Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling in Twitter

Francisco Rangel[1,2] Paolo Rosso[2]

[1]Autoritas Consulting, S.A., Spain
[2]PRHLT Research Center, Universitat Politècnica de València, Spain

pan@webis.de http://pan.webis.de

Abstract This overview presents the Author Profiling shared task at PAN 2019. The focus of this year's task is to determine whether the author of a Twitter feed is a bot or a human. Furthermore, in case of human, to profile the gender of the author. Two have been the main aims: *i)* to show the feasibility of automatically identifying bots in Twitter; and *ii)* to show the difficulty of identifying them when they do not limit themselves to just retweet domain-specific news. For this purpose a corpus with Twitter data has been provided, covering the languages English, and Spanish. Altogether, the approaches of 56 participants are evaluated.

CEUR-WS.org/Vol-2380/paper_263.pdf

1 Introduction

Society is increasingly polarised, at least this is what we can infer from the last World Economic Forum's 2017 Global Risk Report[1]. People are organised into separated communities with similar opinions, and with the same stance towards controversial topics. The communication among these communities is non-existent, or it is based on hate speech. Highly partisan entities try to massively influence public opinion[2] [71] through social media, since these new communication media can amplify what occurs in society. Shielded behind anonymity and combined with botnets, these partisan entities can achieve a significantly negative impact[3] [37, 87].

Bots are automated programs which pose as humans with the aim at influencing users with commercial, political or ideological purposes. Malicious bots are strongly related to polarisation due to their aim to spread disinformation and hate speech, trying

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

[1] http://reports.weforum.org/global-risks-2017/part-1-global-risks-2017
[2] https://pan.webis.de/semeval19/semeval19-web/index.html
[3] https://www.bloomberg.com/news/articles/2018-05-21/twitter-bots-helped-trump-and-brexit-win-economic-study-says
https://www.cnbc.com/2019/02/04/twitter-bots-were-more-active-than-previously-known-during-2018-midterms-study.html

for instance to enhance some political opinions or supporting some political candidates during elections [8]. For example, the authors of [88] showed that 23.5% of 3.6 million tweets about the 1 Oct 2017 referendum for the Catalan independence were generated by bots. These bots sent emotional and aggressive messages to pro-independence influencers [88]. About 19% of the interactions were from bots to humans, in form of RTs and mentions, as a way to support them (*echo chamber*), whereas only 3% of humans interacted with bots. Similarly, accordingly to Marc Jones and Alexei Abrahams [44], a plague of Twitter bots is roiling the Middle East[4]. The authors showed that 17% of a random sample of tweets mentioning Qatar in Arabic were produced by bots in May 2017 and they raised up to 29% one year later. The authors highlighted the prevalence of automated Twitter accounts deploying hate speech, especially in relation to sectarianism and the Gulf Cooperation Council (GCC) crisis. In 2016, a large bot network producing tens of thousands of anti Shia tweets were detected on regional hashtags in the Gulf[5]. During the outbreak of the Gulf Crisis in 2017, thousands of bots were found to be promoting highly polarising anti-Qatar hate speech. Regarding the U.S. presidential election, Bessi and Ferrara [8] showed that, in the week before election day, around 19 million bots tweeted to support Trump or Clinton[6]. In Russia fake accounts and social bots have been created to spread disinformation[7] [65], and around 1,000 Russian trolls would have been paid to spread fake news about Hillary Clinton[8].

Bots could artificially inflate the popularity of a product by promoting it and/or writing positive ratings, as well as undermine the reputation of competitive products through negative valuations. The threat is even greater when the purpose is political or ideological (see Brexit referendum or US Presidential elections[9] [39]). Fearing the effect of this influence, the German political parties rejected the use of bots in their electoral campaign for the general elections[10]. In addition, the use of bots is increasing. As shown by a recent analysis of the Pew Research Center[11], an estimated two-thirds of tweeted links to popular websites are posted by automated accounts – not human beings. Therefore, to approach the identification of bots from an author profiling perspective is of high importance from the point of view of marketing, forensics and security.

Author profiling aims at classifying authors depending on how language is shared by people. This may allow to identify demographics such as age and gender. After having addressed several aspects of author profiling in social media from 2013 to 2018, the Author Profiling shared task of 2019 aims at investigating whether the author of a Twitter feed is a bot or a human. Furthermore, in case of human, to profile the gender of the author. Two have been the main aims of this year's task: *i)* to show the feasibility

[4] https://www.washingtonpost.com/news/monkey-cage/wp/2018/06/05/fighting-the-weaponization-of-social-media-in-the-middle-east

[5] https://exposingtheinvisible.org/resources/obtainingevidence/automated-sectarianism

[6] http://comprop.oii.ox.ac.uk/2016/11/18/resource-for- understanding-political-bots/

[7] http://time.com/4783932/inside-russia-social-media-war-america/

[8] http://www.huffingtonpost.com/entry/russian-trolls-fake-news_us_58dde6bae4b08194e3b8d5c4

[9] https://www.theguardian.com/world/2018/jan/10/russian-influence-brexit-vote-detailed-us-senate-report

[10] https://www.voanews.com/europe/merkel-fears-social-bots-may-manipulate-german-election

[11] https://www.pewinternet.org/2018/04/09/bots-in-the-twittersphere/

of automatically identifying bots in Twitter; and *ii)* to show the difficulty of identifying them when they do not limit themselves to just retweet domain-specific news.

The remainder of this paper is organized as follows. Section 2 covers the state of the art, Section 3 describes the corpus and the evaluation measures, and Section 4 presents the approaches submitted by the participants. Sections 5 and 6 discuss results and draw conclusions respectively.

2 Related Work

Pioneer researchers [51, 52] proposed the use of *honeypots* to identify the main characteristics of online spammers. To this end, they deployed social honeypots in MySpace and Twitter as fake websites that act as traps to spammers. They found that the collected spam data contained signals strongly correlated with observable profile features such as contents, friend information or posting patterns. They used these observable features to feed a machine learning classifier with, in order to identified spammers with high precision and low rate of false positives.

Recently, the authors of [28] proposed a framework for collecting, preprocessing, annotating and analysing bots in Twitter. Then, in [29] they extracted several features such as the number of likes, retweets, user replies and mentions, URLs, or follower-friend ratio, among others. They found out that humans create more novel contents that bots, which rely more on retweets or URLs sharing. The authors of [20] approached the bots identification problem from an emotional perspective. They wondered whether humans were more opinionated than bots, showing that sentiment related factors help in identifying bots. They reported an AUC of 0.73 on a dataset regarding the 2014 Indian elections.

Botometer[12] [92] is an online tool for bots detection which extracts about 1,200 features for a given Twitter account in order to characterise the account's profile, friends, social network structure, temporal activity patterns, language, and sentiment. According to the authors of [96], the aim of Botometer was at arming the public with artificial intelligence to counter social bots. Thus, there are several analyses carried out with the help of Botometer. For example, the authors of [12] analysed the effect of Twitter bots and Russian trolls in the amplification around the vaccine debate. Similarly, the authors of [83] analysed with Botometer the spread of low-credibility contents.

Although most of the approaches are based on feature engineering and traditional machine learning classifiers, some authors are moving to deep learning. For example, the authors of [49] proposed a deep neural network based on contextual long short-term memory (LSTM) architecture which is fed with both content and metadata. At tweet-level, the authors reported a high classification accuracy (AUC > 96%), whereas the reported accuracy at user-level is nearly perfect (AUC > 99%). Similarly, the authors of [14] used an LSTM to analyse temporal text data collected from Twitter and reported an F1 score of 0.8732 on the honeypot dataset created by the authors of [61]. The authors built the dataset under the hypothesis that a user who connects to the honeypot could be considered a bot.

[12] https://botometer.iuni.iu.edu

The investigation is less prolific in languages different than English. In Arabic for instance, the authors in [1] collected a corpus from Twitter that was annotated within a crowd sourcing platform. They approached the problem by combining formality, structural, tweet-specific and temporal features. They showed that tweet-specific features helped to improve the accuracy to 92%. Similarly, the authors in [10] built their corpus in two ways. Firstly, they paid for services that bring automatic retweets, favourites, and votes, and labeled the Twitter accounts as bots. Secondly, they manually labeled as bots Twitter accounts that repeated the same content rapidly, post non useful text, or post unrelated contents. The authors combined several features from the tweets (source of the tweet, number of favourites, number of retweets, tweet length, number of hashtags, etc.), and the Twitter account (number of tweets/retweets per hour/day/total, time between two consecutive tweets, number of followers, biography length, etc.). The authors reported an accuracy of 98.68%.

Echoing the importance of automatically detecting bots, DARPA held a 4-week competition [89] with the aim at identifying *influence* bots supporting a pro-vaccination discussion on Twitter. Influence bots can be defined as those bots whose purpose is to shape opinion on a topic, posing in danger the freedom of expression. The organisers provided with a total of 7,038 user accounts, with the corresponding user profile, and a total of 4,095,083 tweets. They also provided with network data with snapshots consisting of tuples (from_user, to_user, timestamp, weight). Six teams participated in the challenge by using features from the tweets contents together with temporal, user profile and network features.

Although there are several approaches to bots identification in social media, almost all of them rely on several characteristics beyond text. As said by the authors of [72], a content-based bot detection model could also be seen as a step towards a multi-platform solution, as it would be less dependent on Twitter-specific social features.

Regarding gender identification, pioneer researchers such as Pennebaker [67] found that in English women use more negations and first persons, because they are more self-conscientious, whereas men use more prepositions in order to describe their environment. On the basis of their psycho-linguistic studies, the authors developed the Linguistic Inquiry and Word Count (LIWC) resource [66].

Notwithstanding initial investigations in author profiling focused mainly on formal texts and blogs [3, 38, 13, 46, 82], recent investigations moved to social media such as Twitter, where language is more spontaneous and less formal. In this line, it is worth mentioning the contribution of different researchers that used the PAN corpora since 2013. The authors of [56] proposed a MapReduce architecture to approach, with 3 million features, the gender identification task on the PAN-AP-2013 corpus, which contains hundreds of thousands of users, whereas the authors of [95] showed the importance of information retrieval-based features for the task of gender identification on the same corpus. The authors of [76, 75] showed the contribution of the emotions to discriminate between genders with their EmoGraph graph-based approach on the PAN-AP-2013 corpus as well as the robustness of the approach against genres and languages on the PAN-AP-2014 corpus. The authors of [7] showed that word embeddings work better than TF-IDF to discriminate gender on the PAN-AP-2016 corpus. It should be highlighted the contribution of the authors of [53, 54, 2] since they obtained the best

results in three editions of PAN from 2013 to 2015 with their second order representation based on relationships between documents and profiles. Finally, the authors of [6] obtained the best results at PAN 2017 with combinations of n-grams.

3 Evaluation Framework

The purpose of this section is to introduce the technical background. We outline the construction of the corpus, introduce the performance measures and baselines, and describe the idea of so-called software submissions.

3.1 Corpus

To build the PAN-AP-2019 corpus[13] we have combined Twitter accounts identified as bots in existent datasets [92, 52, 17, 18, 16] with newly discovered ones on the basis of specific search queries. Firstly, we downloaded and manually inspected the Twitter accounts identified in the previous datasets in order to ensure that the accounts still remain in Twitter. As Twitter has removed millions of bots from its platform,[14] we have looked for new ones on the basis of search queries such as "I'm a bot". Moreover, other bots relying on more elaborated technologies such as Markov chains or metaphors have been considered. For example, the bot *@metaphormagnet* was developed by Tony Veale and Goufu Li [93] to automatically generate metaphorical language, or the bot *@markov_chain* read periodically the latest tweets and uses Markov chains to generate related contents. Once the accounts were identified, we manually annotated them with the agreement of at least two annotators[15]. If some of the annotators disagreed, the Twitter user was discarded.

We have selected humans from the corpora created in previous editions of the author profiling shared task [78, 80]. Nonetheless, we have performed a new manual review of the annotation to ensure quality. Table 1 overviews the key figures of the corpus. The corpus is completely balanced per type (bot / human), and in case of human, it is also completely balanced per gender. Each author is composed of exactly 100 tweets.

[13] We should highlight that we are aware of the legal and ethical issues related to collecting, analysing and profiling social media data [77] and that we are committed to legal and ethical compliance in our scientific research and its outcomes.

[14] https://mashable.com/article/twitter-removing-followers-locked-accounts
https://www.reuters.com/article/us-usa-election-twitter-exclusive/exclusive-twitter-deletes-over-10000-accounts-that-sought-to-discourage-u-s-voting-idUSKCN1N72FA
https://thehill.com/policy/technology/440187-twitter-removes-5000-bot-accounts-promoting-russiagate-hoax

[15] The collected Twitter accounts were previously annotated, on the one hand by the authors who created the datasets, on the other hand by the owner of the Twitter account described as "I'm a bot" and similar queries.

Table 1. Number of authors per language. The corpus is balanced regarding bots vs. humans, and regarding gender in case of humans, and it contains 100 tweets per author.

	(EN) English				(ES) Spanish			
	Bots	Female	Male	Total	Bots	Female	Male	Total
Training	2,060	1,030	1,030	4,120	1,500	750	750	3,000
Test	1,320	660	660	2,640	900	450	450	1,800
Total	3,380	1,690	1,690	6,760	2,400	1,200	1,200	4,800

While annotating bots we found that most of them could be classified into predefined classes. Concretely, we defined the following taxonomy and classified each bot in one of these classes:

- *Template*: the Twitter feed responds to a predefined structure or template, such as for example a Twitter account giving the state of the earthquakes in a region or job offers in a sector.
- *Feed*: the Twitter feed retweets or shares news about a predefined topic, such as for example regarding Trump's policies.
- *Quote*: the Twitter feed reproduces quotes from famous books or songs, quotes from celebrities (or historical) people, or jokes.
- *Advanced*: Twitter feeds whose language is generated on the basis of more elaborated technologies such as Markov chains, metaphors, or in some cases, randomly choosing and merging texts from big corpus.

The information regarding this taxonomy was not released publicly and its only purpose was to analyse more in-depth the error made by the participants of the shared task.

3.2 Performance Measures

The participants were asked to send two predictions per author: *i)* whether the author is a bot or a human; and *ii)* in case of a human, whether the author is male or female. The participants were allowed to approach the task also in one of the languages and to address only one problem (bots or gender). The accuracy has been used for evaluation. For each language, we obtain the accuracy for both problems in both languages separately and average them to obtain the final ranking:

$$ranking = \frac{bots_{en} + bots_{es} + gender_{en} + gender_{es}}{4} \quad (1)$$

3.3 Baselines

In order to assess the complexity of the subtasks per language and to compare the performance of the participants' approaches, we propose the following baselines:

- *BASELINE-majority.* A statistical baseline that always predicts the majority class in the training set. In case of balanced classes, it predicts one of them.

– *BASELINE-random.* A baseline that randomly generates the predictions among the different classes.
– *BASELINE-char n-grams,* with values for n from 1 to 10, and selecting the 100, 200, 500, 1,000, 2,000, 5,000 and 10,000 most frequent ones.
– *BASELINE-word n-grams,* with values for n from 1 to 10, and selecting the 100, 200, 500, 1,000, 2,000, 5,000 and 10,000 most frequent ones.
– *BASELINE-W2V [59, 60].* Texts are represented with two word embedding models: *i)* Continuous Bag of Words (CBOW); and *ii)* Skip-Grams.
– *BASELINE-LDSE [79].* This method represents documents on the basis of the probability distribution of occurrence of their words in the different classes. The key concept of LDSE is a weight, representing the probability of a term to belong to one of the different categories: human / bot, male / female. The distribution of weights for a given document should be closer to the weights of its corresponding category. LDSE takes advantage of the whole vocabulary.

For all the methods we have experimented with several machine learning algorithms (below), although we will report only the best performing one in each case.

– Bayessian methods: Naive Bayes (NB), Naive Bayes Multinomial (NBM), Naive Bayes Multinomial Text (NBMT), Naive Bayes Multinomial Updateable (NBMU), and Bayes Net (BN).
– Logistic methods: Logistic Regression (LR), and Simple Logistic (SL).
– Neural Networks: Multilayer Perceptron (MP), and Voted Perceptron (VP).
– Support Vector Machines (SVM).
– Rule-based methods: Decision Table (DT).
– Trees: Decision Stump, Hoeffding Tree (HT), J48, LMT, Random Forest (RF), Random Tree, and REP Tree.
– Lazy methods: KStar.
– Meta-classifiers: Bagging, Classification via Regression, Multiclass Classifier (MCC), Multiclass Classifier Updateable (MCCU), Iterative Classifier Optimize.

Finally, we have used the following configurations:

– BASELINE-char n-grams:
  • BOTS-EN: 500 characters 5-grams + Random Forest
  • BOTS-ES: 2,000 characters 5-grams + Random Forest
  • GENDER-EN: 2,000 characters 4-grams + Random Forest
  • GENDER-ES: 1,000 characters 5-grams + Random Forest
– BASELINE-word n-grams:
  • BOTS-EN: 200 words 1-grams + Random Forest
  • BOTS-ES: 100 words 1-grams + Random Forest
  • GENDER-EN: 200 words 1-grams + Random Forest
  • GENDER-ES: 200 words 1-grams + Random Forest
– BASELINE-W2V:
  • BOTS-EN: glove.twitter.27B.200d + Random Forest

    - BOTS-ES: fasttext-wikipedia + J48
    - GENDER-EN: glove.twitter.27B.100d + SVM
    - GENDER-ES: fasttext-sbwc + SVM
- BASELINE-LDSE:
    - BOTS-EN: LDSE.v2 (MinFreq=10, MinSize=1) + Naive Bayes
    - BOTS-ES: LDSE.v1 (MinFreq=10, MinSize=1) + Naive Bayes
    - GENDER-EN: LDSE.v1 (MinFreq=10, MinSize=3) + BayesNet
    - GENDER-ES: LDSE.v1 (MinFreq=2, MinSize=1) + Naive Bayes

3.4 Software Submissions

We asked for software submissions (as opposed to run submissions). Within software submissions, participants submit executables of their author profiling softwares instead of just the output (also called "run") of their softwares on a given test set. Our rationale to do so is to increase the sustainability of our shared task and to allow for the re-evaluation of approaches to Author Profiling later on, and, in particular, on future evaluation corpora. To facilitate software submissions, the TIRA experimentation platform was employed [31, 32], which renders the handling of software submissions at scale as simple as handling run submissions. Using TIRA, participants deploy their software on virtual machines at our site, which allows us to keep them in a running state [33].

4 Overview of the Submitted Approaches

This year, 56 teams[16] participated in the Author Profiling shared task and 46 of them submitted the notebook paper[17]. We analyse their approaches from three perspectives: preprocessing, features to represent the authors' texts, and classification approaches.

4.1 Preprocessing

Various participants cleaned the textual contents to obtain plain text. To this end, most of them removed, normalised or masked Twitter specific elements such as URLs, user mentions, hashtags, emojis or reserved words (e.g., RTs, FAV) as well as emails, dates, money or numbers [91, 94, 70, 26, 30, 73, 81, 68, 90, 64, 27, 21, 97, 57]. The authors of [30, 45] applied word segmentation to split hashtags into the corresponding words. The authors of [91, 70, 30, 45, 5, 68, 34, 97, 57] tokenised texts and the authors of [40, 45, 81, 5, 68, 27, 34, 97] applied stemming or lemmatisation, depending on the language. Punctuation marks were removed by the authors of [94, 81, 64, 63, 34, 21, 97]. The authors of [91, 94, 26, 81, 63] lowercased the tweets, removed stopwords [45, 81, 27, 97] and treated character flooding [94, 30, 34]. In order to reduce the dimensionality, LSA was applied by the authors of [74], while the authors of [40, 30] removed words that appear less than a given frequency in the training corpus. Similarly, the authors of [94] removed words with less than a given number of characters. Conversely, the authors of [45, 81] unfolded contractions and acronyms.

[16] The authors of [48] could not finish before the deadline, hence they are considered out-of-competition.

[17] Regretfully, some working notes had to be rejected due to lack of scientific quality.

4.2 Features

As in previous editions of the author profiling task at PAN, participants used a high variety of different features. We can group them into three main groups: *i)* n-grams; *ii)* stylistics; and *iii)* embeddings. Traditional features such as character and word n-grams have been widely used [41, 11, 74, 90]. Both Mahmood *et al.* [57] and Fahim *et al.* [84] used bag-of-words (word unigrams) as text representation. Espinosa *et al.* [21] used character n-grams while Pizarro [69] and Przybyla *et al.* used word n-grams. Combinations of both character and word n-grams where used by the authors of [85, 58, 19, 94, 26]. The authors of [50, 27, 81, 45, 5, 43, 23, 35] weighted the n-grams with tf/idf, whereas Van Halteren *et al.* [91] used character n-grams from tokens. Finally, Gishamer *et al.* [30] used Part-of-Speech (POS) n-grams.

Some authors measured the stylistic variation of the tweets by counting the occurrence of some types of elements [45, 34, 4, 15]. For example, Oliveira *et al.* [63] counted the use of function words, Ikae *et al.* [40] counted the use of articles and personal pronouns, similar to De la Peña [50] who counted verbs, adjectives and pronouns. Puertas *et al.* counted the use of hashtags, mentions, URLs or emojis, Johansson [43] counted the number of words in capital and lower letters, as well as the number of urls, user mentions and RTs, and Giachanou and Ghanem [26] counted the use of punctuation marks such as exclamations and questions, the number of terms in capital letters, the use of mentions, links and hashtags, or the occurrence of words with character flooding. Retweet ratios, tweets length, and ratios of unique words were also used by authors such as Martinc *et al.* [58], Przybyla *et al.* [72], Van Halteren [91] and Fernquist *et al.* [23].

Some authors [15] also used emotional features. Giachanou and Ghanem [26] used emotional words, Oliveira *et al.* [63] used the emoticons and both of them used sentiments or/and polarity words. Polignano *et al.* [70], Fagni *et al.* [22], Halvani *et al.* [36] and Onose *et al.* [64] used different embedding-based features to represent text. Similarly, López-Santillan *et al.* [55] and Staykovsky *et al.* [86] combined word embeddings with tf-idf, whereas Joo *et al.* [45] used document-level embeddings. Apart from the previous approaches, Gamallo *et al.* [25] used lexicon-based features and Fernquist *et al.* [23] combined different compression algorithms. Finally, it is worth to mention the DNA-based approach by Kosmajac *et al.* [47].

4.3 Classification Approaches

Regarding the classification approaches, most participants used traditional approaches, mainly Support Vector Machines (SVM) [94, 15, 22, 69, 42, 35, 5, 34, 85, 57, 21, 63, 27, 74]. Some authors ensembled SVM with Logistic Regression [30, 62]. The last authors also included in the ensemble SpaCy and Random Forest. The authors of [26] used SVM only for the gender identification subtask, using Stochastic Gradient Descent (SGD) for the bots vs. human discrimination. Participants also used other traditional approaches such as Logistic Regression [90, 9, 72], SGD [11], Random Forest [43], Decision Trees [81], Multinomial BayesNet [81], Naive Bayes [25], Adaboost together with SVM [5], CatBoost [23], kNN [40], and Mutilayer Perceptron [86].

Only few participants approached the task with deep learning methods. The authors of [19, 68] combined Convolutional Neural Networks (CNNs) with Recurrent Neural

Networks (RNNs). CNNs have been used by the authors of [70, 24] and RNNs by the authors of [9, 64], the first author only used RNNs for the gender identification subtask, and the second one together with hierarchical attention. Finally, the authors of [45] used a BERT model, the authors of [36, 50] used Feedforward Neural Networks, and the authors of [97] a voted LSTM.

5 Evaluation and Discussion of the Results

Although we recommended to participate in both subtasks, bots and gender profiling, some participants approached only one problem, or / and in just one language: English or Spanish. Therefore, we present the results separately in order to take into account this fact.

5.1 Global Ranking

In Table 3 the overall performance per language and users' ranking are shown. The best results have been obtained in English for both bots (95.95% vs. 93.33% in Spanish) and gender (84.17% vs. 81.72% in Spanish) profiling. The best results per language and problem are highlighted in bold font. The overall best result (88.05%), as well as the best result for both tasks in Spanish (93.33% and 81.72%), have been obtained by Pizarro [69]. He has approached the task with a Support Vector Machine with character and word n-grams features. The best result for bots discrimination in English (95.95%) has been obtained by Johansson [43]. He has approached the task with Random Forest and several features such as term frequencies together with aggregated stats (tweets length, number of capital letters, lower letters, URLs, mentions, RTs ratios, etc.). In case of gender identification in English, the best result (84.32%) has been obtained by Valencia *et al.* [90]. They have approached the task with n-grams and Logistic Regression. It should be highlighted the high results obtained by the word and character n-grams baselines, even greater than word embeddings [59, 60] and the Low Dimensionality Statistical Embedding (LDSE) [79]. In this vein, it is worth to mention that the four teams with the highest performance [69, 85, 5, 42] used combinations of n-grams with SVM and the fifth one [23] used CatBoost. The first time a deep learning approach appears, concretely a CNN, is in the eleventh position [70].

Table 2. Statistics on the accuracy per task and language.

Stat	**Bots vs. Human**		**Gender**		**Average**
	EN	ES	EN	ES	
Min	0.4163	0.4744	0.2511	0.2567	0.3784
Q1	0.8719	0.8307	0.7220	0.6986	0.7906
Median	0.9057	0.8698	0.7731	0.7206	0.8176
Mean	0.8615	0.8408	0.7279	0.7017	0.7932
SDev	0.1239	0.1020	0.1385	0.1031	0.0975
Q3	0.9159	0.8954	0.8174	0.7564	0.8432
Max	0.9595	0.9333	0.8432	0.8172	0.8805
Skewness	-2.4679	-2.5676	-1.7444	-2.5298	-2.7257
Kurtosis	7.9291	9.1006	5.2570	10.5175	10.7926
Normality (p-value)	2.20e-16	4.81e-12	2.86e-12	5.13e-08	4.78e-11

Table 3. Accuracy per language and global ranking as average per language.

Ranking	Team	Bots vs. Human		Gender		Average
		EN	ES	EN	ES	
1	Pizarro	0.9360	**0.9333**	0.8356	**0.8172**	**0.8805**
2	Srinivasarao & Manu	0.9371	0.9061	0.8398	0.7967	0.8699
3	Bacciu et al.	0.9432	0.9078	0.8417	0.7761	0.8672
4	Jimenez-Villar et al.	0.9114	0.9211	0.8212	0.8100	0.8659
5	Fernquist	0.9496	0.9061	0.8273	0.7667	0.8624
6	Mahmood	0.9121	0.9167	0.8163	0.7950	0.8600
7	Ipsas & Popescu	0.9345	0.8950	0.8265	0.7822	0.8596
8	Vogel & Jiang	0.9201	0.9056	0.8167	0.7756	0.8545
9	Johansson & Isbister	**0.9595**	0.8817	0.8379	0.7278	0.8517
10	Goubin et al.	0.9034	0.8678	0.8333	0.7917	0.8491
11	Polignano & de Pinto	0.9182	0.9156	0.7973	0.7417	0.8432
12	Valencia et al.	0.9061	0.8606	**0.8432**	0.7539	0.8410
13	Kosmajac & Keselj	0.9216	0.8956	0.7928	0.7494	0.8399
14	Fagni & Tesconi	0.9148	0.9144	0.7670	0.7589	0.8388
	char nGrams	0.9360	0.8972	0.7920	0.7289	0.8385
15	Glocker	0.9091	0.8767	0.8114	0.7467	0.8360
	word nGrams	0.9356	0.8833	0.7989	0.7244	0.8356
16	Martinc et al.	0.8939	0.8744	0.7989	0.7572	0.8311
17	Sanchis & Velez	0.9129	0.8756	0.8061	0.7233	0.8295
18	Halvani & Marquardt	0.9159	0.8239	0.8273	0.7378	0.8262
19	Ashraf et al.	0.9227	0.8839	0.7583	0.7261	0.8228
20	Gishamer	0.9352	0.7922	0.8402	0.7122	0.8200
21	Petrik & Chuda	0.9008	0.8689	0.7758	0.7250	0.8176
22	Oliveira et al.	0.9057	0.8767	0.7686	0.7150	0.8165
	W2V	0.9030	0.8444	0.7879	0.7156	0.8127
23	De La Peña & Prieto	0.9045	0.8578	0.7898	0.6967	0.8122
24	López Santillán et al.	0.8867	0.8544	0.7773	0.7100	0.8071
	LDSE	0.9054	0.8372	0.7800	0.6900	0.8032
25	Bolonyai et al.	0.9136	0.8389	0.7572	0.6956	0.8013
26	Moryossef	0.8909	0.8378	0.7871	0.6894	0.8013
27	Zhechev	0.8652	0.8706	0.7360	0.7178	0.7974
28	Giachanou & Ghanem	0.9057	0.8556	0.7731	0.6478	0.7956
29	Espinosa et al.	0.8413	0.7683	0.8413	0.7178	0.7922
30	Rahgouy et al.	0.8621	0.8378	0.7636	0.7022	0.7914
31	Onose et al.	0.8943	0.8483	0.7485	0.6711	0.7906
32	Przybyla	0.9155	0.8844	0.6898	0.6533	0.7858
33	Puertas et al.	0.8807	0.8061	0.7610	0.6944	0.7856
34	Van Halteren	0.8962	0.8283	0.7420	0.6728	0.7848
35	Gamallo & Almatarneh	0.8148	0.8767	0.7220	0.7056	0.7798
36	Bryan & Philipp	0.8689	0.7883	0.6455	0.6056	0.7271
37	Dias & Paraboni	0.8409	0.8211	0.5807	0.6467	0.7224
38	Oliva & Masanet	0.9114	0.9111	0.4462	0.4589	0.6819
39	Hacohen-Kerner et al.	0.4163	0.4744	0.7489	0.7378	0.5944
40	Kloppenburg	0.5830	0.5389	0.4678	0.4483	0.5095
	MAJORITY	0.5000	0.5000	0.5000	0.5000	0.5000
	RANDOM	0.4905	0.4861	0.3716	0.3700	0.4296
41	Bounaama & Amine	0.5008	0.5050	0.2511	0.2567	0.3784
42	Joo & Hwang	0.9333	-	0.8360	-	0.4423
43	Staykovski	0.9186	-	0.8174	-	0.4340
44	Cimino & Dell'Orletta	0.9083	-	0.7898	-	0.4245
45	Ikae et al.	0.9125	-	0.7371	-	0.4124
46	Jeanneau	0.8924	-	0.7451	-	0.4094
47	Zhang	0.8977	-	0.7197	-	0.4044
48	Fahim et al.	0.8629	-	0.6837	-	0.3867
49	Saborit	-	0.8100	-	0.6567	0.3667
50	Saeed & Shirazi	0.7951	-	0.5655	-	0.3402
51	Radarapu	0.7242	-	0.4951	-	0.3048
52	Bennani-Smires	0.9159	-	-	-	0.2290
53	Gupta	0.5007	-	0.4044	-	0.2263
54	Qurdina	0.9034	-	-	-	0.2259
55	Aroyehun	0.5000	-	-	-	0.1250

As can be seen in Figure 1 and Table 2, the results for the bots vs. human task in English are higher and slightly less sparse than in Spanish. Although the average is similar for both languages (86.15% vs. 84.08%), in the case of English the median is 90.57% with an inter-quartile range of 4.4%, whereas in the case of Spanish the median is 84.08% with an inter-quartile range of 6.47%. Nevertheless, in the case of English, the standard deviation is higher (12.39% vs. 10.20%), due to the higher number of outliers (see Figure 2 and 3). In the case of gender, the average results in English (72.79%) are also higher than in Spanish (70.17%). In this case, the sparsity is higher in the case of English, with an inter-quartile range of 9.54% vs. 5.78% in Spanish. Due to the several outliers in the case of English, the standard deviation (13.85%) is also higher than in Spanish (10.31%). We can conclude that notwithstanding most systems obtained better results in the case of English for bot tasks, several systems obtained low accuracy and reduced the average as well as increased the sparsity.

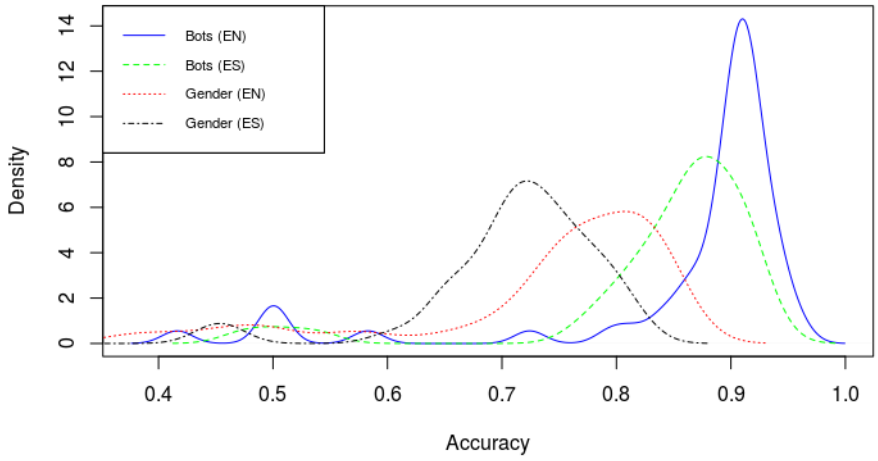

Figure 1. Density of the results for both tasks in the different languages.

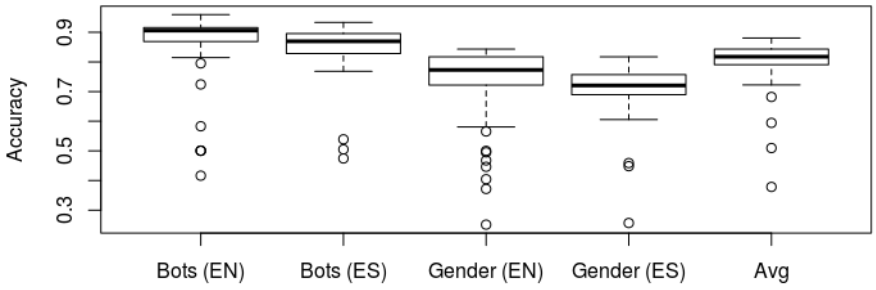

Figure 2. Distribution of results for both tasks in the different languages.

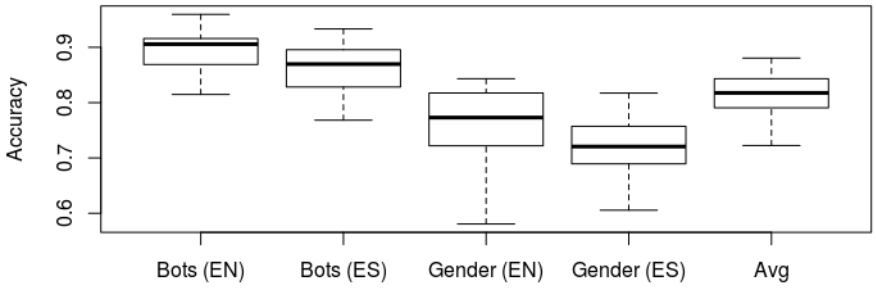

Figure 3. Distribution of results for both tasks in the different languages (without outliers).

5.2 Error Analysis

In this section we perform an in-depth error analysis. Firstly, confusion matrices are plotted and analysed. Then, we explore when a bot is wrongly classified as a human, taking into account the type of bot as well as the predicted gender. Finally, we analyse the humans that wrongly were identified as bots, also from the gender perspective.

Confusion Matrices We have aggregated all the participants' predictions for the bots vs. human discrimination task, except baselines, and plotted the respective confusion matrices for English and Spanish in Figures 4 and 5, respectively. In the case of English, the highest confusion is from bots to humans (17.15% vs. 7.86%). Nonetheless, in the case of Spanish the confusion is similar in both cases (14.45% vs. 14.08%, respectively for humans to bots and bots to humans).

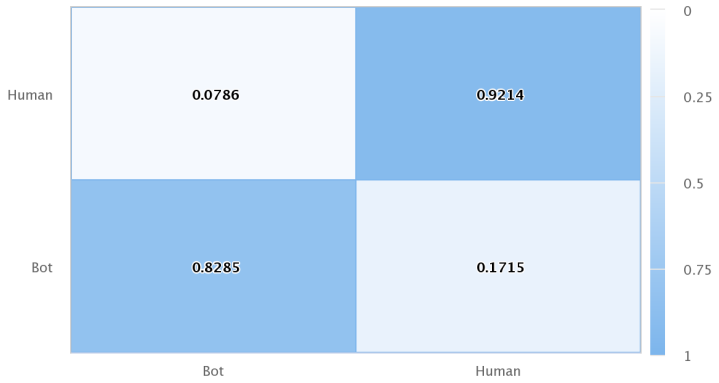

Figure 4. Aggregated confusion matrix for bots vs. human discrimination in English.

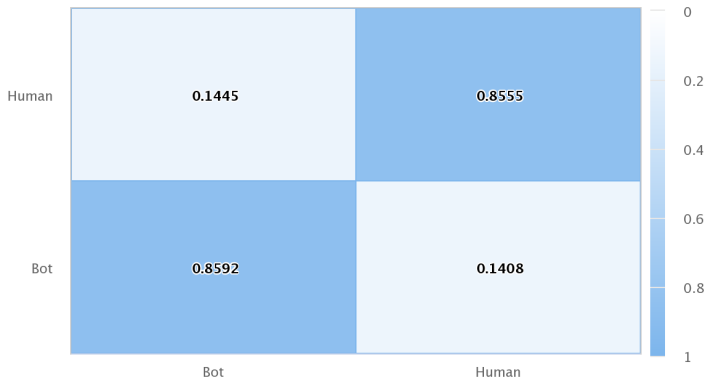

Figure 5. Aggregated confusion matrix for bots vs. human discrimination in Spanish.

In Figures 6 and 7 we have aggregated the predictions for the gender identification task, respectively for English and Spanish. In both languages, bots are mainly confused for males, although the difference with females is lower in the case of English (9.83% vs. 7.53%) than in Spanish (8.5% vs. 5.02%). Similarly, also males are more confused with bots than females, and again this difference is lower in the case of English (8.85% vs. 3.55%) than in Spanish (18.93% vs. 11.61%). Within genders, the confusion is very similar in the case of English (27.56% from males to females vs. 26.67% from females to males), whereas the difference is much higher in Spanish (21.03% from males to females vs. 11.61% from females to males).

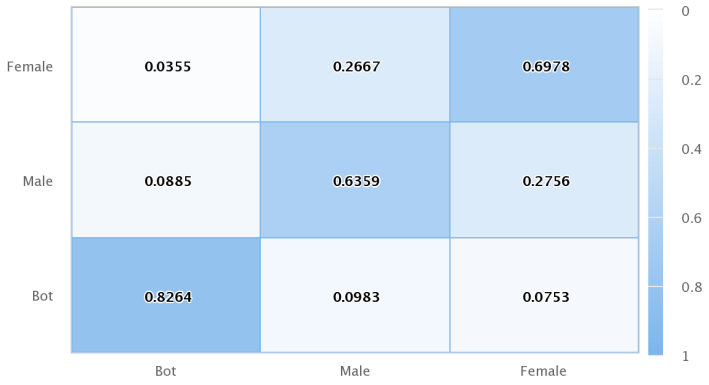

Figure 6. Aggregated confusion matrix for gender identification in English.

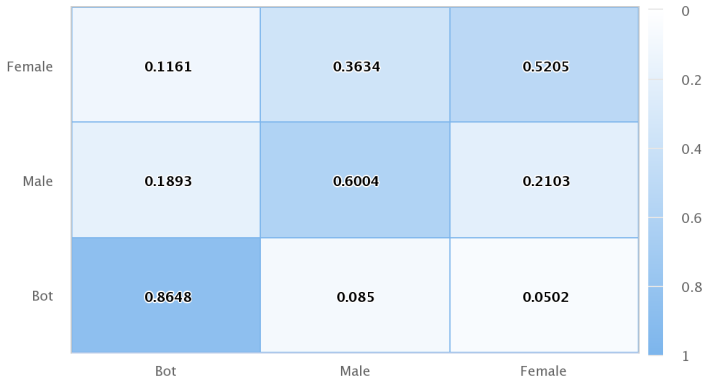

Figure 7. Aggregated confusion matrix for gender identification in Spanish.

Errors per Bot Type For each participant, we have obtained the number of errors per bot type. Then, we have obtained the basic statistics shown in Table 4 and represented their distribution in Figure 8. In both languages, as it was expected, **the number of errors is higher in case of advanced bots (average error rate of 30.11% and 32.38% respectively for English and Spanish)**. It is worth mentioning that in both languages, although specially in the case of Spanish, the inter-quartile range is higher also for advanced bots (26.88% in the case of English, 36.5% in case of Spanish), meaning a high variability in the systems' behaviour. In the case of English, quote bots were identified with similar number of errors (as well as its variability) than template errors. On the contrary, in Spanish quote bots were almost equally difficult to be identified than

advanced bots for most systems (median of 19.64% vs. 23%), similarly to what occurs with feed bots and advanced bots in English (median of 21.16% vs. 24.37%).

Table 4. Statistics of the errors per bot type.

Stat	Template		Feed		Quote		Advanced	
	EN	ES	EN	ES	EN	ES	EN	ES
Min	0.0000	0.0000	0.0000	0.0267	0.0000	0.0000	0.0000	0.0000
Q1	0.0091	0.0173	0.1605	0.0694	0.0552	0.1304	0.1187	0.1250
Median	0.0364	0.0538	0.2116	0.0811	0.0897	0.1964	0.2437	0.2300
Mean	0.1264	0.1320	0.2789	0.1428	0.1794	0.2651	0.3011	0.3238
SDev	0.2641	0.2194	0.2299	0.2029	0.2601	0.2175	0.2617	0.2771
Q3	0.0841	0.1327	0.2744	0.1161	0.1345	0.3625	0.3875	0.4900
Max	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Skewness	2.8051	2.9892	2.3214	3.6109	2.4970	1.9761	1.4789	0.8724
Kurtosis	9.4586	11.8837	7.7310	15.1821	7.9773	7.2431	4.7429	2.9126
Normality (p-value)	2.2e-16	2.2e-14	1.03e-15	2.2e-16	2.20e-16	2.49e-06	9.25e-07	0.0021

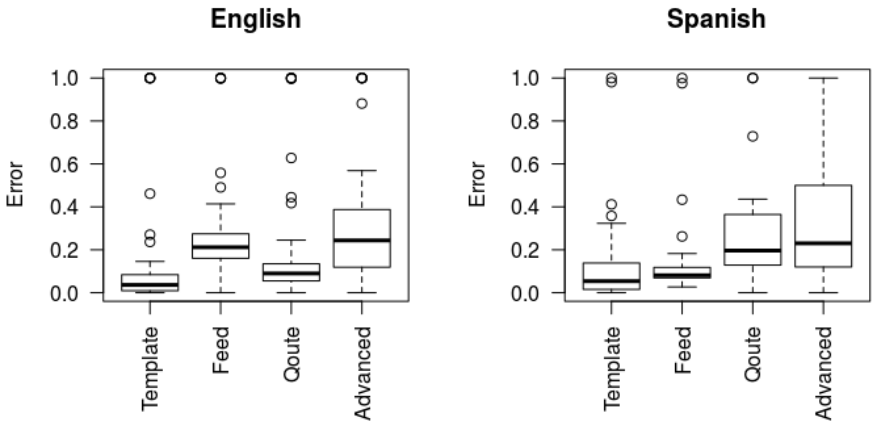

Figure 8. Distribution of the errors per bot type (English on the left, Spanish on the right).

In Figure 9 the number of systems failing in each of the predictions per bot type is shown for English. It can be seen that the highest sparsity occurs with feed bots, where most of the instances were properly predicted by most of the systems, but with several instances where most of the systems failed. This is similar in the case of quote bots, but with a higher number of systems failing on average. In the case of advanced bots, although the number of systems failing is more concentrated, there are no instances where at least five to ten systems failed.

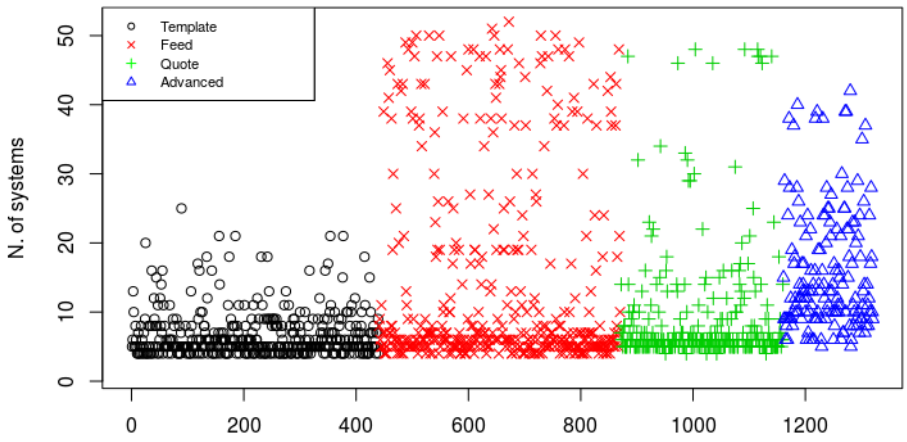

Figure 9. Errors per bot type in English.

Figure 10 represents the instances where at least half of the systems failed. In the case of template bots, only one instance was wrongly predicted by at least half of the systems. The highest number of systems failed in feed and quote bots. In the case of advanced bots, two groups of instances can be seen. In the first group, between twenty-five and thirty systems failed in the prediction. In the second one, between thirty-five and forty-five systems failed (almost all of them).

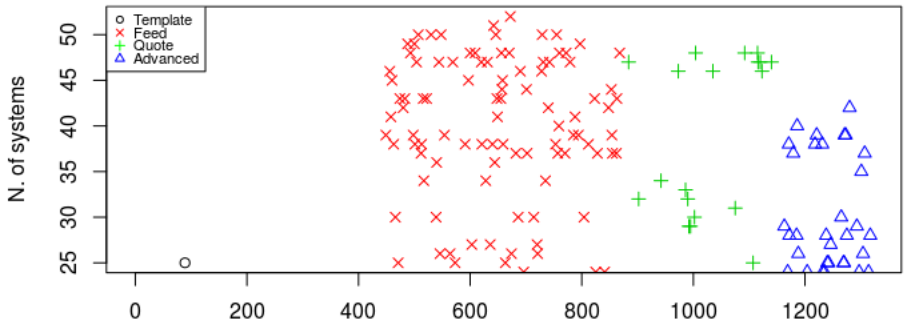

Figure 10. Errors per bot type in English when at least half of the systems failed.

Similarly in Spanish, the highest number of systems failed for the feed and quote bots (Figure 11), whereas the lowest sparsity occurs with advanced bots. In the case of

template bots, most of the instances were wrongly predicted by less than ten systems, whereas in the case of advanced bots, only few instances were properly predicted by all the systems, failing at least five in the vast majority of them.

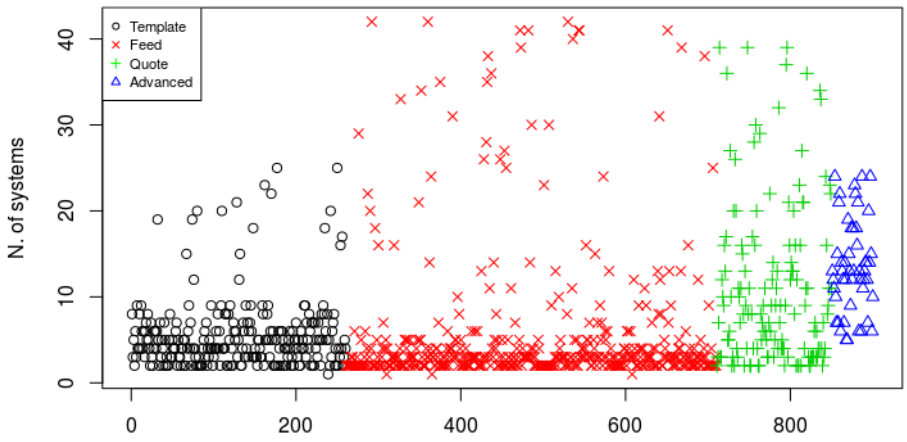

Figure 11. Errors per bot type in Spanish.

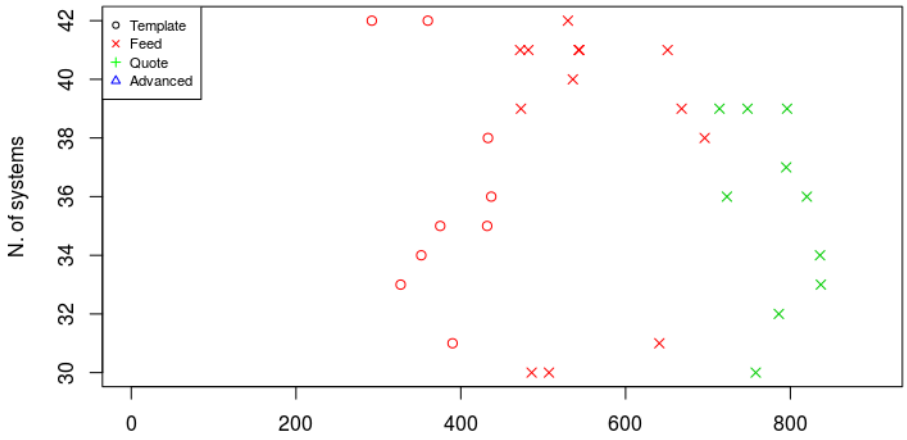

Figure 12. Errors per bot type in Spanish when at least half of the systems failed.

Looking at Figure 12 we can observe the instances where at least half of the systems failed. In the case of template bots, there is more sparsity than for English, with

instances where between twenty and twenty-five systems failed. This is similar to the case of advanced bots where there is a group of instances where between twenty and twenty-five systems failed. In none of these two kinds of bots there is an instance with more than twenty-five systems failing. In case of feed and quote bots, the number of instances with more than half of the systems failing is much higher. There are also a series of instances where almost all the systems failed.

In Tables 5 and 6 we can observe examples of wrongly classified bots as humans. The tables show the author id, the real Twitter account, and the type of bot. The last column shows the number of systems that failed in the classification and the total number of systems.

In the case of English, we can see that even the Twitter account reflects the kind of bot. For example, the @MessiQuote is a Twitter account that search for quotes from Messi and automatically tweets them, the bio of @NasaTimeMachine says "I'm a bot tasked with finding cool old photos from this day in NASA history. Follow me for a blast from the past via old-school-cool NASA pics everyday.", and the @markov_chain account is described as "I am a fan of Markov Chains. Every ten minutes I read the latest Tweets and work out what I'm going to tweet from them. Yes, I am a Twitter bot :)". However, as can be seen most of the systems failed detecting them.

Table 5. Top failing examples in English.

Author Id.	Twitter Account	Type	N. Systems
caf6d82d5dca1598beb5bfac0aea4161	@NasaTimeMachine	template	21 / 53
@wylejw You must be cool, I'll follow you!			
4c27d3c7a10964f574849b6be1df872d	@rarehero	feed	52 / 53
Get a doll, drape fabric and spray the hell out of it with Fabric Quick Stiffening Spray ... https://t.co/C9Ub6xXZWI via @duckduckgo			
8d08e3a0e1fea2f965fd7eb36f3b0b07	@MessiQuote	quote	48 / 53
.@PedroPintoUEFA: "Messi is unstoppable and we should feel privileged to be watching a player who may be the best of all time." https://t.co/TmCR6qCzO2			
6a6766790e1f5f67813afd7c0aa1e60d	@markov_chains	advanced	42 / 53
I have transferred to the local library go you! Just be Crazy John's prepaid sim card.			

In the case of Spanish, the bio of @Joker32191969 is "Bot AMLOVER", a word game that references the Mexican president Antonio Manuel López Obrador (AMLO) and which automatically publishes tweets supporting the president. Similarly, @Con_Sentimiento is a Twitter bot which automatically publishes love quotes, or @Online_DAM auto-defines itself as *"Official Distributor of Tamashii Nations and Megahouse in Mexico. My name is Tav-o and I am a Sociopat Bot, evil twin of Ultraman"*.[18]

[18] The official bio is in Spanish: "Distribuidor Oficial Tamashii Nations y Megahouse en México. Me llamo Tav-o soy un Robot Sociópata, gemelo malvado de Ultraman."

Table 6. Top failing examples in Spanish.

Author Id.	Twitter Account	Type	N. Systems
d0254a9765c8637b044dd2fa3788a103	@Online_DAM	template	25 / 42
¡Nueva edición Out of da Box! Presentando a Sailor Urano y Neptuno. http://t.co/KloUlv4l7j			
1416c9615d30d0e6fbe774496ffa5d0f	@Joker32191969	feed	42 / 42
RT @MiguelGRodri: @fernandeznorona @lopezobrador_@Taibo2Sitienederecho, pero hay que se prudente, que no mame en pleno proceso electora...			
d58008f7878fc9dd7cde1febaec65201	@Con_Sentimiento	quote	39 / 42
"Todo el mundo puede ser un capítulo, no todos llegan a ser historia."			
ded97b0a2efad0ba098311fe467b5136	@ClintHouseDosch	advanced	18 / 42
Acabo de llenar un sobremesa y lo de La Manada, y la universidad hemos llegado a degenerado a Danny DeVito y lo sucedido			

It is worth mentioning that most systems did not have much problems in identifying two advanced bots that we expected the systems to fail. Concretely, @metaphormagnet and @emailmktsales. The first one was developed by Tony Veale and Goufu Li [93] to automatically generate metaphorical language, and in the worst case, only 16 of 53 systems failed (e2cb393082f76b316bcd350d094ae100). The second one was developed by the first author of this overview with the aim at generating automatically contents related to marketing posing as human posts. In the worst case, only 13 of 53 systems failed (317354a53c2b7725f316421f8578cad0).

Bot to Human per Gender Errors Tables 7 and 8 show the statistics of the bots wrongly classified as humans, taking into account the predicted gender. As stated in Figures 4 and 5, on average the misclassification occurred mainly towards males. This is especially true in the case of feed bots in English (12.49% vs. 9.60%), and advanced bots in both English (13.90% vs. 9.48%) and Spanish (21.86% vs. 6.19%). It is worth mentioning that the only case were errors go towards females was in the case of quote bots in Spanish, where the average is almost double (6.75% vs. 15.29%) and the difference is even greater in case of the median (1.43% vs. 11.79%).

Table 7. Statistics of the errors per bot type and gender in English.

Stat	Template		Feed		Quote		Advanced	
	Male	Female	Male	Female	Male	Female	Male	Female
Min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q1	0.0000	0.0000	0.0512	0.0419	0.0345	0.0000	0.0187	0.0125
Median	0.0091	0.0045	0.1046	0.0744	0.0414	0.0172	0.0812	0.0687
Mean	0.0389	0.0370	0.1249	0.0960	0.0755	0.0537	0.1390	0.0948
SDev	0.1380	0.1397	0.1432	0.1402	0.1459	0.1474	0.1925	0.1458
Q3	0.0204	0.0182	0.1512	0.1070	0.0690	0.0379	0.1750	0.1250
Max	0.9954	1.0000	0.9953	1.0000	0.9966	1.0000	1.0000	1.0000
Skewness	6.4717	6.3896	4.3197	5.2448	5.1936	5.3702	2.8814	4.6631
Kurtosis	45.2338	44.3030	26.9656	34.0622	31.9446	33.8018	12.4398	29.2938
Normality (p-value)	2.2e-16	2.2e-16	4.3e-10	2.2e-16	2.2e-16	2.2e-16	1.7e-11	7.1e-13

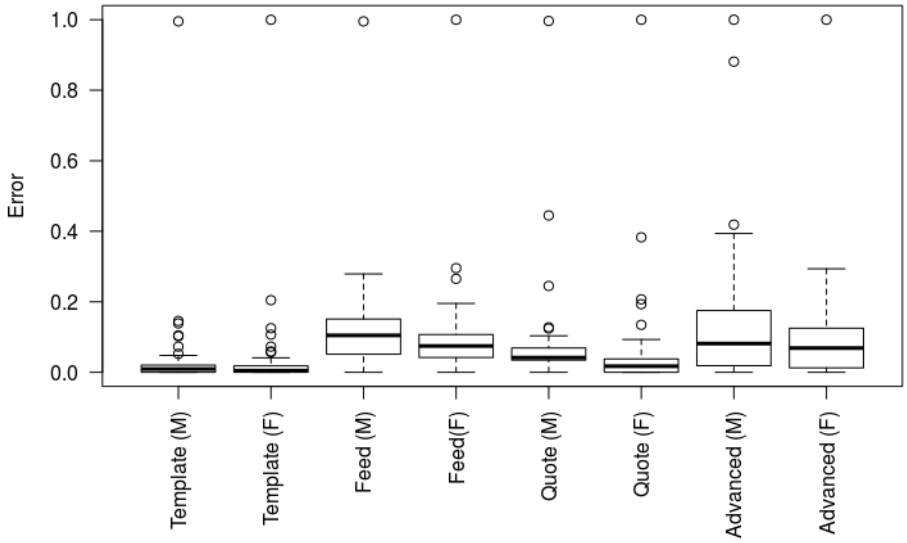

Figure 13. Errors per human in English.

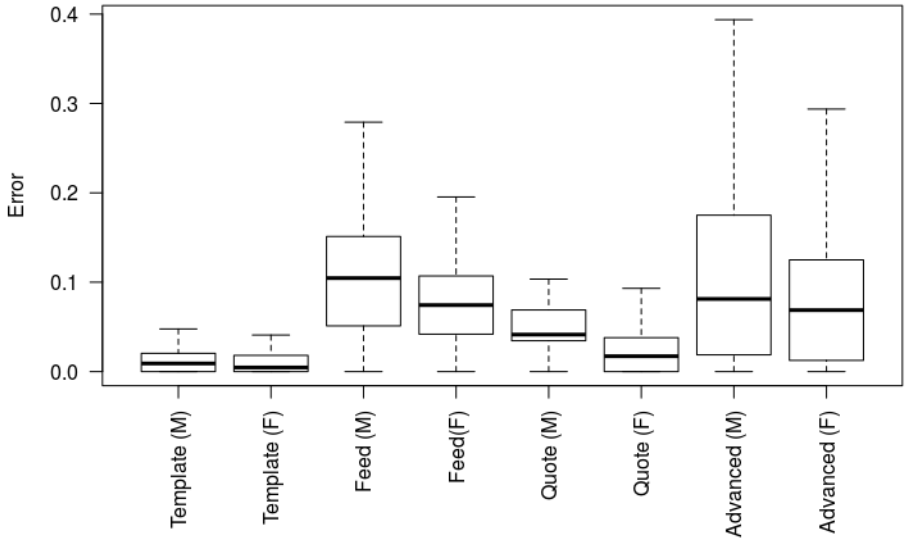

Figure 14. Errors per human in English.

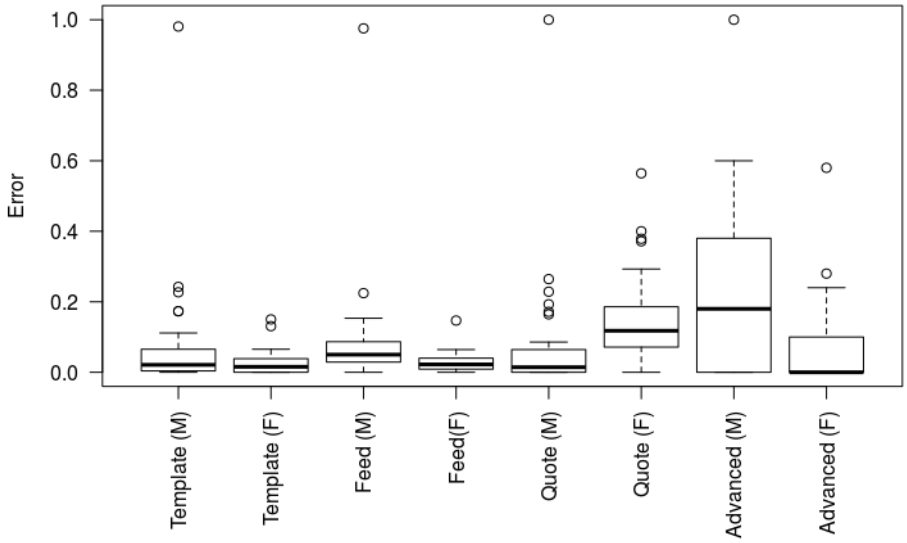

Figure 15. Errors per human in Spanish.

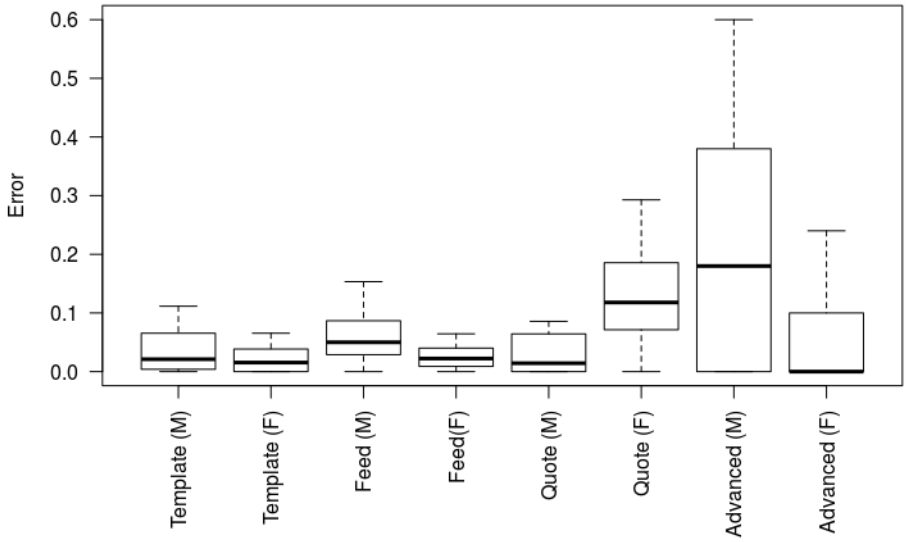

Figure 16. Errors per human in Spanish.

Table 8. Statistics of the errors per bot type and gender in Spanish.

Stat	Template		Feed		Quote		Advanced	
	Male	Female	Male	Female	Male	Female	Male	Female
Min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q1	0.0038	0.0000	0.0294	0.0094	0.0000	0.0750	0.0000	0.0000
Median	0.0211	0.0154	0.0500	0.0222	0.0143	0.1179	0.1800	0.0000
Mean	0.0689	0.0258	0.0776	0.0273	0.0675	0.1529	0.2186	0.0619
SDev	0.1567	0.0321	0.1482	0.0258	0.1615	0.1225	0.2338	0.1161
Q3	0.0654	0.0385	0.0811	0.0400	0.0643	0.1857	0.3800	0.0900
Max	0.9808	0.1500	0.9756	0.1467	1.0000	0.5643	1.0000	0.5800
Skewness	4.9026	2.1976	5.4956	2.4720	4.8096	1.2499	1.1546	2.5727
Kurtosis	28.7533	8.6738	33.7374	11.9685	27.8880	4.6554	4.1529	10.7066
Normality (p-value)	2.2e-16	7.6e-07	2.2e-16	0.0002	2.2e-16	0.0013	7.8e-05	2.4e-15

Human to Bot Errors Figures 17 and 19 show the number of systems which wrongly classified humans as bots, differentiating between genders. As can be seen, for both English and Spanish, the highest number of systems failed with male instances.

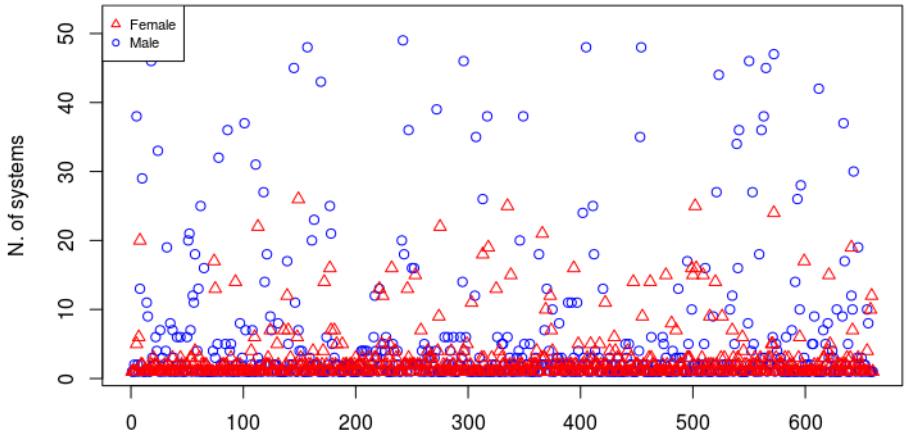

Figure 17. Number of systems wrongly classifying humans as bots, per gender, in English.

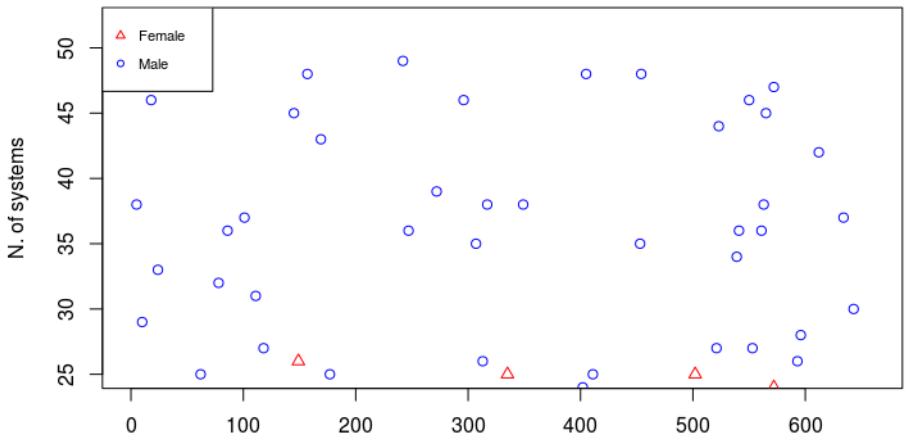

Figure 18. Number of systems wrongly classifying humans as bots, per gender, in English when at least half of the systems failed.

Figures 18 and 20 represent the instances where at least half of the systems failed. In the case of English it can be seen that almost all the instances correspond to male users. In the case of Spanish, although also males appear in the top failing instances, the distribution between genders is more homogeneous.

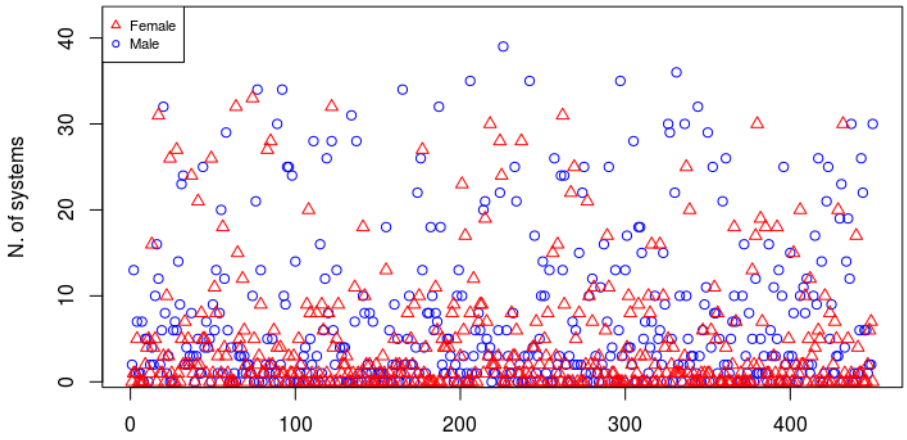

Figure 19. Number of systems wrongly classifying humans as bots, per gender, in Spanish.

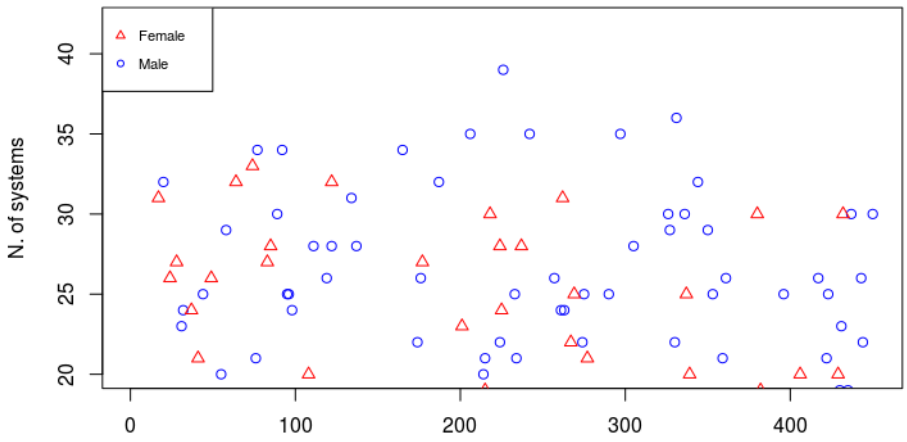

Figure 20. Number of systems wrongly classifying humans as bots, per gender, in Spanish when at least half of the systems failed.

In Tables 9 and 10 we have compiled some of the instances with the highest number of systems failing in the prediction.

Table 9. Top failing examples in English.

Author Id.	Twitter Account	Gender	N. Systems
63e4206bde634213b3a37343cf76e900	@Ask_KFitz	male	49 / 53
#Electric Imp Smart Refrigerator https://t.co/qigh5Womd7 https://t.co/JNVsRKvRQ8			
b11ffeeed0b38eb85e4e288f5c74f704	@iqbalmustansar	male	45 / 53
Trend - What's Dominating Digital Marketing Right Now? - https://t.co/dWp7ovqzCM			
ba0850ae38408f1db832707f1e0258fd	@CharBar_tweets	female	26 / 53
Hollywood boll #bowling #legs #Sundayfunday https://t.co/cLq9ZlNM38			
d64be10ecfbbb81d0c6e5b3115c335a5	@RheaRoryJames	female	25 / 53
RT @realDonaldTrump: Employment is up, Taxes are DOWN. Enjoy!			

Table 10. Top failing examples in Spanish.

Author Id.	Twitter Account	Gender	N. Systems
a22edd53bb04de0c06a52df897b13dd0	@carlosguadian	male	39 / 42
Tres días para analizar el presente y futuro de la Administración pública: lo que trae el Congreso NovaGob 2018 - NovaGob 2018 https://t.co/Ofc4cDTeym #novagob2018			
cf520c8e810a6a9bae9171d6f23c29be	@kicorangel	male	35 / 42
Google prepara una versión de pago para Youtube http://t.co/UvZdao68wc			
8e4340e95667c8add31f427a09dd3840	@EmaMArredondoM	female	30 / 42
@andrespino007 ¿Se ha preguntado cómo alguien llega a ser científico? Pequeña muestra chilena: https://t.co/fLjJsV0I0J			
6730bdf9686769c4a8a79d2f766a7f67	@Annie$_H$ *go*	female	24 / 42
Wow!! Nuevamente rebasamos expectivas... https://t.co/PJ8bHA1SrG			

If we go to the Twitter accounts, we can see that all of them can be easily confused with feed bots since their content consists mainly of sharing news or retweeting them. For instance, as can be seen in Figure 21, even Botometer assigns a score of 3.9 out of 5 in both contents and sentiment features to the @Ask_KFitz user, which would give a false positive.

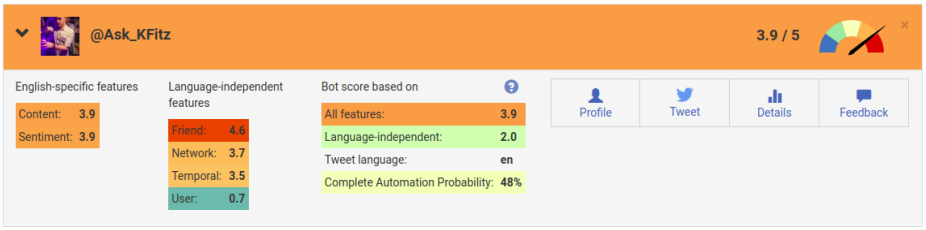

Figure 21. Botometer prediction for @Ask_Fitz user.

Notwithstanding the predictions, they are actually humans. For example, the user @kicorangel is the first author of this overview, who mainly uses Twitter to share news of his interest. Similarly, the user @carlosguadian is a friend of the first author who uses Twitter with the same purpose.

5.3 Best Results

In Table 11 we summarise the best results per language and task. We can observe that for both tasks the best results have been obtained in English, although with a slight difference. In case of bots vs. human, the best accuracy range from 93.33% in Spanish to 95.95% in English, while in case of gender identification it ranges between 81.72% in Spanish and 84.17% in English.

Table 11. Best results per language and problem.

Language	Bots vs. Human	Gender
English	0.9595	0.8417
Spanish	0.9333	0.8172

The best results in bots detection in English (95.95%) have been obtained by Johansson [43] who used Random Forest with a variety of stylistic features such as term occurrences, tweets length or number of capital and lower letters, URLs, user mentions, and so on. The best results in gender identification in English (84.32%) have been achieved by Valencia *et al.* [90] with Logistic Regression and n-grams. In Spanish, Pizarro [69] achieved the best results in both bots (93.33%) and gender identification (81.72%) with combinations of n-grams and Support Vector Machines.

6 Conclusion

In this paper we presented the results of the 7th International Author Profiling Shared Task at PAN 2019, hosted at CLEF 2019. The participants had to discriminate from Twitter authors between bots and humans, and in case of humans, to identify their gender. The provided data cover English and Spanish languages.

The participants used different features to address the task, mainly: *i)* n-grams; *ii)* stylistics; and *iii)* embeddings. With respect to machine learning algorithms, the most used one was Support Vector Machines. Nevertheless, few participants approached the task with deep learning techniques. In such cases, they used Convolutional Neural Networks, Recurrent Neural Networks, and FeedForward Neural Networks. According to the results, traditional approaches obtained higher accuracies than deep learning ones. The four teams with the highest performance [69, 85, 5, 42] used combinations of n-grams with SVM and the fifth one [23] used CatBoost. The first time a deep learning approach appears in the ranking, concretely a CNN, is in the eleventh position [70].

The best results have been obtained in English for both bots detection (95.95% vs. 93.33%) and gender identification (84.17% vs. 81.72%). The best results in bots detection in English have been obtained with a variety of stylistic features and Random Forest [43], whereas in Spanish were obtained with combinations of n-grams and Support Vector Machines [69]. Regarding gender, the best results in Spanish were achieved by the previous author, and the best results in English were obtained with n-grams and Logistic Regression [69].

The error analysis shows that the highest confusion is from bots to humans (17.15% vs. 7.86% in English, 14.45% vs. 14.08% in Spanish), and mainly towards males (9.83% vs. 7.53% in English, 8.5% vs. 5.02%). Similarly, males are also more confused with bots than females (8.85% vs. 3.55% in English, 18.93% vs. 11.61% in Spanish). Within genders, the confusion is similar in English (27.56% from males to females vs. 26.67% from females to males), whereas the difference is much higher in Spanish (21.03% from males to females vs. 11.61% from females to males).

The error analysis per bot type shows that the highest error on average was produced in case of advanced bots (30.11% and 32.38% respectively for English and Spanish). In the case of English, the systems failed in a similar rate on quote and template bots (12.64% and 17.94%), while the error is higher in the case of feed bots (27.89%). However, in Spanish template and feed bots obtained similar rate of error (13.20% and 14.28%), while quote bots error raises up to 26.51%. No matter the type of bot, the highest confusion is towards male, except in case of quote bots in Spanish (15.29% towards females vs. 6.75% towards males).

Looking at the results, the error analysis and the given misclassified examples, we can conclude that: *i)* it is feasible to automatically identify bots in Twitter with high precision, even when only textual features are used; but *ii)* there are specific cases where the task is difficult due to the language used by the bots (e.g., advanced bots), or due to the way the humans use the platform (e.g., to share news). In both cases, although the precision is high, a major effort needs to be made to take into account false positives.

Acknowledgements

Our special thanks goes to all PAN participants for providing high-quality submission, and to The Logic Value[19] for sponsoring the author profiling shared task award. The work of Paolo Rosso was partially funded by the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

Bibliography

[1] Hind Almerekhi and Tamer Elsayed. Detecting automatically-generated arabic tweets. In *AIRS*, pages 123–134. Springer, 2015.

[2] Miguel-Angel Álvarez-Carmona, A.-Pastor López-Monroy, Manuel Montes-Y-Gómez, Luis Villaseñor-Pineda, and Hugo Jair-Escalante. Inaoe's participation at pan'15: author profiling task—notebook for pan at clef 2015. 2015.

[3] Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *TEXT*, 23:321–346, 2003.

[4] Shaina Ashraf, Omer Javed, Muhammad Adeel, Haider Ali, and Rao Muhammad Adeel Nawab. Bots and gender prediction using language independent stylometry-based approach. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[5] Andrea Bacciu, Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, Valerio Neri, and Julinda Stefa. Bot and gender detection of twitter accounts using distortion and lsa. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[6] Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. N-gram: New groningen author-profiling model. *arXiv preprint arXiv:1707.03764*, 2017.

[7] Roy Bayot and Teresa Gonçalves. Multilingual author profiling using word embedding averages and svms. In *Software, Knowledge, Information Management & Applications (SKIMA), 2016 10th International Conference on*, pages 382–386. IEEE, 2016.

[19] https://thelogicvalue.com

[8] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. *First Monday, vol. 21 (11)*, 2016.

[9] $\mathrm{FlÃ}^3raBolonyai, JakabBuda, andEszterKatona. Botornot : Atwo - levelapproachinauthorprofiling.notebookforpanatclef2019. InLinda Cappellato and Nicola Ferro and Davi$

[10] Bayan Boreggah, Arwa Alrazooq, Muna Al-Razgan, and Hana AlShabib. Analysis of arabic bot behaviors. In *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pages 1–6. IEEE, 2018.

[11] Rabia Bounaama and Mohammed Amine Abderrahim. Tlemcen university at pan @ clef 2019: Bots and gender profiling task. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[12] David A Broniatowski, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American journal of public health*, 108(10):1378–1384, 2018.

[13] John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[14] Chiyu Cai, Linjing Li, and Daniel Zengi. Behavior enhanced deep bot detection in social media. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 128–130. IEEE, 2017.

[15] Andrea Cimino and Felice dell'Orletta. A hierarchical neural network approach for bots and gender profiling. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[16] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Fame for sale: Efficient detection of fake twitter followers. *Decision Support Systems*, 80:56–71, 2015.

[17] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 963–972. International World Wide Web Conferences Steering Committee, 2017.

[18] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing*, 15(4):561–576, 2017.

[19] Rafael Dias and Ivandre Paraboni. Combined cnn+rnn bot and gender profiling. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[20] John P Dickerson, Vadim Kagan, and VS Subrahmanian. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 620–627. IEEE Press, 2014.

[21] Daniel Yacob Espinosa, Helena Gómez-Adorno, and Grigori Sidorov. Bots and gender profiling using character bigrams. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[22] Tiziano Fagni and Maurizio Tesconi. Profiling twitter users using autogenerated features invariant to data distribution. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[23] Johan Fernquist. A four feature types approach for detecting bot and gender of twitter users. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[24] Michael FÃrber, Agon Qurdina, and Lule Ahmedi. Identifying twitter bots using a convolutional neural network. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[25] Pablo Gamallo and Sattam Almatarneh. Naive-bayesian classification for bot detection in twitter. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[26] Anastasia Giachanou and Bilal Ghanem. Bot and gender detection using textual and stylistic information. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[27] Hamed Babaei Giglou, Mostafa Rahgouy, Taher Rahgooy, Mohammad Karami Sheykhlan, and Erfan Mohammadzadeh. Author profiling: Bot and gender prediction using a multi-aspect ensemble approach. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[28] Zafar Gilani, Liang Wang, Jon Crowcroft, Mario Almeida, and Reza Farahbakhsh. Stweeler: A framework for twitter bot analysis. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 37–38. International World Wide Web Conferences Steering Committee, 2016.

[29] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. Of bots and humans (on twitter). In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 349–354. ACM, 2017.

[30] Flurin Gishamer. Using hashtags and pos-tags for author profiling. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[31] Tim Gollub, Benno Stein, and Steven Burrows. Ousting ivory tower research: towards a web framework for providing experiments as a service. In Bill Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *35th International*

ACM Conference on Research and Development in Information Retrieval (SIGIR 12), pages 1125–1126. ACM, August 2012. ISBN 978-1-4503-1472-5.

[32] Tim Gollub, Benno Stein, Steven Burrows, and Dennis Hoppe. TIRA: Configuring, executing, and disseminating information retrieval experiments. In A Min Tjoa, Stephen Liddle, Klaus-Dieter Schewe, and Xiaofang Zhou, editors, *9th International Workshop on Text-based Information Retrieval (TIR 12) at DEXA*, pages 151–155, Los Alamitos, California, September 2012. IEEE.

[33] Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Recent trends in digital text forensics and its evaluation. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative (CLEF 13)*, pages 282–302, Berlin Heidelberg New York, September 2013. Springer.

[34] Régis Goubin, Dorian Lefeuvre, Alaa Alhamzeh, Jelena Mitrović, and Elöd Egyed-Zsigmond. Bots and gender profiling using a multi-layer architecture. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[35] Yaakov HaCohen-Kerner, Natan Manor, and Michael Goldmeier. Bots and gender profiling of tweets using word and character n-grams. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[36] Oren Halvani and Philipp Marquardt. An unsophisticated neural bots and gender profiling system. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[37] Zakaria el Hjouji, D Scott Hunter, Nicolas Guenon des Mesnards, and Tauhid Zaman. The impact of bots on opinions in social networks. *arXiv preprint arXiv:1810.12398*, 2018.

[38] Janet Holmes and Miriam Meyerhoff. *The handbook of language and gender*. Blackwell Handbooks in Linguistics. Wiley, 2003.

[39] Philip N Howard, Samuel Woolley, and Ryan Calo. Algorithms, bots, and political communication in the us 2016 election: The challenge of automated political communication for election law and administration. *Journal of information technology & politics*, 15(2):81–93, 2018.

[40] Catherine Ikae, Sukanya Nath, and Jacques Savoy. Unine at pan-clef 2019: Bots and gender task. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[41] Adrian Ispas and Mircea Teodor Popescu. Normalized k3rn3l function (rejected). In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[42] Víctor Jimenez-Villar, Javier Sánchez-Junquera, Manuel Montes y Gómez, Luis Villase nor Pineda, and Simone Paolo Ponzetto. Bots and gender profiling

using masking techniques. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[43] Fredrik Johansson. Supervised classification of twitter accounts based on textual content of tweets. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[44] Marc Owen Jones. The gulf information warl propaganda, fake news, and fake trends: The weaponization of twitter bots in the gulf crisis. *International Journal of Communication*, 13:27, 2019.

[45] Youngjun Joo and Inchon Hwang. Author profiling on social media: An ensemble learning model using various features. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[46] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. literary and linguistic computing 17(4), 2002.

[47] Dijana Kosmajac and Vlado Keselj. Twitter user profiling: Bot and gender identification. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[48] György Kovács, Vanda Balogh, Kumar Shridhar, Purvanshi Mehta, and Pedro Alonso. Author profiling using semantic and syntactic features. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[49] Sneha Kudugunta and Emilio Ferrara. Deep neural networks for bot detection. *Information Sciences*, 467:312–322, 2018.

[50] Gretel Liz De la Peña Sarracén and Jose R. Prieto Fontcuberta. Bots and gender profiling using a deep learning approach. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[51] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442. ACM, 2010.

[52] Kyumin Lee, Brian David Eoff, and James Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[53] A. Pastor Lopez-Monroy, Manuel Montes-Y-Gomez, Hugo Jair Escalante, Luis Villasenor-Pineda, and Esau Villatoro-Tello. INAOE's participation at PAN'13: author profiling task—Notebook for PAN at CLEF 2013. In Pamela Forner, Roberto Navigli, and Dan Tufis, editors, *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*, September 2013.

[54] A. Pastor López-Monroy, Manuel Montes y Gómez, Hugo Jair-Escalante, and Luis Villase nor Pineda. Using intra-profile information for author profiling—Notebook for PAN at CLEF 2014. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*, September 2014.

[55] Roberto López-Santillán, Luis Carlos González-Gurrola, Manuel Montes y Gómez, Graciela Ramírez-Alonso, and Olanda Prieto-Ordaz. An evolutionary approach to build user representations for profiling of bots and humans in twitter. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[56] Suraj Maharjan, Prasha Shrestha, Thamar Solorio, and Ragib Hasan. A straightforward author profiling approach in mapreduce. In *Advances in Artificial Intelligence. Iberamia*, pages 95–107, 2014.

[57] Asad Mahmood and Padmini Srinivasan. Twitter bots and gender detection using tf-idf. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[58] Matej Martinc, BlaÅ$\frac{3}{4}$ Å krlj, and Senja Pollak. Fake or not: Distinguishing between bots, males and females. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[59] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at International Conference on Learning Representations (ICLR'13)*, 2013.

[60] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems pp. 3111–3119*, 2013.

[61] Fred Morstatter, Liang Wu, Tahora H Nazer, Kathleen M Carley, and Huan Liu. A new approach to bot detection: striking the balance between precision and recall. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 533–540. IEEE, 2016.

[62] Amit Moryossef. Ensembling classifiers for bots and gender profiling. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[63] Rodrigo Ribeiro Oliveira, Cláudio Moisés Valiense de Andrade, José Solenir Lima Figuerêdo, Jo ao B. Rocha-Junior, Rodrigo Tripodi Calumby, Iago Machado da Conceição Silva, and Almir Moreira da Silva Neto. Bot and gender identification: Textual analysis of tweets. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[64] Cristian Onose, Dumitru-Clementin Cercel, and Claudiu-Marcel Nedelcu. Bots and gender profiling using hierarchical attention networks. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and*

Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org, 2019.

[65] Christopher Paul and Miriam Matthews. The russian âfirehose of falsehoodâ propaganda model. *Rand Corporation*, pages 2–7, 2016.

[66] James W. Pennebaker. *The secret life of pronouns: what our words say about us*. Bloomsbury USA, 2013.

[67] James W. Pennebaker, Mathias R. Mehl, and Kate G. Niederhoffer. Psychological aspects of natural language use: our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[68] Juraj Petrik and Daniela Chuda. Bots and gender profiling with convolutional hierarchical recurrent neural network. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[69] Juan Pizarro. Using n-grams to detect bots on twitter. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[70] Marco Polignano, Marco Giuseppe de Pinto, Pasquale Lops, and Giovanni Semeraro. Identification of bot accounts in twitter using 2d cnns on user-generated contents. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[71] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

[72] Piotr Przybyła. Detecting bot accounts on twitter by measuring message predictability. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[73] Edwin Puertas, Luis Gabriel Moreno-Sandoval, Flor Miriam Plaza del Arco, Jorge Andres Alvarado-Valencia, Alexandra Pomares-Quimbaya, and L.Alfonso Ure na López. Bots and gender profiling on twitter using sociolinguistic features. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[74] Radarapu Rakesh, Yogesh Vishwakarma, Akkajosyula Surya Sai Gopal, , and Anand Kumar M. Bot and gender identification from twitter. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[75] Francisco Rangel and Paolo Rosso. On the multilingual and genre robustness of emographs for author profiling in social media. In *6th international conference of CLEF on experimental IR meets multilinguality, multimodality, and interaction*, pages 274–280. Springer-Verlag, LNCS(9283), 2015.

[76] Francisco Rangel and Paolo Rosso. On the impact of emotions on author profiling. *Information processing & management*, 52(1):73–92, 2016.

[77] Francisco Rangel and Paolo Rosso. On the implications of the general data protection regulation on the organisation of evaluation tasks. *Language and Law= Linguagem e Direito*, 5(2):95–117, 2018.

[78] Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In *Cappellato L., Ferro N., Goeuriot L, Mandl T. (Eds.) CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1866.*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2016.

[79] Francisco Rangel, Paolo Rosso, and Marc Franco-Salvador. A low dimensionality representation for language variety identification. In *17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing'16*. Springer-Verlag, LNCS(9624), pp. 156-169, 2018.

[80] Francisco Rangel, Paolo Rosso, Manuel Montes-y-Gómez, Martin Potthast, and Benno Stein. Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier, editors, *Working Notes Papers of the CLEF 2018 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2018.

[81] Usman Saeed and Farid Shirazi. Bots and gender classification on twitter. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[82] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205. AAAI, 2006.

[83] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):4787, 2018.

[84] Muhammad Hammad Fahim Siddiqui, Iqra Ameer, Alexander Gelbukh, and Grigori Sidorov. Bots and gender profiling on twitter. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[85] Mahendrakar Srinivasarao and Siddharth Manu. Bots and gender profiling using character and word n-grams. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[86] Todor Staykovski. Stacked bots and gender prediction from twitter feeds. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[87] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots sustain and inflate striking opposition in online social systems. *arXiv preprint arXiv:1802.07292*, 2018.

[88] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49):12435–12440, 2018.

[89] VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. The darpa twitter bot challenge. *Computer*, 49(6):38–46, 2016.

[90] Alex I. Valencia Valencia, Helena Gomez Adorno, Christopher Stephens Rhodes, and Gibran Fuentes Pineda. Bots and gender identification based on stylometry of tweet minimal structure and n-grams model. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[91] Hans van Halteren. Bot and gender recognition on tweets using feature count deviations: notebook for pan2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[92] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Eleventh international AAAI conference on web and social media*, 2017.

[93] Tony Veale and Guofu Li. Specifying viewpoint and information need with affective metaphors: A system demonstration of the metaphor magnet web app/service. In *Proceedings of the ACL 2012 System Demonstrations*, pages 7–12. Association for Computational Linguistics, 2012.

[94] Inna Vogel and Peter Jiang. Bot and gender identification in twitter using word and character n-grams. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.

[95] Edson Weren, Anderson Kauer, Lucas Mizusaki, Viviane Moreira, Palazzo de Oliveira, and Leandro Wives. Examining multiple features for author profiling. In *Journal of Information and Data Management*, pages 266–279, 2014.

[96] Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61, 2019.

[97] Iliya Zhechev, Zdravko Andonov, and Ivan Bozhilov. Bot and gender profiling based on voting lstm. notebook for pan at clef 2019. In *Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org*, 2019.