 14,000 text samples," *Discourse Processes*, vol. 45, no. 3, pp. 211–236, 2008.

[7] J. Pennebaker, M. Mehl, and K. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology*, vol. 54, no. 1, pp. 547–577, 2003.

[8] A. Piolat, R. J. Booth, C. K. Chung, M. Davids, and J. W. Pennebaker, "La version franaise du liwc : modalits de construction et exemples d'application," *Psychologie franaise*, vol. 56, pp. 145–159, 2011.

[9] P. Taylor, K. Karlsson, M. Gustafsson Sendn, and J. Pennebaker, "Translating liwc- dictionaries into swedish, work in progress."

[10] M. Hosseini and Z. Tammimy, "Recognizing users gender in social media using linguistic features," *Comput. Hum. Behav.*, vol. 56, no. C, pp. 192–197, Mar. 2016.

[11] M. Koppel, S. Argamon, and A. Shimoni, "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, vol. 17, no. 3, 2003.

[12] A. Mukherjee and B. Liu, "Improving gender classification of blog authors," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '10, 2010, pp. 207–217.

[13] A. Bartle and J. Zheng, "Gender classification with deep learning," Technical report, The Stanford NLP Group., Tech. Rep., 2015.

[14] J. S. Alowibdi, U. A. Buy, and P. Yu, "Language independent gender classification on twitter," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13, 2013, pp. 739–743.

[15] "Google blogs." [Online]. Available: https://www.blogger.com/

[16] J. Pennebaker, R. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," in *Austin, TX: University of Texas at Austin*, 2015.

[17] F. Heylighen and J.-M. Dewaele, "Variation in the contextuality of language: An empirical measure," *Foundations of Science*, vol. 7, p. 293340, 2002.