UniNE at PAN-CLEF 2019: Bots and Gender Task
Notebook for PAN at CLEF 2019

Catherine Ikae, Sukanya Nath, Jacques Savoy

Computer Science Department, University of Neuchatel, Switzerland
{Catherine.Ikae, Sukunya.Nath, Jacques.Savoy}@unine.ch

Abstract. When participating in the "bots and gender" subtask (both in English and Spanish), our aim is to automatically detect different text sources (sequence of tweets sent by a bot or a human). When a text is identified as being sent by humans, the system must determine the author's gender (author profiling). To solve these questions, we focus on a simple classifier (k-NN, $k = 5$) usually able to produce a correct answer but not in an efficient way. Thus, we apply a feature selection procedure to reduce the number of terms (around 200 to 500). We also propose to apply a Zeta model to reduce the number of decisions taken by the k-NN classifier. In this case, we focus on terms used in one category and ignored or used rarely by the second. In addition, the Type-Token Ratio of the lexical density (LD) presents some merit to discriminate between tweets sent by a bot (TTR < 0.2, LD ≥ 0.8) or humans (TTR ≥ 0.2, LD < 0.8).

1 Introduction[1]

In the last two decades, UniNE has participated in different CLEF evaluation campaigns with the objective of creating new test collections on the one hand and, on the other, to promote research in different NLP domains. This year, our team takes part in the CLEF-PAN in the subtask "bots and gender profiling" using both the English and Spanish corpus (Rangel & Rosso, 2019).

Within this track, given a set of tweets, the computer must identify whether this sequence was sent by a bot or a human. In the latter case, the author gender must be determined. This author profiling question is not new (Schwartz *et al.*, 2016) and has been the subject of previous evaluation campaigns (Potthast *et al.*, 2019a). This problem presents interesting questions from a linguistics point of view because the web offers new forms of communication (chat, forum, e-mail, social networks, etc.). It was recognized (Crystal, 2006) that such communication channels might be viewed as new forms between the classical oral and written usage. In addition, CLEF-PAN campaigns allow us to access large text corpora to verify stylistic assumptions and to detect new facets in our understanding of gender differences (Pennebaker, 2011).

The rest of this paper is organized as follows. Section 2 describes the text datasets while Section 3 describes our feature selection procedure. Section 4 exposes our combined Zeta and *k*-NN classifier and Section 5 shows some of our evaluation results. A conclusion draws the main findings of our experiments.

2 Corpus

When faced with a new dataset, a first analysis is to extract an overall picture of the data, their relationships, and to detect and explore some simple patterns related to the different categories. A statistical overview of these PAN datasets is provided in Section 2.1 while Section 2.2 focuses on the emoticon distribution across the different categories. The distribution of the Type-Token ratio values is exposed in Section 2.3. Section 2.4 proposes to use the lexical density to discriminate between bots and humans. Finally, Section 2.5 exposes a brief overview of the distribution of positive and negative emotions.

2.1 Overall Statistics

To design and implement our classification system, a training corpus was available in the English and Spanish languages. As depicted in Table 1, the training data contains the same number of documents (one document = a sequence of 100 tweets) in the bots and human categories. In the latter case, one can find exactly the same number of documents written by men and women (1,030 in English, 750 in Spanish).

These values are obtained by concatenating the two subsets made available by the organizers, namely the `train` and `dev` parts. To be precise, the `train` subset is composed of 2,880 English documents and the `dev` by 1,240 items (for a grand total of 4,120). For the Spanish corpus, one can count respectively 2,080 and 920 documents (total: 3,000).

As each document is not a single tweet (but usually 100), the mean number of tokens per document is around 2,097 for the English language (median: 1,920; sd: 961.4; min: 100; max: 5,277). For the Spanish language, the mean length is 1,889 (median: 1,925.5; sd: 619.2; min: 100; max: 4,933). In this computation, the punctuation symbols and emoticons (or sequences of them) count as tokens. For example, from the expression "Paul's books!!!", our tokenizer returns {paul ' s book !!!}. As we can see, a light stemmer was applied, removing only the plural form '-s' (Harman, 1981). This choice is justified to keep the word meaning as close as possible to the original one (which is not the case, for example, with Porter's stemmer reducing "organization" to "organ").

	English		Spanish	
	Bots	Human M / F	Bots	Human M / F
Nb. doc.	2060	1,030 / 1,030	1,500	750 / 750
Nb tweets	205,919	102,842 / 102,930	149,968	75,000 / 75,000
Mean length	2,097	2,014 / 2,123	1,889	1,964 / 1,821
|Voc|	101,826	95,323 / 102,689 human: 162,384	119,965	95,590 / 89,141 human: 147,109

Table 1: Overall statistics about the training data in both languages

Looking at the mean length for both genders, Table 1 does not corroborate the common assumption that “women are more talkative than men”. For the English language, the mean is slightly higher for women (2,123 vs. 2,014) but not for the Spanish corpus (1,821 vs. 1,964).

As text categorization problems are known for having a large and sparse feature set (Sebastiani, 2002), Table 1 indicates the number of distinct terms per category (or the vocabulary size denoted by |Voc|) which is 101,826 for the English bots category. Moreover, and for both languages, the vocabulary size is larger for the human category than for the bots (English: 101,826 vs. 162,384; Spanish: 119,965 vs. 147,109). The texts sent by bots are certainly composed with a smaller vocabulary and the same or similar expressions are often repeated.

🚑 #JOB 🚑 #medical Anesthesiologist https://t.co/t8C84NGQuI 👈 #hiring #health 🏥 https://t.co/HlAmnmpjPZ 🏥.   🚑 #JOB 🚑 #medical Mental Health Nurse https://t.co/i9PEEOOxz2 👈 #hiring #health 🏥 https://t.co/HlAmnmpjPZ 🏥.
11:21 Of the Izharites, the Hebronites, the family of the LORD, that I am a brother to wife.   9:2 And he called for their land to Assyria unto this day have I drawn thee.
🍀🍀🍀🍀🍄🌻🍀🍄🌺   🍀🌺🍀🍀🍄🍀🍄🍀🍄

Table 2a: Examples of two tweets sent by three distinct bots

RT @EdinburghUni: The future of Scotland’s international relations will be discussed at ‘Global Heritage, Global Ambitions: Scotland’s Inte…   Indeed Murray.... https://t.co/fUZ3dqGL1U
Getting ready for Easter ! Growing up in Québec my sweet memories of Easter are from la cabane à… https://t.co/OrIEL6cBSH   Diner tonight ... Nettles a la crème 🌿🌿🌿🌿 https://t.co/eE4vycXV9h

Table 2b: Examples of tweets sent by two women

Happy 1 year with the most amazing girlfriend I could ask for❤️ https://t.co/94QD5vM2KJ   RT @CuntsWatching: "No idea he cut hair" 😂😂😂 https://t.co/qAgs7rRGR3
RT @Jam10Moir: When yeh forget to take the hanger aff yer jersey https://t.co/Gd5SIrS3vA   @UbuntuBhoy It's a hard life.   @DR_Kronenbourg I nearly fucking was 😂

Table 2c: Examples of tweets sent by two men

Of course, the tweets produced by bots are not really generated by computers but correspond to retweets or tweets showing text excerpts extracted from a larger corpus. To illustrate this, Table 2a exposes six examples of tweets generated by three bots, while Table 2b and 2c present four tweets written by two women and men.

2.2 Emoticons

An interesting aspect of web communication (Crystal, 2006) is the frequent usage of emoticons to denote an author's emotions (e.g., 😱, 😊) or to shorten the message (e.g., 🙏, 👌, 💻). Table 3 shows the most frequent emoticons per category and language. From this table, it is not fully clear how we can simply detect a pertinent pattern to be suitable for automatic classification. One can infer that humans employ more frequently such symbols compared to bots. On the other hand, women show a higher usage of emoticons but without showing an important difference about the emoticon types. When analyzing the sequence of emoticons, the most frequent one is "😂😂😂" follows by "😂😂".

	English			Spanish		
Rank	Bots	Male	Female	Bots	Male	Female
1	👉 55	😂 406	😂 457	😂 102	😂 236	😍 328
2	👈 51	👍 193	😍 316	🏻 62	😍 165	😂 310
3	😂 39	🏻 172	😭 270	😍 62	👏 153	🏻 234
4	👀 36	👌 150	🏻 259	👉 60	🏻 141	😭 210
5	🔥 33	😍 150	👏 248	🇪 55	🤔 129	👏 179

Table 3: The most frequent emoticons per category and language

2.3 Type-Token Ratio

As bots could be deployed to send a repetitive message (maybe with a slight modification), one can assume that the TTR value (the number of distinct word-types divided by the number of word-tokens) should be smaller than for a sequence of tweets written by a human. Of course, the text genre has a clear impact on this estimation, with a lower TTR value for an oral production compared to a written message. As a comparison basis, the TTR achieved by Trump was 0.297 vs. 0.362 for Hillary Clinton (oral form, primaries debates) (Savoy, 2018). Over all candidates, Trump achieved the lowest value, depicting a candidate owning a reduced vocabulary and repeating the same expressions again and again. These examples indicate that values smaller than 0.25 or 0.2 represent a clear lower limit for a message.

Based on the training set (English language), the TTR values have been computed for documents sent by bots and humans. The two resulting distributions are depicted in Figure 1. In this case, one can see that messages sent by bots tend to contain the same or similar expressions resulting in a lower TTR value, even lower than 0.2 (usually producing a boring message). A similar picture can be obtained with the Spanish language (see the Appendix).

With the English training data, one can count 398 documents generated by a bot having a TTR value smaller than 0.2 (over 2,060 or 19.3%). On the other hand, only 13 documents having a TTR smaller than 0.2 have been written by humans. For the Spanish corpus, one can find 843 documents generated by bots with a TTR values smaller than 0.2 (over 1500, or 56.2%), and none by human beings.

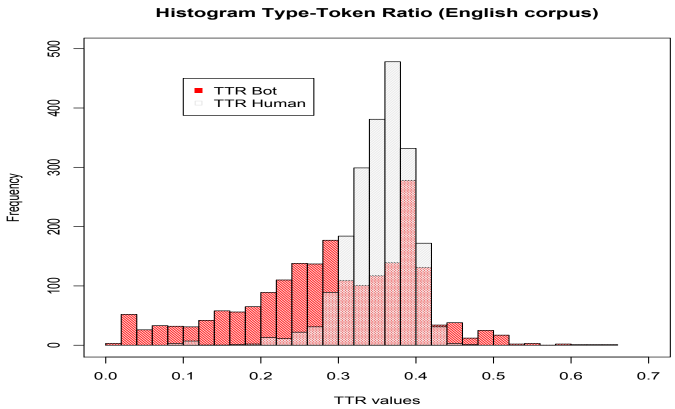

Figure 1: Distribution of the TTR values for the bots vs. human (English corpus)

2.4 Lexical Density

The lexical density measures the percentage of content words in a text. This percentage can also be estimated by considering the number of functional words in a text and assuming that a word could be either a content word or a functional one (see Eq. 1). In our implementation, the English language has 571 functional words while for Spanish such a wordlist counts 350 entries.

$$\mathrm{LD(T)} = {content(T)}\big/{len(T)} = 1 = - {function(T)}\big/{len(T)} \qquad (1)$$

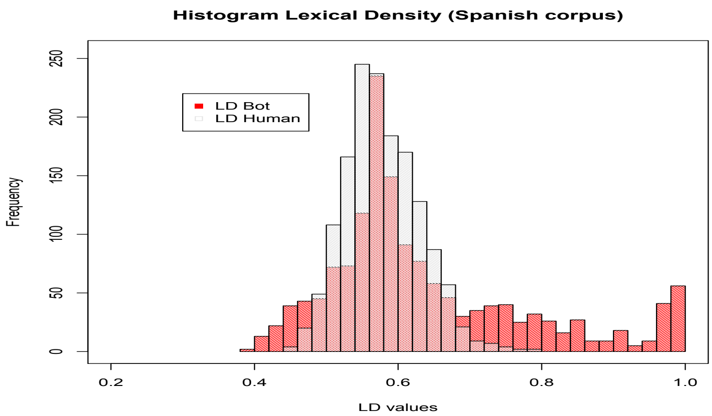

Figure 2: Distribution of the lexical density values for the bots vs. human (Spanish corpus)

As shown in Figure 2, bots tend to present a higher LD value than the set of tweets sent by humans. For example, by assuming that the maximum value for a document written by a human is 0.8, the system can consider documents having a larger value as sent by bots. On the training set, one can count 322 English documents or 216 Spanish ones (sent by bots) for one single English document written by a human (and none in the Spanish corpus).

2.5 Emotion Distribution

With the English language, we have a list of words corresponding to positive (159 entries) and negative (151 entries) emotions (extracted from the LIWC (Linguistic Inquiry and Word Count) (Tausczik & Pennebaker, 2010)). According to Pennebaker's findings (2011), one can expect a larger number of emotional words in tweets written by woman. According to the data depicted in Table 4, such a difference does exist but it is rather small. Moreover, when analyzing only the emotions expressed with words, the mean is rather low (2.5%) but even smaller for bots (1.86%). One can also consider that emotions are also provided by emoticons and thus we need to take account of both the emoticons and words indicating emotions.

	Positive		Negative	
	Bots	Male / Female	Bots	Male / Female
Mean	0.0186	0.0239 / 0.0252	0.0043	0.0057 / 0.0056
Median	0.0152	0.0233 / 0.0239	0.0032	0.0052 / 0.005
Stdev	0.0164	0.009 / 0.0106	0.006	0.0032 / 0.0034

Table 4: Distribution of the positive and negative emotions (English language)

3 The Feature Selection

According to our point of view, the key function of a successful classifier is to be able to generate a good feature set. Moreover, we also want to understand the proposed attribution and be able to explain it in plain English. Therefore, one of our main objectives is to reduce the feature space into one to three orders of magnitude compared to a solution based on all possible isolated words. As shown in Table 5, the vocabulary size (|Voc|) is large for all categories and languages. If text categorization can be characterized by such huge feature spaces, they are also sparse (when considering isolated words, *n*-grams of words or letters). Many terms occur just once (*hapax legomenon*) or twice (*dis legomena*). Ignoring those words reduces the vocabulary size by around 50%.

To reduce this feature set and based on the training data, terms (isolated words or punctuation symbols in this study) having a tweet frequency (*df*) smaller than a predefined threshold (fixed at 9 in our experiments) are ignored (Savoy, 2015). With the English bots corpus (see the first two rows in Table 5), this filter reduced the feature space from 101,826 to 15,478 dimensions (a reduction of 84.8%). Higher reduction rates can be achieved with the human vocabulary (English: 162,384 to 14,728 (90.9%); Spanish: 147,109 to 13,866 (90.6%)).

Using the term frequency difference, we can observe the term more employed in each category. For example, the terms "urllink" (replacing the sequence "http://aref"), "job", "developer", "and", "hiring" or "swissmade" appear more frequently in tweets sent by bots than by humans. As other examples, we can mention that men used more frequently: “the”, “that” “it”, “he”, “a”, “is” and the punctuation symbols “.”, “,”. Woman tweets contain more “rt” (retweets), “you”, “to”, “my”, “your”, “thank”, “me”, “love” and the punctuation symbols “:”, “&”. These short examples tend to confirm part of Pennebaker’s (2011) findings, indicating that definite articles are more frequently used by men while personal pronouns and emotions tend to appear more often in female messages.

	English		Spanish	
	Bots	Humans M / F	Bots	Humans M / F
|Voc|	101,826	95,323 / 102,689 human: 162,384	119,965	95,590 / 89,141 human: 147,109
with *df* > 9	15,478	9,122 / 9,227 human: 14,728	16,725	11,488 / 9,848 human: 13,866
with *tf*	8,699	5,768 / 5,417 human: 10,173	*8,732*	5,415 / 4,552 human: 9,283
Voc Uniq	345	373	391	364

Table 5: Vocabulary size with different feature selection strategies

To go further in this space reduction, one can then add a final third step by applying a feature selection procedure. For example, one can reduce the feature space to a value between 200 to 500, allowing a manageable space to explain the proposed decision. For example, previous studies indicate that odds ratio, mutual information or occurrence frequency tends to produce effective reduced term sets for different text categorization tasks (Sebastiani, 2002), (Savoy, 2015).

In addition, our classifier will also consider terms used infrequently in one category and ignored or used rarely by the other (Zeta model) (Burrows, 2007), (Craig & Kinney, 2009). To achieve this, terms appearing only in a single category are extracted and ranked according to their term frequency (*tf*) or document frequency (*df*). Instead of considering all of them, only the top 200 most frequent ones (based on the *tf* and *df* statistics) are judged useful to discriminate between two classes. The two wordlists (each containing 200 entries) are merged to generate the final terms able to discriminate between the two categories. The size of those lists is depicted in Table 5 under the label “Voc Uniq” (e.g., English bots: 345, English human: 373). For example, within the bots category, one can find terms such as: “camber”, “cincinnati”, “cooperative” or “norwalk”. The male category is characterized by terms such as: “outwildtv”, “obstruction”, “avalanche” or “golfer” while in tweets written by women, one can find “gown”, “allergy”, “👭” or “ballet”.

It is also interesting to analyze the distribution of definite articles and some pronouns (Pennebaker, 2011) in both languages as depicted in Table 6a (English) and 6b (Spanish). In those tables, the number of documents in each gender is the same and represents the half of those appearing in the column “Bots”. Thus, looking at the frequencies, one can expect a pattern such as 2:1:1 when the term occurrence frequency is the same through the different categories.

The frequencies depicted in Table 6a confirm Pennebaker’s findings. Definite articles (“the”, “a”) are more frequent with male writers, and personal pronouns (“i”, “you”, “me”, etc.) are more often used by women. Exceptions can be found. The

English pronoun “he” is clearly employed more often by men. Bots frequently adopt the pronouns “you” or “we” and use more infrequently “she”. Is the bot style more feminine?

	Bots *tf/df*	Male *tf/df*	Female *tf/df*
the	95,860 / 1,881	54,272 / 1,030	48,311 / 1,030
a	67,654 / 1,895	32,316 / 1,030	31,773 / 1,030
i	24,367 / 1,568	24,160 / 1,022	29,598 / 1,014
you	39,521 / 1,654	16,916 / 1,030	20,811 / 1,030
she	1,403 / 628	1,347 / 557	2,114 / 685
he	4,633 / 1,092	5,269 / 897	3,165 / 765
we	13,331 / 1,463	6,661 / 995	7,483 / 985
swissmade	439 / 6	0 / 0	0 / 0
swiss	62 / 47	15 / 15	15 / 14
spain	68 / 46	89 / 64	44 / 39
italy	79 / 64	92 / 57	113 / 59
portugal	25 / 19	37 / 29	22 / 19
germany	147 / 96	105 / 78	66 / 56
france	213 / 147	158 / 112	145 / 103

Table 6a: Some occurrence statistics for the English corpus (*tf* / *df*)

	Bots *tf/df*	Male *tf/df*	Female *tf/df*
el	56,700 / 1310	26,180 / 750	19,505 / 750
un	21,627 / 1256	19,233 / 749	8,930 / 749
una	12,070 / 1191	6,293 / 742	5,817 / 745
unos	1,155 / 367	355 / 263	276 / 211
unas	710 / 254	156 / 129	163 / 146
yo	2,174 / 552	1,892 / 662	3,073 / 662
tu	4,178 / 686	1,687 / 577	2,569 / 648
ella[s]	562 / 319	381 / 253	534 / 323
ello[s]	674 / 362	442 / 281	429 / 265
nosotro[s]	511 / 276	260 / 190	282 / 198
vosotro[s]	118 / 74	43 / 32	32 / 26

Table 6b: Some occurrence statistics for the Spanish corpus (*tf* / *df*)

The Spanish corpus also confirms Pennebaker’s conclusions. Definite articles (“el”, “un”, “una”, etc.) appear more frequently with men, and personal pronouns (“yo”, “tu”, “ella”, etc.) are more associated with the woman’s style. The Spanish pronoun “nosotros” (we) or “vosotros” (you, plural) are usually not present but this indication is often implicit with the verbal suffixes (e.g., “podemos” we can). (A linguist will also infer that frequencies of such pronouns will be rather small due to their spelling composed of 8 letters, not a length reflecting the less effort principle).

Finally, when analyzing the popularity of some countries (see Table 6a), one can see that “france” is the most popular while “swissmade” appears only with bots. For the other names, “italy” is more associated with women, while all the others are with men (except “swiss” that is associated with bots) (due to soccer, a sport popular in Spain, Italy and Germany?).

4 Proposed Text Classification Strategy

Our solution is based on a three-stage function. In the first, the needed variables are initialized (function `preProcessing()` in Figure 3) and they correspond to the unique vocabulary used in the two categories (`VocUnC1, VocUnC2`) and to the document representations belonging to the two categories (`PtC1, PtC2`).

Based on the training data, the system extracts the vocabulary (isolated terms with their frequency) appearing in both categories (function `defineVoc`()). From them, one can determine the terms appearing frequently in one category but absent (or occurring rarely) in the second (in our implementation, such a term can appear up to three times (`min=3`) in the second category). To rank them, the term frequency (*tf*) or the tweet frequency (*df*) statistics are applied. Instead of returning two wordlists, the system selects the top 200 most frequent ones in the underlying category and merges them (function `topVoc()`). In Steps #5 and #6, the system represents the documents belonging to Category #1 or #2 as vectors (generating the `PtC1` and `PtC2` variables).

After this initialization, each document belonging to the test sample can be processed (see function `binaryClassifier()` in Figure 3). In Step #1, the Zeta model is applied. This function counts the number of distinct terms appearing in `VocUnC1` (denoted `N1`) and in `VocUnC2` (or `N2`). If (`N1` > `N2`+θ), the test identifies the given document as belonging to Category #1. On the other hand, if (`N2` > `N1`+θ), it is assumed that the document must be labeled with the second category (e.g. Human). If the Zeta reaches a decision (e.g., `dec=1` for Bot, `dec=2`, for Human), this value is returned.

```
preProcessing(trainDoc)
1   vocC1 = defineVoc(trainDoc)
2   vocC2 = defineVoc(trainDoc)
3   VocUnC1 = topVoc(vocC1, vocC2, top=200, min=3)
4   VocUnC2 = topVoc(vocC2, vocC1, top=200, min=3)
5   PtC1 = definePoints(trainDoc, C1)
6   PtC2 = definePoints(trainDoc, C2)
    return(VocUnC1, VocUnC2, PtC1, PtC2)

binaryClassifier(newD, VocUnC1,VocUnC2,PtC1,PtC2):
    decision = 0
1   dec = Zeta(newD, VocUnC1, VocUnC2, θ=3)
2   if (dec == 1) or (dec == 2):  return(dec)
3   aTTR = TTR(newD)
4   if (aTTR < 0.2): return(dec=1)
5   dec = k-NN(newD, PtC1, PtC2, k=13)
    return(dec)
```

Figure 3: The main steps of our automatic attribution system

Otherwise, Zeta is unable to achieve a clear decision (`dec=0`). For those cases, the TTR value (Type-Token Ratio) is computed (Step #3 and4). When this value is smaller than 0.2, the decision is taken as "Bot" (`dec=1`). In addition, we might have computed the lexical density value and returned "Bot" if this value is larger than 0.8. This step was not included in our final submission (due to time constraints).

In general, the Zeta model (together with the TTR value) cannot always propose a clear answer. In this case, the system calls the *k*-NN function (with the new document, and the set of points corresponding to Category #1 (`PtC1`) or #2 (`PtC2`)). In our experiment, the *k* value was fixed to 13 and the distance between two text surrogates is computed according to the Manhattan function (Kocher & Savoy, 2017).

When the document type is found to be sent as human, the system re-applies the `binaryClassifier()` function but with Category #1 corresponding to male and Category #2 to female (but ignoring the TTR computation).

5 Evaluation

Table 7 depicts the accuracy rate achieved with our model under different conditions and for both the type (bot vs. human) and the gender (male vs. female). These results were achieved with the English corpus using the `dev` test set. In the first row, all words have been used to build the document surrogates. In the second line, the vocabulary size was reduced to consider only terms having a *df* value larger than 9. In the next row labelled "FS", our feature selection is applied. Finally, the last five lines correspond to a feature space reduced to 100, 200, 300, 400, or 500 terms selected by the information gain function (Sebastiani, 2002). When applying our nearest neighbor approach, Table 7 indicates the mean accuracy rates achieved considering *k*=13 or *k*=5 neighbors.

To compute the accuracy rates, only the `train` subset is used to define the needed wordlists and document surrogates (in other words, based on 2,880 English documents, and 1,240 Spanish ones). During the evaluation, only the `dev` subset was needed to derive the performance values (or with 2,080 English documents, and 980 Spanish ones).

	English (`dev` set)			
	k = 13		*k* = 5	
	type	gender	type	gender
All voc	0.8807	0.7161	0.8863	0.7161
with *df* > 9	0.**9032**	0.7436	0.8976	0.7339
with *tf*	0.9024	0.**7557**	0.8984	0.7420
100 IG	0.8927	0.7105	0.8960	0.7129
200 IG	0.8807	0.7081	0.8911	0.7145
300 IG	0.8831	0.7048	0.8815	0.6992
400 IG	0.8895	0.7177	0.8871	0.6992
500 IG	0.8960	0.7282	0.8911	0.7056

Table 7: Evaluation of under different feature selection strategies (English corpus)

The important conclusion that can be drawn from Table 7 is that it is possible to reduce the feature set to a few hundred words and to still have a good overall effectiveness. Considering *k*=13 neighbors tends to produce better results (and this solution is less prone to over-fitting).

Table 8 reports our official results achieved with the TIRA system (Potthast *et al.*, 2019b) using the first (test set 1) or the second (test set 2). These evaluations correspond to our feature selection (FS) with the inclusion of the Zeta test and TTR filter. More information can be found in (Rangel & Rosso, 2019).

	TIRA test set 1 *k*=5		TIRA test set 1 *k*=13		TIRA test set 2 *k*=13	
Classifier	type	gender	type	gender	type	gender
FS+Zeta test+TTR	0.8939	0.7689	0.8939	0.7992	**0.9125**	**0.7371**

Table 8: Official Evaluation of under different feature selection strategies (English corpus)

6 Conclusion

Using the CLEF-PAN datasets of the “bots and gender profiling” written in English and Spanish, we were able to achieve the following main findings. First, the text genre associated with bots can be viewed as repetitive, showing a low TTR value (usually lower than 0.25). After fixing a threshold (e.g., 0.2) for this value, one can detect 9.6% to 55% tweet sequences sent by bots (see Figure 1) with a low error rate (around 3%). For a large majority however (90% for the English corpus), documents present a higher TTR value and no decision can be reached with this simple rule. Similarly, one can compute the lexical density value and one can see that values larger than 0.8 correspond very often to bot tweets.

Second, analyzing the emoticon distribution, or the most frequent ones, we can infer that humans tend to employ them more frequently than bots. In tweets sent by machines, the used emoticons indicate directions or appear to draw reader attention (see Table 3). If humans have adopted the emoticons in their web communications, it is not clear whether we can easily distinguish their usage between men and women.

Third, our attribution approach is based on a cascade classifier. In a first step, the Zeta classifier is used to determine the category (bots vs. human, male vs. female) based on terms occurring infrequently in the first class and never (or very rarely) in the second. When the test sample is strongly correlated to the training set, such a strategy works well and can accurately determine close to 85% when a decision can be computed. As the main drawback, this approach fails to propose an answer when the vocabulary appearing in the new document is not associated clearly with one of the predefined wordlists. In such cases, a second classifier must be used (k-NN in our experiments, with $k = 5$, Manhattan distance).

Fourth, removing terms occurring rarely or in a few documents corresponds to our first step in the proposed reduction procedure. In addition, we impose that terms appearing more frequently in a given category must be selected for that class. This strategy can be further improved by applying a term filter (e.g., mutual information, odds ratio (Sebastiani, 2002), (Savoy, 2015)). After this step, the number of terms could be limited from 200 to 500. This last step is usually accompanied with an effectiveness decrease (around 3% to 8%, depending on the collection).

References

Burrows, J.F. All the way through: Testing for authorship in different frequency strata. *Literary and Linguistic Computing*, 22(1), 27-47 (2007).

Craig, H., and Kinney, A.F. *Shakespeare, computers, and the mystery of authorship*. Cambridge University Press, Cambridge (2009).

Crystal, D. *Language and the internet*. Cambridge University Press, Cambridge (2006).

Harman, D. How effective is suffixing? *Journal of the American Society for Information Science*, 42(1), 7–15 (1991).

Kocher, M., Savoy, J. Distance measures in author profiling. *Information Processing & Management*, 53(5), 1103–1119 (2017).

Pennebaker, J.W. *The secret life of pronouns*. Bloomsbury Press, New York (2011).

Potthast, M., Rosso, P., Stamatatos, E., Stein, B. A decade of shared tasks in digital text forensics at PAN. *Proceedings ECIR 2019*, Springer LNCS # 11437, 291–303 (2019a).

Potthast, M., Gollub, T., Wiegmann, M., Stein, B. TIRA integrated research architecture. In N. Ferro, C. Peters (eds), Information retrieval evaluation in a changing world – Lessons Learned from 20 years of CLEF. Springer, Berlin (2019b).

Rangel, F., & Rosso, P. *Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling*. In: Cappellato L., Ferro N., Müller H, Losada D. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, (2019).

Savoy, J. Comparative evaluation of term selection functions for authorship attribution. *Digital Scholarship in the Humanities*, 30(2), 246–261 (2015).

Savoy, J. Analysis of the style and the rhetoric of the 2016 US presidential primaries. *Digital Scholarship in the Humanities*, 33(1), 143–159 (2018).

Sebastiani, F. Machine learning in automatic text categorization. *ACM Computing Survey*, 34(1), 1–27 (2002).

Schwartz, H.A, Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seliman, M.E.P., Ungar, L.H. Personality, gender, and age in the language of social media. *PLOS One*, 8(9) (2013).

Tausczik, Y.R., & Pennebaker, J.W. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1): 24-54 (2010).

Appendix

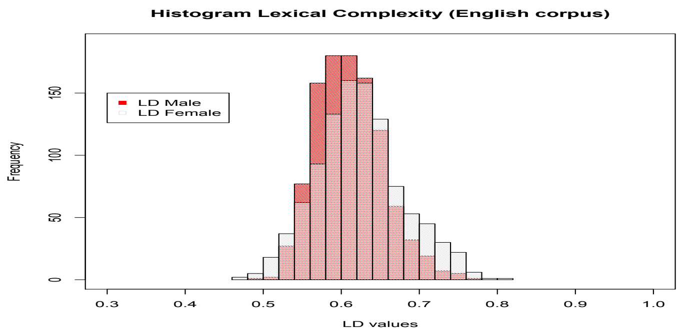

Figure A.1: Distribution of the lexical density values for the male vs. female (English corpus)

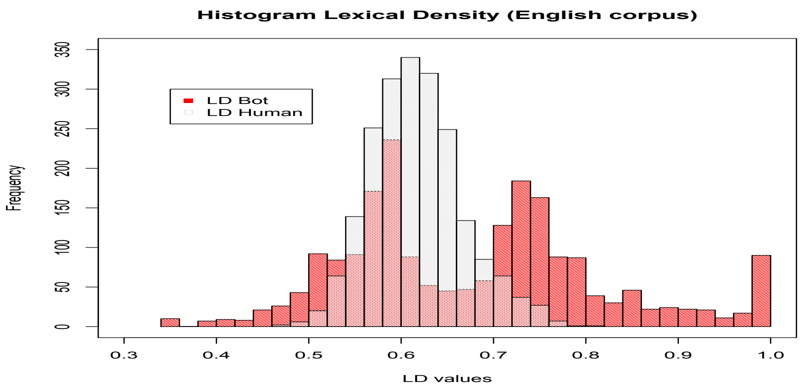

Figure A.2: Distribution of the lexical density values for the bots vs. human (English corpus)

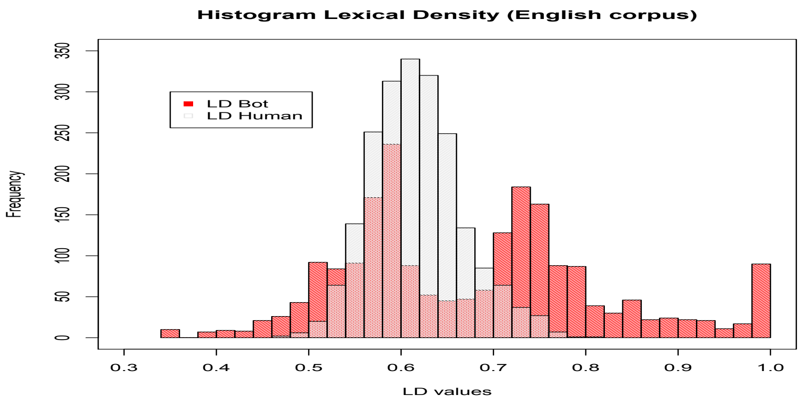

Figure A.3: Distribution of the lexical density values for the bots vs. human (English corpus)