Author Profiling Using Semantic and Syntactic Features

György Kovács§‡, Vanda Balogh†, Purvanshi Mehta¤, Kumar Shridhar¤, Pedro Alonso§, and Marcus Liwicki§*

§*Embedded Internet Systems Lab*, Luleå University of Technology, Luleå, Sweden (gyorgy.kovacs@ltu.se, pedro.Alonso@ltu.se, marcus.liwicki@ltu.se)
‡*MTA-SZTE Research Group on Artificial Intelligence*, Szeged, Hungary
†*Institute of Informatics*, University of Szeged, Szeged, Hungary, (bvanda@inf.u-szeged.hu)
¤*MindGarage*, Kaiserslautern, Germany (purvanshi.mehta11@gmail.com, shridhra.stark@gmail.com)

Key contributions

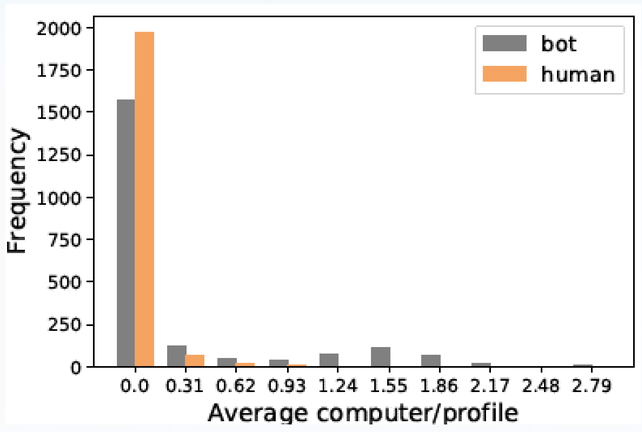
- Examined a wide set of features for both tasks
- Examined a variety of machine learning algorithms for the bot detection class
- Attained 99% accuracy on validation and ~89% on the test set for bot detection

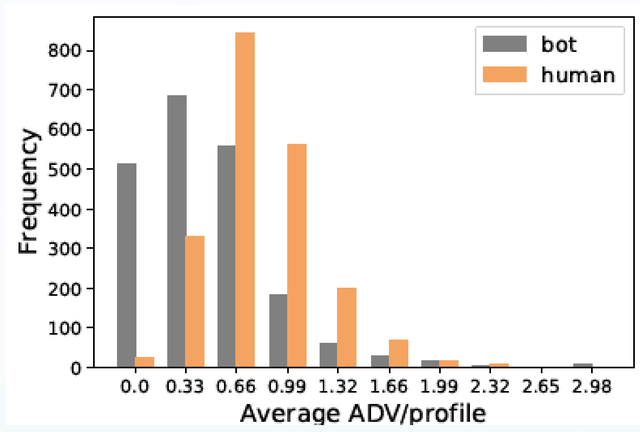
Task

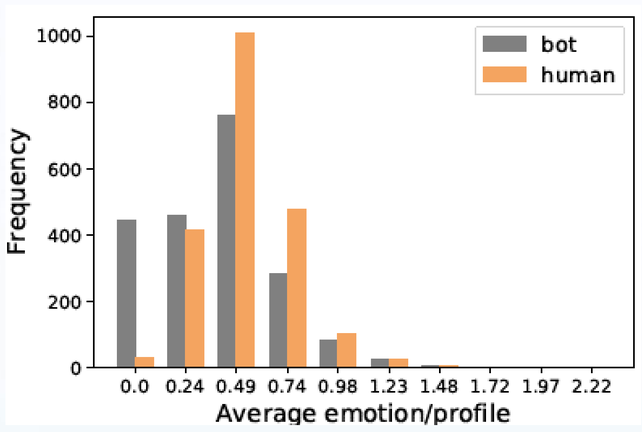
- Bot detection and gender classification
- Classify twitter profiles based on tweets
- English language

Features

- **URL features**: average no. of URLs
- **Emoticons**: average emoticon count
- **Stylistic features**:
  - Character flooooodings
  - Average no. of capital letters per word
  - No. of sentences
  - No. of tokens
  - Flesch reading-ease score
  - Tokens repeated more than twice
  - Maximum repetition count of token
- **POS tags**: average no. of spaCy [1] POS tags per profile
- **Topic features**: Prevalence of words in the profile belonging to SEMCAT [2] and SemCor [3] categories (133 altogether)







Methods [4]

- **AdaBoost (AB)**: combination of several "weak" learners (e.g decision stumps) into a robust classifier
- **Random Forest (RF)**: combination of several decision tress trained on different subsets of the training set
- **Bagging Classifier (BC):** trains base classifiers (k nearest neighbours) on random subsets of the training data, and aggregates their decisions
- **Support Vector Machines (SVMs):** apply hyperplanes to separate the different classes in feature space
- **Recurrent Neural Networks (RNNs)**: class of neural networks where the output from the previous step is used as an input in the current step. Here, a variant is used that is designed to overcome the issue of vanishing gradients: Long Short Term Memory (LSTM)

Results

Bot classification on development set

Classifier	F1-score
AdaBoost	99
RandomForest	97
Bagging Classifier	97
Gradient Boost	98
SVMs	94
BiDirectional LSTM	83

Gender classification on development set

Classifier	F1-score
AdaBoost	89
RandomForest	83
Bagging Classifier	83
Gradient Boost	84

PAN evaluation

Classification Accuracy using AdaBoost

Task	Dev	Test
Bot/Human	99%	89%
Female/Male	94%	36%

Conclusions and future work

- An efficient syntactic and semantic feature extractor is introduced
- Several types of features included in the examination
  - URL, emoticons, tokens, capital letters
  - Syntactic features extracted using POS tags
  - Semantic features extracted using the SEMCAT and SemCor dataset
- Future work
  - Analyze the use of languages in tweets
  - Examine the use of separate feature sets for the two tasks
  - Combination of topic modelling with emotions detected in tweets

[1] Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear (2017)
[2] 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
[3] 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)
[4] 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)

*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