Author Profiling Using Semantic and Syntactic Features

Notebook for PAN at CLEF 2019

György Kovács[1,2], Vanda Balogh[3], Purvanshi Mehta[4], Kumar Shridhar[4], Pedro Alonso[1], and Marcus Liwicki[1]

[1]Embedded Internet Systems Lab, Luleå University of Technology, Luleå, Sweden
[2]MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
[3]Institute of Informatics, University of Szeged, Szeged, Hungary
[4]MindGarage, Kaiserslautern, Germany
gyorgy.kovacs@ltu.se, bvanda@inf.u-szeged.hu, purvanshi.mehta11@gmail.com, shridhar.stark@gmail.com, pedro.alonso@ltu.se, marcus.liwicki@ltu.se

Abstract In this paper we present an approach for the PAN 2019 Author Profiling challenge. The task here is to detect Twitter bots and also to classify the gender of human Twitter users as male or female, based on a hundred select tweets from their profile. Focusing on feature engineering, we explore the semantic categories present in tweets. We combine these semantic features with part of speech tags and other stylistic features – e.g. character floodings and the use of capital letters – for our eventual feature set. We have experimented with different machine learning techniques, including ensemble techniques, and found AdaBoost to be the most successful (attaining an F1-score of 0.99 on the development set). Using this technique, we achieved an accuracy score of 89.17% for English language tweets in the bot detection subtask.

1 Introduction

With the increasing use of social media [5], and its growing effect on our lives it is becoming more and more important to provide automatic methods that are capable of processing social media content. For one, it is paramount for companies interested in targeted advertisement to automatically identify certain traits of users, such as age, location, personality, and gender, even if the users do not report these traits themselves (although this application admittedly raises many ethical implications and challenges). More important is however the identification of fake news, and the detection of social media bots. With the growing role of social media as a primary news source [1], and the increasing effect of social media bots on political discourse [9] (in particular, their ability to effectively spread a large amount of misinformation in critical times [18]), it is vital to have the ability to monitor or even filter out such accounts. This, however, first requires the ability to efficiently identify such accounts. For this reason, when working on the *bots and gender profiling* PAN challenge [23,22], our main area of focus was the bot detection task.

1.1 Related Work

Social media analytics has a wide range of applications from understanding customer sentiment to determining the political orientation of a crowd. Another area of application of social media analytics that is growing rapidly is that of bot identification and fake news detection. The methods deployed in these tasks range from the use of various classical machine learning algorithms [8] to the more recent deep learning approaches [29].

Decision trees have been a popular choice in the task of bot vs human classification. For example, Botometer [30], a popular bot detection tool, uses random forests to identify twitter bots. Hall et al. [13] also applies random forests to remove bots from Wikipedia pages. One good quality of decision trees is that they work well with many languages, as their power to classify stance and gender in Spanish is shown in the works of Vinayakumar et al. [29] for the Ibereval 2017 task [27]. Besides decision trees, other well-known machine learning algorithms have also been used for the task, namely Support Vector Machines (SVMs) [7,29], Logistic Regression [7], and K-Nearest Neighbours [10]. Convolutional Neural Networks (CNNs) [14], Recurrent Neural Networks (RNNs) [7], and combinations of the two [3] have also been used for opinion detection in social media.

1.2 Bot and Gender Profiling

The research problem to be undertaken in this work is the PAN 2019 bot and gender profiling task. As the challenge is described in detail in accompanied overview papers [6,23], we only give a short description of the task here, and for more detail we refer the reader to the aforementioned publications. In this challenge, each team performs the task of classifying twitter profiles based on a randomly selected set of a hundred tweets, as bots or humans. Furthermore, in case an author is identified as human, the additional task is to identify the gender of said human as male or female. For submissions and evaluation, the PAN task uses TIRA virtual machines where teams upload and run their software[21]. The author profiling challenge is organised for both English and Spanish language tweets, but due to the time restraints, here we only tackle the problem for English. However, given sufficient time, the methods described in this paper could also be applied to Spanish language as well.

Data Partitioning While testing is carried out on a held-out dataset that is not publicly available, the training data of 4120 twitter profiles was publicly released, and is available in xml format. The classes here are balanced, which means that half of the profiles belong to bots, while the other half belong to human twitter users. Conversely, half of the human authors are female, and the other half are male. For our experiments we partition this data into training and validation sets, using a randomly selected 67% of the data for training purposes, and 33% for meta-parameter optimization, as well as for validating our trained models.

2 Methods

Motivated by the positive results of classical machine learning approaches mentioned in Section 1.1, we explore how these methods would fit the task at hand. In our final submission we rely only on our best performing model (i.e. AdaBoost), however we find it important for future research in the topic to share our experiments with other methods as well. Hence, in this section we discuss three widely-used methods, namely AdaBoost, Random Forest, and Recurrent Neural Networks.

2.1 AdaBoost

Boosting [25,24] is a popular family of algorithms for ensemble learning. The main idea behind these algorithms is to combine several "weak learners" (i.e. classifiers that may perform poorly, but still perform better than random guessing) into a "strong learner", or in other words, a robust classifier. Here, we used one early, successful boosting algorithm published by Freund and Schapire [12]. AdaBoost builds its strong learner on top of the weak learners by weighting each classifier according to its performance. To compute such weights, weak classifiers are trained on the training set, allowing to calculate the probability of error. Each classifier is weighted according to such probabilities and included in the AdaBoost model.

2.2 Random Forest

Random Forest [2] is a supervised machine learning classifier where bootstrapping method is used to partition features into multiple training subsets. It trains individual decision trees for each training subset in the training data. The final classification is given by collecting decisions from all the trees and choosing the final class having maximum scores. The scoring can be done by assigning equal votes to the final decisions of all trees or using a weighted strategy that can be adopted to assign unequal weights to the final decisions of the resulting trees.

2.3 Recurrent Neural Networks

Deep learning attempts to model high-level abstractions in data. Here, we deploy a popular deep learning architecture, namely Recurrent Neural Networks (RNNs). RNNs are particularly suited for tasks where the output is not just dependent on the present input, but also on past input several time steps removed. The contextual meaning within a tweet and the order of tweets carry some extra information prompting the need to employ methods that have the potential to exploit these dependencies. As these dependencies may be long term (spanning up to a hundred tweets), a vanilla RNN may face the issue of vanishing gradient. Because of this, we use the Long Short Term Memory (LSTM) variant in our work to counter this problem.

3 Features

Based on the results of preliminary experiments using neural networks, our focus was on combining classical machine learning algorithms with carefully engineered features. Here, the same set of features are employed for the bot detection and the gender prediction tasks. We calculated most of these features for each tweet independently, then averaged them over a profile. When the computations were carried out differently, we state this explicitly. During our experiments, we noticed that some features share the same value for all Twitter profiles. Later on, these features were dropped. Lastly, after feature extraction we scaled our final set of features using scikit-learn's StandardScaler [19].

3.1 URL Features

We experimented with several features based on the URLs present in tweets, particularly domain-based features (e.g. the ratio of the most commonly linked domains, the ratio of links leading to twitter, the ratio of the most commonly linked twitter profiles). However, as the majority of URLs present in the tweets were first processed by link shortening services, this required Internet access, which is not available in the TIRA virtual machine [20,21]. Hence in the final feature set we confine ourselves to the use of the average number of URLs present in a twitter profile.

3.2 Emoticon Features

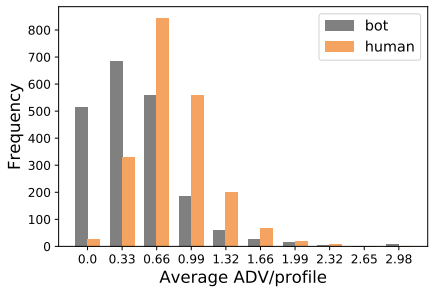
Another feature used in our experiments is the number of emoticons (or emojis) present in each tweet. For the extraction of this feature we use the freely available emoji for Python project [16]. Following the work of Zhenpeng et al. [4] we have also experimented with the use of more high level features based on the emoji-use of twitter profiles. This includes both the emoji frequency and emoji preference features of the original publication (for more details, see [4]). In our preliminary experiments however, these features did not significantly improve the results of either task. Thus, in our final submission we only use the average emoticon count per tweet in our feature set.

3.3 Stylistic Features

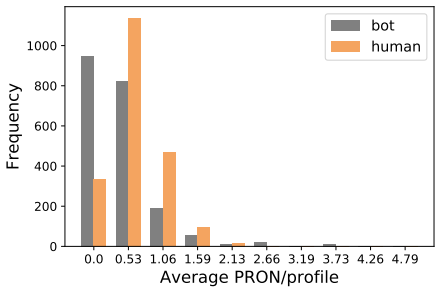
For each tweet we detect and count character floodings, capital letters, sentences and tokens. The average number of capital letters per word is also taken into consideration alongside the Flesch reading-ease score (FRES) [11], calculated as follows:

$$\mathrm{FRES}(text) = 206.835 - 1.015 \left(\frac{\#\mathrm{words}(text)}{\#\mathrm{sentences}(text)} \right) - 84.6 \left(\frac{\#\mathrm{syllables}(text)}{\#\mathrm{words}(text)} \right).$$

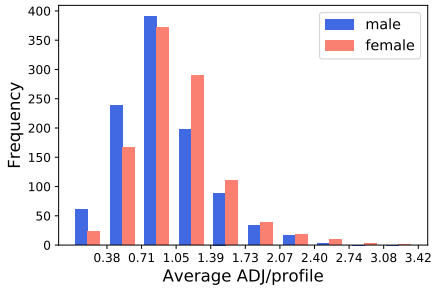
Furthermore, on tweet and profile level, we count the number of tokens that are repeated more than two times and among the repetitive tokens we report the maximum number of repetitions. For example for the following tweet: "*Hairy cats like other cats that are not hairy. However, hairy dogs like cats that are not hairy.*" the tokens that are repeated more than two times are *hairy* and *cats*, so the number of tokens repeated is 2 and the token *hairy* is repeated most times, 4 times. Altogether, we have 10 stylistic features.



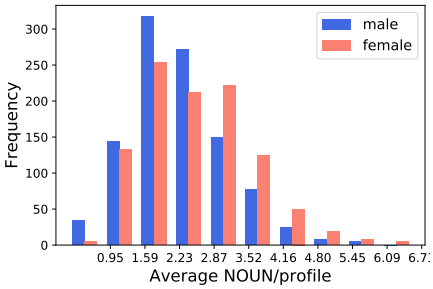
(a) Average number of adverbs (e.g. *very, tomorrow, up, who, there*) used among bot and human profiles



(b) Average number of pronouns (e.g. *I, you, he, myself, themselves, someone*) used among bot and human profiles



(c) Average number of adjectives (e.g. *big, nice, green, last*) among male and female profiles



(d) Average number of nouns (e.g.*girl, dog, book, beauty*) used among male and female profiles

Figure 1: Histograms on the average use of a certain type of POS per twitter profile comparing bots to humans and males to females, respectively

3.4 POS Tags

We count the POS tags for each tweet using spaCy's POS tagger [15] including a total number of 19 POS tags. Indeed, the average number of POS tags per profile could be important – Figures 1a and 1b illustrate that humans tend to use more pronouns and adverbs than bots in their tweets. Furthermore, as Figures 1c and 1d indicate, females on average include more adjectives and nouns in their tweets than males do.

3.5 Topic Features

Our motivation is to explore the semantic topics and categories an author tends to tweet about. For this reason, we employ the SEMCAT [26] and the SemCor [17] datasets on lemmatized words. The SEMCAT (SEMantic CATegories) dataset contains more than 6,500 English words grouped under 110 semantic categories describing diverse types of relations. SemCor is a WordNet-annotated corpus that captures, among others, semantic category annotations for verbs and nouns. We use the SemCor dataset constructed

category	sample words
car	auto buggy car hybrid jeep limo
clothes	apparel bikini fashion fur jeans ring
family	children engaged engagement family love wife
food	breakfast carbohydrate chocolate cook hungry restaurant
money	atm bank currency euro investor withdraw
weather	biosphere cyclone degree humidity meteorology unstable

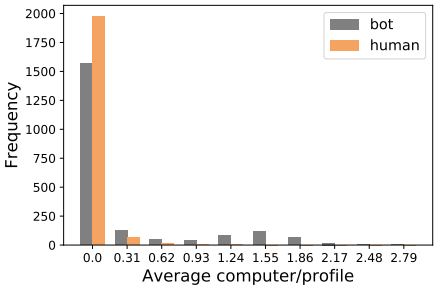
(a) SEMCAT

category	sample words
animal	cow dog eggs fur horn tail
body	artery bathe neck nucleus relax shave
communication	counsel debate description horn interview session
food	beer honey lamb leg produce ration
location	aegean area baltimore china location neighborhood
time	0 acceleration calendar future youth yr

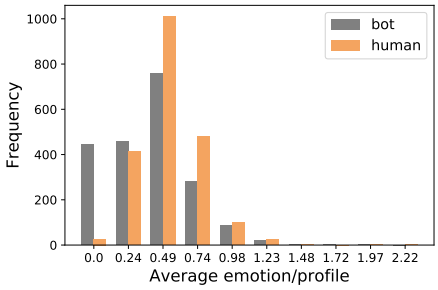
(b) SemCor

Table 1: Representative categories and their 5 sample words from two datasets

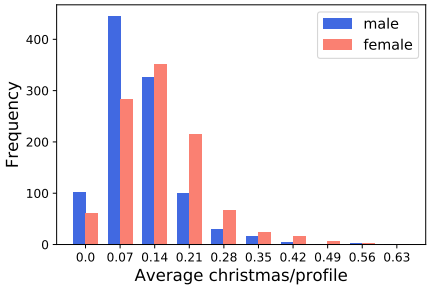
by Tsvetkov *et al.* [28], where words appearing less than 5 times are omitted. This leaves us with more than 4,000 words and 41 categories. Table 1 shows representative SEMCAT and SemCor categories and their sample words. The categories (and their words) are not differentiated based on their source datasets, which means that we work with a total number of 133 topic features. As illustrated in Figure 2a, there are more bot profiles that use a lot of *computer* related words on average, whereas, as Figure 2b shows, humans tend to tweet more about *emotion*s. By comparing males with females,



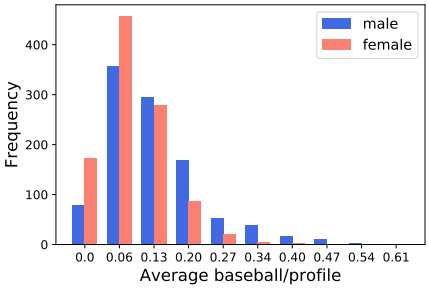
(a) Average use of *computer* related words for bot and human profiles



(b) Average use of *emotion* related words for bot and human profiles



(c) Average use of *christmas* related words for male and female profiles



(d) Average use of *baseball* related words for male and female profiles

Figure 2: Histograms on the average number of words related to a certain type of semantic category per twitter profile comparing bots with humans and males with females

Figure 2c indicates that females describe more *christmas* related words in their tweets, while males tweet more about *baseball*, as shown in Figure 2d.

4 Results and Discussion

After concatenating all features, each twitter profile was described by a 159 dimensional feture vector. As discussed in Section 1.2, to carry out our experiments we first split the dataset into a train and validation set in a 2:1 ratio. We thus created a train set with 2760 examples, and a validation set with 1360 examples. We split these data sets further, to create separate training and validation sets for the two sub-tasks, namely bot detection (a two-class classification task with bot and human labels) and gender classification (a two-class classification task with male and female labels). Lastly, we combined the two models to perform a three class classification task with bot, male, and female labels. In this section we discuss experimental results in this order. First, the results of the bot vs human classification task are discussed. This is followed by the discussion of the results on the gender classification task, and the results of the three class classification task. Lastly, we conclude this section by presenting the results we attained on the held out official test set. It should also be noted that the results reported here as well as the cod for our experiments are available on github[1].

4.1 Bot vs Human Classification

We benchmarked the Bot vs Human Classification task using six popular classification methods. The resulting precision, recall, and F1 scores are listed in Table 2. As can be seen in Table 2, classical machine learning algorithms performed much better than Bi-directional LSTMs. Furthermore, all ensemble methods – random forest, AdaBoost, bagging classifier, gradient boost classifier – resulted in higher scores than those attained using SVMs. Table 2 also shows that the best performance was achieved when using one of the two boosting methods, AdaBoost performing slightly better. For this reason in the remaining tasks our focus was on ensemble tasks, and we did not carry out experiments with LSTMs or Support Vector Machines.

Classifiers	**Precision**	**Recall**	**F1 Score**
Random Forest	97	97	97
AdaBoost	**99**	**99**	**99**
Bagging Classifier	97	97	97
Gradient Boost Classifier	98	98	98
Support Vector Machines	94	94	94
Bi-directional LSTM	-	-	83

Table 2: Precision, Recall and F1-score (in percent) average score on Bot vs Human Classification Task on the validation set using various classification methods.

[1] https://github.com/purvanshi/Gender-and-bot-detection

Class	Precision	Recall	F1-score
female	83	84	83
male	81	87	84
weighted average	83	85	83

(a) Random Forest Classifier

Class	Precision	Recall	F1-score
female	88	92	90
male	90	89	89
weighted average	88	91	89

(b) AdaBoost Classifier

Class	Precision	Recall	F1-score
female	84	85	84
male	80	87	83
weighted average	83	86	83

(c) Bagging Classifier

Class	Precision	Recall	F1-score
female	82	85	84
male	83	84	84
weighted average	82	85	84

(d) Gradient Tree Boosting

Table 3: Precision, Recall and F1-score (in percent) of the gender classification task on the validation set using various ensemble methods for classification.

4.2 Gender Classification Task (Male vs Female)

A markedly higher performance resulting from the use of decision tress on the initial bot detection task supported our earlier decision about focusing on classical machine learning algorithms. Thus for later tasks we only carried out experiments using the four ensemble methods that provided higher scores. In these further experiments we first examined the capability of these ensemble methods to differentiate between twitter profiles that belong to male and female users. The resulting precision, recall, and F1 scores are listed in Table 3.

When comparing the resulting scores in Table 3 to those in Table 2 we see that all algorithms result in markedly higher scores when applied for bot detection than when the same algorithms are applied for gender classification. This may suggest that the task of gender classification is more difficult than that of bot detection. It can also signify, however, that the two tasks require a different set of features, or different machine learning methods. Another possible explanation for this phenomenon may be that we have twice as much data available for the task of bot detection than we do for the task of gender classification. A more thorough investigation of this question is for future work, as the present experimental results are not sufficient to provide a definitive answer.

Table 3 also shows that with each classifier we have similar scores – at most 1% F1–score difference – for the male the female class. It can be observed as well that recall scores tend to be slightly higher than precision scores with the exception of AdaBoost where the precision score for the male class is slightly higher than the recall score for the same class. Lastly, we can also notice that while the weighted average of F1-scores is very similar for three of the methods, it is significantly higher for AdaBoost. We also reported higher scores for AdaBoost on the bot detection class as well, the difference here, however is much more pronounced.

Class	Precision	Recall	F1-score
bot	98	95	96
female	83	83	83
male	81	84	82
weighted average	90	89	89

(a) Random Forest Classifier

Class	Precision	Recall	F1-score
bot	100	98	99
female	88	92	90
male	90	88	89
weighted average	94	94	94

(b) AdaBoost Classifier

Class	Precision	Recall	F1-score
bot	99	94	97
female	84	85	84
male	80	86	83
weighted average	90	90	90

(c) Bagging Classifier

Class	Precision	Recall	F1-score
bot	98	97	98
female	82	84	83
male	83	83	83
weighted average	90	90	90

(d) Gradient Tree Boosting

Table 4: Classification results on Three class classification task.

4.3 Three Class Classification Task (Bot vs Male vs Female)

As a final experiment on the validation set, we evaluated the performance of decision tree classifiers on the three class classification task (bot vs male vs female). The resulting scores are listed in Table 4, which indicates that for each classifier the bot class has significantly higher scores – above 90%, while the male and the female classes have scores around 80–85% may indicate male vs female classification being more difficult than the bot detection task. The resulting scores in Table 4 also show that AdaBoost can attain a markedly higher performance than the other three decision tree based classification methods we used in our experiments.

4.4 Discussion

In all three experiments, we found the F1 scores provided by the AdaBoost Classifier to be the highest (producing +3% higher scores on average than the average score gotten using the other decision tree based classifiers). Another interesting observation is the similar performance of the other three decision tree based methods used which we suspect may be an indication that no one feature is generally better than the other. We have also found that deep learning based methods (bidirectional LSTMs, in particular) did not perform well on the task. This might be due to the limited amount of data available. This issue was accentuated by the restrictions of the competition that limited the use of extra data for the competition, which prevents the use of transfer learning that may alleviate the problem of data scarcity.

4.5 TIRA Evaluation

Lastly, we evaluated our best performing method (AdaBoost) on the official test set of the competition. Given that according to the regulations of the competition, the results of only one (the last) run were to be shared by the organisers, here we used AdaBoost only (as in our preliminary experiments on the development set it was the best performing method).

Task	Validation	Test
Bot detection	99.04%	89.17%
Gender classification	93.75%	35.87%

Table 5: Accuracy scores got using AdaBoost for the bot detection and gender classification task, using our development set, and the official test set

The resulting accuracy scores are listed in Table 5, which indicates there is a marked drop in performance from the validation set to the test set. This drop in performance is less pronounced on the task of bot detection, as the performance of AdaBoost on the Test set is still close to 90%. One possible explanation for this can be if the bots in the two sets were of different domain. In Section 3.5 for example we discuss the prevalence of computer related topic words in the tweets of bot profiles, this however may be due to the overrepresentation of bots in the training set that advertised positions in the IT industry. The drop is much more striking in the case of gender classification. We should note here, however that due to an error in the process of generating output (the algorithm mistakenly outputs a male or female label for the gender task, even if it identified the profile as a bot bofore), our ceiling here is only 50%, and thus we do not think this score is representative of the generalisation capabilities of our model. Overall, we can say however that as it pertains to the generalisation ability of our model, there is much room for improvement still.

5 Conclusions and Future Work

In this paper we proposed an efficient way to extract semantic and syntactic features from twitter profiles. For this we take use of the URLs, emoticons, tokens, and capital letters used in the tweets as different features. The syntactic features were extracted using POS tags. We used semantic categories employing the SEMCAT and semcor datasets which altogether capture 133 categories. We present the results on binary (human - bot, male - female) and multi label (bot, male, female) classification tasks using various machine learning and deep learning techniques. The use of languages in tweets could be analyzed or can be used as another feature. In this work we used the same features for bot and gender detection, although different semantic features could be used. The topic modelling task could also be combined with the emotions used in the tweets.

6 Acknowledgements

This work was supported by the National Research, Development and Innovation Office of Hungary through the Artificial Intelligence National Excellence Program (grant no.: 2018-1.2.1-NKP-2018-00008). Furthermore this research was also supported by the project "Integrated program for training new generation of scientists in the fields of computer science", no EFOP-3.6.3-VEKOP-16-2017-0002. The project has been supported by the European Union and co-funded by the European Social Fund.

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of Economic Perspectives 31(2), 211–236 (2017)
2. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (Oct 2001), https://doi.org/10.1023/A:1010933404324
3. Cai, C., Li, L., Zengi, D.: Behavior enhanced deep bot detection in social media. In: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI). pp. 128–130. IEEE (2017)
4. Chen, Z., Lu, X., Ai, W., Li, H., Mei, Q., Liu, X.: Through a gender lens: Learning usage patterns of emojis from large-scale android users. In: Proceedings of the 2018 World Wide Web Conference. pp. 763–772. WWW '18 (2018)
5. Chou, W.y.S., Hunt, Y.M., Beckjord, E.B., Moser, R.P., Hesse, B.W.: Social media use in the United States: Implications for health communication. J Med Internet Res 11(4) (Nov 2009)
6. Daelemans, W., Kestemont, M., Manjavancas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., Stein, B., Tschuggnall, M., Wiegmann, M., Zangerle, E.: Overview of PAN 2019: Author Profiling, Celebrity Profiling, Cross-domain Authorship Attribution and Style Change Detection. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D., Heinatz, G., Cappellato, L., Ferro, N. (eds.) Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Springer (Sep 2019)
7. Daneshvar, S., Inkpen, D.: Gender Identification in Twitter using N-grams and LSA: Notebook for PAN at CLEF 2018. In: CEUR Workshop Proceedings. vol. 2125 (2018)
8. Dickerson, J.P., Kagan, V., Subrahmanian, V.S.: 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. pp. 620–627. ASONAM '14, IEEE Press, Piscataway, NJ, USA (2014)
9. Ferrara, E.: Disinformation and social bot operations in the run up to the 2017 french presidential election. First Monday 22 (06 2017)
10. Ferrara, E., Varol, O., Menczer, F., Flammini, A.: Detection of promoted social media campaigns. In: tenth international AAAI conference on web and social media (2016)
11. Flesch, R.: A new readability yardstick. Journal of Applied Psychology 32(3), 221–233 (1948)
12. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55(1), 119–139 (Aug 1997), http://dx.doi.org/10.1006/jcss.1997.1504
13. Hall, A., Terveen, L., Halfaker, A.: Bot detection in wikidata using behavioral and other informal cues. Proc. ACM Hum.-Comput. Interact. 2(CSCW), 64:1–64:18 (Nov 2018)
14. el Hjouji, Z., Hunter, D.S., des Mesnards, N.G., Zaman, T.: The impact of bots on opinions in social networks. CoRR abs/1810.12398 (2018), http://arxiv.org/abs/1810.12398
15. Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear (2017)
16. Kim, T., Wurster, K.: Emoji for python. https://pypi.org/project/emoji/ (2019)
17. Miller, G.A., Leacock, C., Tengi, R., Bunker, R.T.: A semantic concordance. In: Proceedings of the Workshop on Human Language Technology. pp. 303–308. HLT '93, Association for Computational Linguistics, Stroudsburg, PA, USA (1993)
18. N. Howard, P., Kollanyi, B.: Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU referendum. SSRN Electronic Journal (06 2016)
19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

20. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the reproducibility of PAN's shared tasks:. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation. Multilinguality, Multimodality, and Interaction. pp. 268–299. Springer International Publishing (2014)
21. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF. Springer (2019)
22. Rangel, F., Rosso, P., Franco, M.: A low dimensionality representation for language variety identification. In: Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing '16), Springer-Verlag, LNCS(9624). pp. 156–169 (2018)
23. 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 (2019)
24. Rokach, L.: Ensemble-based classifiers. Artificial Intelligence Review 33(1), 1–39 (Feb 2010)
25. Schapire, R.E.: The strength of weak learnability. Mach. Learn. 5(2), 197–227 (Jul 1990), https://doi.org/10.1023/A:1022648800760
26. Senel, L.K., Utlu, I., Yücesoy, V., Koç, A., Çukur, T.: Semantic structure and interpretability of word embeddings. CoRR abs/1711.00331 (2017), http://arxiv.org/abs/1711.00331
27. Taulé, M., Martí, M.A., Pardo, F.M.R., Rosso, P., Bosco, C., Patti, V.: Overview of the task on stance and gender detection in tweets on catalan independence. In: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017), Murcia, Spain, September 19, 2017. pp. 157–177 (2017)
28. Tsvetkov, Y., Faruqui, M., Ling, W., Lample, G., Dyer, C.: Evaluation of word vector representations by subspace alignment. In: Proc. of EMNLP. pp. 2049–2054 (2015)
29. Vinayakumar, R., Kumar, S.S., Premjith, B., Poornachandran, P., Padannayil, S.K.: Deep stance and gender detection in tweets on catalan independence@ibereval 2017. In: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages. pp. 222–229 (09 2017)
30. Yang, K.C., Varol, O., Davis, C., Ferrara, E., Flammini, A., Menczer, F.: Arming the public with Artificial Intelligence to counter social bots. Human Behavior and Emerging Technologies p. e115 (02 2019)