Using Hashtags and POS-Tags for Author Profiling

Notebook for PAN at CLEF 2019

Flurin Gishamer

ZHAW Zurich University of Applied Sciences
gishamer@pm.me

Abstract This paper investigates automatic methods to separate human created from bot created Tweets, and in the case of human, determine the gender of an author. The novel contribution is the investigation of 2 research questions, firstly the usability of part of speech tags and secondly the usability of hashtags as additional features. It therefore extends the models presented by Daneshvar et al. and Basile et al. in the course of past Author Profiling Tasks @ Pan. The results are evaluated as part of the Author Profiling Task @ Pan 2019. It will be shown that the segmentation of hashtags as well as using POS-Tags n-grams can increase the accuracy when classifying bot and gender on the PAN Twitter-dataset. By adding these features and combining them in an ensemble classifier, it was possible to achieve accuracies of 94% for bots and 84% for gender for the English language on the official test set. However, with 79% for bots and 71% for gender, the performance on the Spanish part of the dataset differs significantly. Possible reasons for this shall be examined in the evaluation of the system.

CEUR-WS.org/Vol-2380/paper_171.pdf

1 Introduction

The ubiquity of social media, in private communication and media coverage calls for strategies to validate both identity of users as well as the validity of the shared content to prevent misuse and manipulation of public opinion.

In [1] Shao et al. state that the deliberate spreading of false information, so-called fake news, is a serious concern. Guess, Nagler and Tucker, who conducted a representative online survey on Facebook users behavior in connection with fake news, say that "The vast majority of Facebook users in our data did not share any articles from fake news domains" [2]. If one compares this with the finding from Chu, Gianvecchio, Wang and Jajodia in [3] that 24% of the Tweets generated on Twitter originate from bots, and relates it with the statement of Shao et al. that "social bots played a disproportionate role in spreading articles from low-credibility sources" [1], it can be concluded that the identification of bot profiles on social media is an important and promising approach to prevent the spreading of fake news, and thus the manipulation of public opinion.

The present work deals with the identification of features, which are suitable to improve the accuracy of existing methods. The focus lies on the identification of bots as well as the identification of gender from authors on Twitter. Both POS-tags as well as the information contained in hashtags are considered.

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.

1.1 Research Questions

The present work examines two central questions, namely:

1. "Does the syntactic structure of Twitter Tweets reveal information about the author's identity with respect to gender/bot, moreover, if so, are such patterns universal, i.e. are these patterns independent of content ?"
   Part of speech tags were chosen to represent the syntactic structure. To consider the sequential nature of the data, POS tags bi- and tri-grams were used as features..

2. "Do hashtags in Twitter Tweets contain information about the identity of the author and can this information improve the accuracy of gender/bot-classification when looking at the individual words they comprise?"
   Due to the special nature of hashtags, where users are forced to use a single word, therefore using compound words, the approach of segmenting hashtags and using the resulting words as features was chosen.

The goal is to develop a model that can classify Tweets via the use of an enriched body of features, analyzing them in parallel and combining them.

1.2 Author Profiling Task @ PAN 2019

In [4] author profiling is described as: "the analysis of shared content in order to predict different attributes of authors such as gender, age, personality, native language, or political orientation."

The task described by Rangel et al. in [5] is concerned with the identification of an authors gender and additionally if the author is either human or bot.

PAN is a series of workshops concerned with digital text forensics [6] and is carried out as part of the CLEF conference which is concerned with the systematic evaluation of information access systems.

The common basis of all participants is a dataset containing 100 Tweets per author, which are combined in one file per author with a corresponding label assigned to it. A label can have the following values: bot/human, and in the case of humans, female/male.

The dataset of the Author Profiling Task 2019 includes the languages English and Spanish. The English dataset contains 4120 authors, of which 2060 are bots, and the remaining 2060 are divided into 1300 female and 1300 male authors. The Spanish dataset contains 3000 authors, of which 1500 are bots, and the remaining 1500 are divided into 750 female and 750 male authors.

Consequently, the gender of the author or whether the author is human should be inferred based on a set of short messages which are available in purely textual form (without meta information or additional content such as images). The evaluation of all submitted systems is carried out on the online platform Tira [7].

To obtain the final scores, the results of all participants are ranked by accuracy. A detailed description as well as results and comparisons of systems submitted to the Author Profiling Task @ Pan 2019 can be found in the official overview paper [5].

2 Related Work

A data set concerned with bot-detection is the honeypot data set. It was introduced by Morstatter et al. in [8], and as the name suggests was created using so-called honeypot bots. They say that honeypot bots such that "... any user in the network that connects to a honeypot will be considered as a bot" [8]. To identify bots, they developed an extension of the AdaBoost algorithm which they call BoostOr, and used the honeypot dataset to evaluate it. According to Morstatter et al. BoostOr "focuses more on the mislabeled bots and downweights mis-labeled regular users." [8].

In [9] Cai, Li and Zengi Introduce their Behaviour enhanced deep bot detection model. It is an artificial neural network architecture which uses an LSTM, to learn a representation of the sequence of an authors Twitter history. They evaluated their model on the honeypot dataset presented in [8], and report in [9] that this model, called BeDM reaches an F1 score of 87.32% as opposed to the BoostOr model presented in [8] which, according to Cai et al. reaches an F1 score of 86.10%

In [10] Basile et al. presented a model with a "simple SVM system (using the scikit-learn LinearSVM implementation) that uses character 3- to 5-grams and word 1- to 2-grams with tf-idf weighting" through which they achieved the best result in the Author Profiling Task @ Pan 2017. In the following year, several of the best-ranked systems (when only considering textual features) employed similar strategies with respect to n-grams and the classification algorithm used.

According to [4] the best result in the combined Author Profiling Task @ Pan 2018 was achieved by Takahashi et al. [11] Their text component consists of a bi-directional recurrent neural network whose output leads via two successive pooling layers into a fully connected layer. As features, they used word vectors. In this system, however, the result of the textual features is supplemented with information from images, which are analyzed using a convolutional neural network. The achieved accuracy averaged over all three languages was according to [4] 78.72%.

Rangel et al. state in [4] that the system of Daneshvar and Inkpen was able to achieve the best results when only textual features were considered. The features were similar to [10], with the addition of word 3-grams for the English part of the dataset and subsequent Latent Semantic Analysis for all languages. The classification algorithm used was a support vector machine [12].

According to [4] the accuracy averaged over all three languages for Daneshvar's model was 81.70%. This result is noteworthy as it shows that the best score of the 2018 task was achieved without the use of the provided images.

According to [4] Tellez et al. achieved the second-best result when considering only textual features, with a value of 80.99%. Similar to [10] and [12], the Bag of Words approach was chosen using the tf-Idf weighting scheme, and support vector machines for classification. Note, however, the additional use of skip grams [13].

In [14] Reuter, Pereira-Martins and Kalita present a pipeline to segment Hashtags. It is a combination of several approaches, where the use of maximum known matching seems to be worth mentioning, which tries to determine a metric for the length of matches, and the result which delivers the longest match is rated highest.

In [15] Declerck and Lendvai mention that Spanish sources contain fewer hashtags than German ones, and that the camelCase notation is mainly used in English sources.

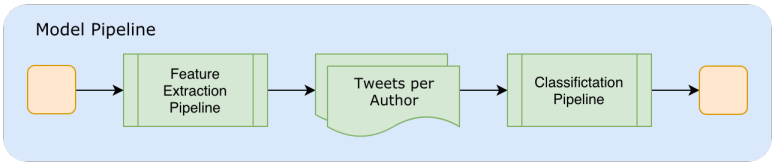

Figure 1. Components of the Pipeline

Their approach segments hashtags written in camelCase notation in a first step, and then uses them as a decision basis for segmenting hashtags written in lower case letters.

In [16] and [4] it is stated that participants either normalized Tweets by removing hashtags altogether, or used ratios of hashtags with respect to Tweets.

To the author's best knowledge, the approach of replacing hashtags with words extracted by segmentation has not yet been used in a model submitted to the Pan workshop.

The use of POS tags as features in the form of n-grams has already been discussed by Martinc et al. [17] in the Author Profiling Task @ Pan 2017, but only trigrams were considered here. Furthermore, a single instance of a Logistic Regression Classifier was employed for classification, to which a combination of differently weighted features was provided.

In [18] López-Santillán, Gonzalez-Gurrola and Ramfrez-Alonso introduce a model in which they create embeddings of POS tags, using the same procedure as is used for word embeddings. Here they chose the skip-gram approach. In addition to the word embeddings the obtained document vector is then enriched with these POS-embeddings.

The LDSE baseline by Rangel, Rosso and Franco is described under [19] and is concerned with the Task of Language variety identification.

3 Model Overview

The model proposed in this paper is based on and extends the model presented by Daneshvar et al. in [12] and Basile et al. in [10]. It comprises two main components, of which the first is the preprocessing pipeline, which is responsible for tokenization and preprocessing of the Tweets.

As is shown in Figure 1, the classification pipeline consists of a text and a POS-part, of which the text component is similar to the implementation in [12] [10] [13], with the addition of hashtag segmentation and handling of compound emojis. The second component is concerned with the classification of POS-tags.

The text and POS-part are combined in an ensemble classifier, which uses a support vector machine as meta classifier.

The resulting output is a prediction which is either bot/female/male, written to an XML file per author.

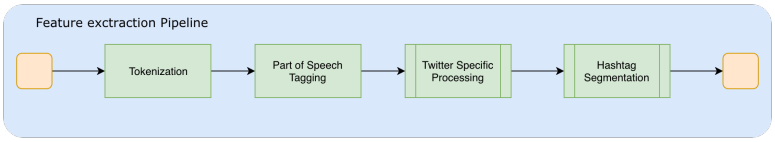

Figure 2. Feature Extraction Pipeline

4 Methodology

This section provides a detailed discussion of the proposed model concerning the research questions presented in the introduction. It first describes the preprocessing pipeline and its specifics, with a particular focus on the peculiarities of Twitter Tweets, and then the detailed structure of the classification pipeline.

4.1 Feature Extraction

As can be seen in Figure 2 during the preprocessing phase, the concatenated Tweets per author are tokenized and Twitter specific replacements, hashtag-segmentation and part of speech tagging is performed simultaneously.

All of the preprocessing is performed using the spaCy NLP Framework as introduced in [20] by Honnibal and Montani. For the English part, the en_cor_web_sm language model was used, which incorporates a convolutional neural network trained on the OntoNotes corpus, consisting of blog articles, news, and comments [21].

For the Spanish part, the es_core_news_sm language model was used which is trained on the AnCora and WikiNER corpus instead, which comprises news and media content [22]. The Twitter specific functionality was implemented via the extension of custom pipeline objects provided by spaCy.

Twitter Specific Preprocessing

Substitutions
Similar to [12] the first step was to perform several substitutions on the concatenated Tweets per author:

Domain names:	URLURL
End of a tweet:	SNTDLM
E-mail addresses:	EMAILEMAIL
Twitter handles:	USERMENTION
Line breaks:	LNFD

To obtain accurate part of speech tags (POS tags), the replacement SNTDLM was used to indicate the end of a sentence to the tagger explicitly. As with [12], sequences of the same letters, which occurred more than three times, were replaced with a sequence of 3 letters, resulting in a replacement of the following form:

heeey, heeeeeey, heeeeeeey $\rightarrow$ heeey

Custom POS tags
The jargon used in social media has some constructs that do not occur in verbal communication or classical texts. In order to consider this, the POS tags have been enhanced with the following elements:

Domain names:	URL
Emojis:	EMJI
E-mail addresses:	EML
Hashtags:	HSHT
Twitter Handles:	HNDL

Emojis
Emojis can be modified in several ways, e.g. there is a skin-tone modifier which can be used to change the skin color of Emojis. Additionally, the combination of several Emojis is possible, e.g. in the family Emoji 👨‍👩‍👧, which consist of 👨, 👩 and 👧: Combining several emojis is generally achieved by creating a sequence with the so-called Zero-Width-Joiner (ZWJ). When using a whitespace tokenizer, this is problematic in several ways: firstly during tokenization, such sequences are cut at the ZWJ, and secondly, the ZWJ remains in the resulting token stream. Therefore the tokenizer was adapted to recognize compound emojis and treat them as one token. For the POS tagger this means that regardless of the length of a sequence of emojis, the result is always one POS-tag.

Hashtag Segmentation
To determine whether hashtags contain information about the identity of an author, when classifying bot/female/male, the procedure of segmenting composite hashtags into individual words was chosen, resulting in replacements of the form:

#makeamericagreatagain	→ make america great again
#roomforrent	→ room for rent

If the hashtag consist of a single word, a wordlist lookup is first performed to avoid divisions such as the following:

#iconic	→ i conic
#handsome	→ hand some

The Viterbi algorithm who was first presented by Viterbi in [23] and more specific an adaptation of it by Bacon [24] was used to segment composite hashtags into individual words. In [25] it is described as follows: "the VA may be viewed as a solution to the problem of maximum a posteriori probability (MAP) estimation of the state sequence of a finite-state discrete-time Markov process". To calculate the probability of the word under consideration, the algorithm needs to access word frequency lists. During the development of the model, such lists were generated from the Pan Dataset, but it was found that word frequency lists based on the OpenSubtitles corpus by Lison and Tiedemann [26] gave superior results. Hence the final model uses them instead.

In the actual algorithm a test is performed first, if the length of a hashtag is less than 3 characters, or it is contained within the provided wordlist, the word without the pound character is returned.

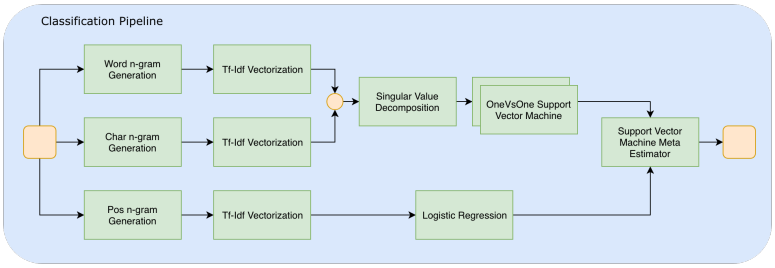

Figure 3. Classification Pipeline

Then a nested loop is executed which steps through the string, considering each substring contained in the hashtag, and using a function to assign it a probability.

The mentioned function takes a word as argument and returns its probability, which is calculated by dividing its frequency by the total number of word occurrences within the provided word-frequency list (this information is contained data variable). The words with the highest probability found are then returned.

4.2 Classification

The classification pipeline shown in Figure 3 consists of an ensemble that combines the results of the text and POS components using an SVM meta-learner to make the final prediction. The ensemble was implemented with the ML-Ensemble framework developed by Flennerhag which facilitates parallel computations [27]. For the single components such as the tf-idf vectorizer, singular value decomposition or the linear svm classifier the sci-kit learn framework was used.

Experiments were conducted with both, a soft- and hard-voting approach, it was found that the ensemble achieves the best results using a soft-voting approach.

N-Grams

As proposed in [10] word 1- to 2-grams in addition to character 3- to 5-grams were used in the text-component. As suggested in [12], for English, character 3-grams were also included. For the POS-Pipeline grid-search was performed which indicated that a combination of word 2- and 3-grams are the optimal setting.

Text Component

The text component uses both word n-grams and character n-grams as features. Each of which is transformed separately into tf-Idf vectors, where only tokens with a term frequency greater than or equal to 2 are considered. The set of resulting document vectors is the source material for latent semantic analysis. This part of the pipeline is essentially an extension to the systems presented under [12] [10]. Experiments with logistic regression have been carried out. However, a support vector machine with a linear kernel has proved to be the most effective choice, just like in [12]. In order to enable multi-class classification different strategies were considered out of which the best results were achieved with a One vs. One approach.

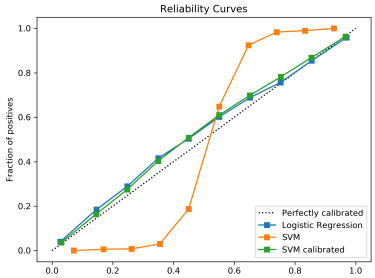

Figure 4. Reliability Curves

POS Component
In the POS component, n-grams were also generated in a first step, but only on the token level (no character n-grams were used). The use of n-grams was chosen to allow the classifier to at least fundamentally analyze the information which lies in the sequence of the data. Inspired by the text component, latent semantic analysis was also experimented with but showed no improvement in accuracy. Interestingly, in contrast to the text component, the accuracy increased when using logistic regression over a support vector machine. Hence the final version uses it.

Probability Calibration
In [28] Platt et al. explain that "Posterior probabilities are also required when a classifier is making a small part of an overall decision, and the classification outputs must be combined for the overall decision." [28]. He continues to point out that support vector machines output an uncalibrated value which is not a probability. The ensemble of the proposed model uses a meta-classifier with a soft-voting approach, which means that it receives as input class-probabilities instead of hard labels. Therefore the SVM-calssifier must be calibrated, as opposed to the logistic regression classifier of the POS-component, which according to Niculescu-Mizil and Caruan [29] already predicts well-calibrated probabilities. The calibration of the SVM was performed over the holdout set of 3 folds; this step was directly included in the training process. In Figure 4 one can see the reliability curve and the effect of calibration on the SVM-classifier.

5 Evaluation

This section presents and evaluates the results that the proposed model was able to achieve on both the training and the test data. Special attention will be paid to the performance of the POS-component and the differences between the Spanish and English parts of the data-set.

5.1 Results on the Training Data

In accordance with [12] 60% of the PAN training data was used to train the models, and 40 % was used to evaluate them. In addition, 10-fold cross-validation was employed during training. The following models were evaluated on the training data:

Table 1. Model Comparison on Training Data

Model	Features	En	Es
Text Component	N-Grams	0.926	0.902
Text Component	N-Grams, Hashtags	0.944	0.914
Ensemble	N-Grams, Hashtags, POS-tags	0.950	0.915

Table 2. Final Model on Training Data

	En	Es
Bot	1.00	0.99
Gender	0.89	0.84

Table 3. POS-Component on Training Data

	Precision		Recall		F1-Score	
	En	Es	En	Es	En	Es
Bot	0.96	0.94	0.96	0.93	0.96	0.93
Gender	0.70	0.68	0.71	0.69	0.71	0.69

Table 4. Final Model on Test Data

	En	Es
Bot	0.935	0.792
Gender	0.840	0.712

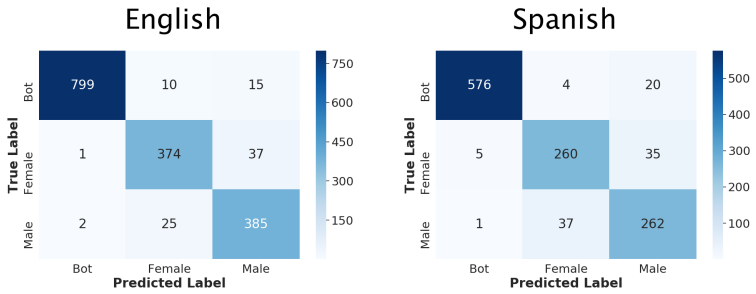

Figure 5. Confusion Matrix

1. Text component without hashtag-segmentation.
2. Text component with hashtag-segmentation.
3. Ensemble with text-component and pos-component

Looking at the results in Table 1, which lists the mean accuracies for each examined model for English and Spanish, it is noticeable that although a wide range of features has been considered, the differences are all below 2%. Nevertheless, the combination of hashtag-segmentation and POS classification leads to a measurable improvement in the overall result.

Both hashtags and POS n-grams have a higher influence on the English part of the data-set. The hashtags in the English part improve the accuracy from 92.6% to 94.4%, whereas the accuracy in the Spanish part increases from 90.2% to 91.4%. The differences for the POS component are even more pronounced. Looking at the differences when adding the POS component, it can be seen that the increase in precision in the English part is 0.5%, whereas in the Spanish part it is only 0.1%. Table 2 shows the results with respect to precision for the final model on the training data by language and class, here it shows, that the difference between class bot and gender for the Spanish part is 5% higher than for the English part:

Considered in isolation, with a mean accuracy of 81.3 % for Spanish, the POS component has a significantly lower accuracy than the text component, but it improves the overall result when used in an ensemble. The results for precision, recall, and F1-Score of the POS-component can be seen in Table 3

Since a One vs. One approach was chosen to implement the multiclass-classification, each of the three classes bot/female/male has its own instance per component. The confusion matrix in Figure 5 now shows that in both languages the number of bots wrongly classified as men was much higher than the number of bots classified as women. In English 10 women compared to 15 men, and in Spanish even 4 women compared to 20 men.

5.2 Results on the Test Data

In order to evaluate the proposed model on the official test data set, it was first trained on the entire training data and in a second step evaluated on the test data. The results are listed in Table 4:

What is particularly noteworthy about the results on the test data are the significant differences between English and Spanish. The fact that the model achieved better results on the English data-set could already be observed during training but was amplified on the test data-set. Where on the training data-set the accuracy for Spanish was 84.5% for the gender classification task compared to 89.5% for English, on the test data-set, it became 71.2% for Spanish compared to 84% for English.

6 Conclusion

1. **Syntactic structure:** It has been shown that it is possible to use POS-tags to classify Tweets based on their syntactic structure, which means classification is possible without any information about the actual content of a text. In addition, it was determined during the evaluation that these features are suitable to improve the accuracy of a system which until now has only classified on the basis of words.
2. **Hashtags:** The results presented as part of the evaluation show that it is possible to improve accuracy when classifying, by segmenting the hashtags contained in the Tweets into individual tokens/words and replacing them with the original hashtag. It has also been shown that the approach presented, using word-frequency lists and the Viterbi algorithm to perform this segmentation is feasible.

However, it was not possible to determine what caused the large differences in accuracy between Spanish and English. A possible explanation for this are the different corpora used to train the Tokenizer/POS taggers in English and Spanish, and the respective word-frequency lists. Although López-Santillán et al. do not use tf-idf vectors, but embeddings of POS tags as features [18], it is interesting that they also report lower accuracies for Spanish than English.

7 Outlook

It would be interesting to investigate to what extent longer sequences enable an improvement in accuracy using an algorithm that is able to better address the relationship between the individual elements. The use of LSTM or GRU networks would be

conceivable here. This would be a further step towards a model that can classify text independent of content.

In the proposed model, the tokens obtained by segmenting the hashtags were treated exactly the same as other tokens. It should be examined whether a further improvement in accuracy could be achieved through a different weighting scheme of the tokens obtained from the hashtag segmentation.

8 Acknowledgements

I would like to express my very great appreciation to Prof. Dr. Martin Braschler for his valuable and constructive contribution to the planning and development of this paper. His feedback as a supervisor has always been of great value to me.

I would also like to thank Saman Daneshvar for providing the source code of his model, which allowed me to focus on my research questions.

References

1. Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.C., Flammini, A., Menczer, F.: The spread of low-credibility content by social bots. Nature communications **9**(1) (2018) 4787
2. Guess, A., Nagler, J., Tucker, J.: Less than you think: Prevalence and predictors of fake news dissemination on facebook. Science advances **5**(1) (2019) eaau4586
3. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is tweeting on twitter: human, bot, or cyborg? In: Proceedings of the 26th annual computer security applications conference, ACM (2010) 21–30
4. Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. In: CEUR Workshop Proceedings, CEUR, CEUR-WS.org (2018)
5. Rangel, F., Rosso, P., Cappellato, L., Ferro, N., Müller, H., Losada, D.: Overview of the 7th author profiling task at pan 2019: Bots and gender profiling. In Cappellato, L., Ferro, N., Losada, D., Müller, H., eds.: CEUR Workshop Proceedings, CEUR, CEUR-WS.org (September 2019)
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 (September 2019)
7. 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)
8. Morstatter, F., Wu, L., Nazer, T.H., Carley, K.M., Liu, H.: 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), IEEE (2016) 533–540
9. 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), IEEE (2017) 128–130

10. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-gram: New groningen author-profiling model. arXiv preprint arXiv:1707.03764 (2017)
11. Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., Ohkuma, T.: Text and image synergy with feature cross technique for gender identification. Working Notes Papers of the CLEF (2018)
12. Daneshvar, S., Inkpen, D.: Gender identification in twitter using n-grams and lsa. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). (2018)
13. Tellez, E.S., Miranda-Jiménez, S., Moctezuma, D., Graff, M., Salgado, V., Ortiz-Bejar, J.: Gender identification through multi-modal tweet analysis using microtc and bag of visual words. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). (2018)
14. Reuter, J., Pereira-Martins, J., Kalita, J.: Segmenting twitter hashtags. Intl. J. on Natural Lang. Computing **5**(4) (2016)
15. Declerck, T., Lendvai, P.: Processing and normalizing hashtags. In: Proceedings of the International Conference Recent Advances in Natural Language Processing. (2015) 104–109
16. Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. In: Working Notes Papers of the CLEF, CEUR, CEUR-WS.org (2017)
17. Martinc, M., Skrjanec, I., Zupan, K., Pollak, S.: Pan 2017: Author profiling-gender and language variety prediction. In: CLEF (Working Notes). (2017)
18. López-Santillán, R., Gonzalez-Gurrola, L., Ramfrez-Alonso, G.: Custom document embeddings via the centroids method: Gender classification in an author profiling task. In: Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018). (2018)
19. Rangel, F., Franco-Salvador, M., Rosso, P.: 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 (2018) 156–169
20. Honnibal, M., Johnson, M.: An improved non-monotonic transition system for dependency parsing. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, Association for Computational Linguistics (September 2015) 1373–1378
21. Honnibal, M., Montani, I.: English · spacy models documentation https://spacy.io/models/en.
22. Honnibal, M., Montani, I.: Spanish · spacy models documentation https://spacy.io/models/es.
23. Viterbi, A.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE transactions on Information Theory **13**(2) (1967) 260–269
24. Bacon, D.: How can i split multiple joined words? http://stackoverflow.com/a/481773/554406.
25. Forney, G.D.: The viterbi algorithm. Proceedings of the IEEE **61**(3) (1973) 268–278
26. Lison, P., Tiedemann, J.: Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. (2016)
27. Flennerhag, S.: Ml-ensemble http://ml-ensemble.com/.
28. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3) (1999) 61–74
29. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning, ACM (2005) 625–632