# Towards a language independent Twitter bot detector

Jonas Lundberg[1], Jonas Nordqvist[2], and Mikko Laitinen[3]

[1] Department of Computer Science, Linnaeus University, Växjö, Sweden
[2] Department of Mathematics, Linnaeus University, Växjö, Sweden
[3] School of Humanities, University of Eastern Finland, Joensuu, Finland

**Abstract.** This article describes our work in developing an application that recognizes automatically generated tweets. The objective of this machine learning application is to increase data accuracy in sociolinguistic studies that utilize Twitter by reducing skewed sampling and inaccuracies in linguistic data. Most previous machine learning attempts to exclude bot material have been language dependent since they make use of monolingual Twitter text in their training phase. In this paper, we present a language independent approach which classifies each single tweet to be either autogenerated (AGT) or human-generated (HGT). We define an AGT as a tweet where all or parts of the natural language content is generated automatically by a bot or other type of program. In other words, while AGT/HGT refer to an individual message, the term bot refers to non-personal and automated accounts that post content to online social networks. Our approach classifies a tweet using only metadata that comes with every tweet, and we utilize those metadata parameters that are both language and country independent. The empirical part shows good success rates. Using a bilingual training set of Finnish and Swedish tweets, we correctly classified about 98.2% of all tweets in a test set using a third language (English).



## 1 Introduction

In recent years, big data from various social media applications have turned the web into a user-generated repository of information in ever-increasing number of areas. Because of the relatively easy access to tweets and their metadata, Twitter[4] has become a popular source of data for investigations of a number of phenomena. These include for instance studies of the Arab Spring [1], various political campaigns [2, 3], of Twitter as a tool for emergency communication [4, 5], and using social media data to predict stock market prices [6]. In linguistics, various mono- [7] and multilingual text corpora of tweets [8] have been built recently and used in a wide range of subfields (e.g. dialectology, language variation and change). The problem of establishing a Twitter text corpus for small languages (e.g., Croatian) is discussed in [10].

[4] www.twitter.com

One special characteristic of Twitter and many other social media applications is the presence of bot accounts, i.e. non-personal and automated accounts that post content to online social networks. A bot refers to a heterogeneous set of account types which post tweets automatically. The popularity of Twitter as an instrument in public debate has led to a situation in which it has become an ideal target of spammers and automated programs. It has been estimated that around 5-10% of all users are bots[5], and that these accounts generate about 20-25% of all tweets posted[6]. For research purposes, bots present a serious problem because they reduce data accuracy and may dramatically skew the results of analyses using social media data.

Consequently, bot detection has been discussed in various papers in computer sciences [9, 11–15]. In sociolinguistics, previous studies have relied on a range of methods when dealing with bots. For instance, Huang et al. [16] recognize their presence but include them in the results (also true for Laitinen et al. [8]). Coats [20] utilizes a method in which material from certain types of devices is excluded.

In computer science, the various bot detection approaches typically apply machine learning based on account properties and/or tweet metadata. A typical method is to focus on classifying whether a user account is a bot or not. These attempts tend to make use of historical (timeline) data to compute properties like tweets per day, or statistical measures (e.g. entropy or $\chi^2$-test, [9, 11, 12]) to identify periodic patterns in the tweeting behavior on an account level. Another approach in previous attempts is to use the actual Twitter text keyed in by the author as an input parameter in the classification. This results in the classifier becoming language dependent since the classifiers are trained on a monolingual set of tweets (English in most cases). While these approaches may result in sufficient precision and recall rates, these approaches have two practical problems. (1) The language dependency requires a new classifier (using a new training set) for each new language. (2) As such, the systems cannot easily classify tweets in real-time, as a part of the Twitter downloading stream, since they make use of historical data that must be downloaded in advance. This makes it difficult to integrate such an application into a digital language infrastructure that makes social media data available for researchers in the humanities.

This paper presents a language independent approach for detecting AGTs. This language independency stems from the fact that the actual Twitter text is not used as an input feature in the classifier. In fact, the algorithm classifies each tweet using only select attributes in the metadata which are available for each tweet. This feature not only makes our approach simple and light, but it also makes it possible to classify tweets in real-time as a part of a Twitter downloading system.

[5] www.nbcnews.com/business/1-10-twitter-accounts-fake-say-researchers-2D11655362

[6] sysomos.com/inside-twitter/most-active-twitter-user-data/

| Example tweet | Comment | Class |
|---|---|---|
| I was out walking 8.02 km with #something #somethingelse https://somewhere.com | This tweet is generated by an app and by adding 'I was out walking' it adds natural language to the tweet. | AGT |
| New year perfect photo frame!! #something #somethingelse @location https://somewhere.com | This tweet is generated by an app but not considered an AGT since it does not add any natural language. The natural language was originally produced in the app by the user. | HGT |

**Fig. 1.** Examples of AGTs and non-AGTs.

## 2 Language independent AGT detection

### 2.1 The dataset

The dataset to be used here is collected using the same parameters as in the Nordic Tweet Stream (NTS) corpus [8, 17]. The NTS uses the Twitter Streaming API to collect tweets by specifying a geographical region covering the five Nordic countries. This corpus is a real-time monitor corpus designed for sociolinguistic studies of variability in the Nordic region. Our research carried out using the material has primarily been related to charting the use of English in the area, investigating its grammatical variability, and modelling social networks in multilingual settings [18, 19]. The data stream has specific characteristics that influence bot-recognition tools. First, it consists of high velocity data, as we capture nearly 40,000 tweets per day. Second, an additional characteristic is heterogeneity, and we work with a natural language stream that is highly multilingual. To illustrate, in the first 301 days of streaming, there were nearly 70 languages present, but 20 most frequent languages made up of 98.2% of the material. The most frequently used languages were English, Swedish, and Finnish, and the ensuing work focuses on these languages to develop tools for future work.

### 2.2 Defining autogenerated tweets

We follow [9] and define autogenerated tweets (AGT) as tweets where all or part of the natural language content is generated automatically by a bot, an application or any other type of program. Moreover, by definition we do not automatically include tweets posted by an application, since we only include those for which the application supplements some natural language content to the tweet. For example, a bot (or an app) that is retweeting a non-AGT is *not* producing a new AGT since it is not adding any natural language. Thus, AGTs in our definition come in two flavors. Tweets generated from pure bot accounts, such as weather bots, job bots, news bots, etc. The second type consists of tweets generated by applications and programs that are maintained and managed by humans. An opposite of an AGT is HGT (a human-generated tweet). Figure 1 above presents two examples of AGTs and HGTs according to our definition. See [9] for more details and examples related to this definition.

### 2.3 Establishing Ground Truth

The three datasets used in this paper are a random sample of: (1) 5,000 English tweets collected during a 10 days period in January 2017, (2) 5,000 Swedish tweets collected during a 10 days period in January 2017, and (3) 5,000 Finnish tweets collected during a 4 days period in April 2018.

The manual AGT/HGT annotation was made by persons very knowledgable in the languages they handled. Native speakers for the Swedish and Finnish datasets. In addition to the AGT definition and a lot of examples, each group of persons where given an excel sheet which, for each tweet, contained the user name, the actual Twitter text and a web link of type

https://twitter.com/anyuser/status/930432195436609536

giving the annotater a chance to see the tweet in a context, among other tweets published by same user. The web link gives the annotater a very good understanding of what type of user that was publishing the tweet.

The Finnish and Swedish are mainly used for training whereas the English tweets are utilized to evaluate the language independence of the classifier. While this is the first attempt to test the algorithm, the results of the pilot study using these three languages should be interpreted with some degree of caution. We plan on expanding the set of languages to other unseen languages in our future work. More importantly, the raw data used here can be made available upon request to those interested to allow replicability and encourage future comparisons.

### 2.4 Tweet properties used in the classification

The input to the AGT classifier consists of 10 tweet properties attaining numerical and nominal values that can be computed directly using the tweet metadata. These properties are selected as indicators that can be used (one at the time, or in combination) to identify non-human behavior. For instance, one should expect that humans have more followers than bots, or that AGTs tend to contain more URLs. The ten properties are:

- *isReply* - boolean indicating if the tweet is a reply
- *isRetweet* - boolean indicating if the tweet is a retweet
- *accountReputation* - number of followers divided by the number of friends and followers
- *hashtagDensity, urlDensity, mentionDensity* - number of hashtag/URL/mention entities, respectively, divided by the total number of the words in the tweet
- *statusesPerDay* - total number of user's tweets divided by account age in days
- *favoritesPerDay* - number of tweets favorited by user divided by account age
- *deviceType* - nominal variable based on the type of source used to post the tweet:
  1. mobile: Twitter for Iphone, Twitter for Android etc.
  2. web: Twitter Web Client, Tweetbot for Mac etc.
  3. app: Instagram, Tumblr, Foursquared etc.
  4. smm: Hootsuite, TweetDeck, dlvr.it, etc.
  5. bot: Trendsmap Alerting, SpotifyNowPlaying, etc.

6. unknown: not classified sources.

Apart from the nominal *deviceType* property, these properties have been discussed and evaluated in [9, 11, 12, 15, 21].

The tweet metadata contains an attribute `source` that identifies what type of an app or program that was used to post the tweet. We manually classified 150 most frequently used sources in our training set into five categories, 1-5, as defined in the *deviceType* attribute. These 150 sources cover about 97% of the training set, while the remaining (unlabelled) sources were uniformly labelled `unknown`. The device type smm stands for Social Media Management. That is, they are tools for managing content on multiple accounts on social networks.

| Device Type | Swedish | Finnish | English |
|---|---|---|---|
| mobile (1) | 65 (0.1) | 62 (3.3) | 51 (0.01) |
| web (2) | 21 (6.3) | 23 (7.5) | 19 (0.4) |
| app (3) | 2.7 (9.6) | 2.1 (18) | 8.2 (10) |
| smm (4) | 6.6 (40.4) | 5.4 (29) | 0.3 (52) |
| bot (5) | 0.7 (93.5) | 3.5 (96) | 21 (99.9) |
| unknown (6) | 2.5 (49.4) | 3.1 (74) | 0.001 (77) |

**Table 1.** Percentage of used device types and percentage of AGTs in each type.

The *deviceType* property turns out to be the backbone of our AGT classification (see Section 3.3). Table 1 shows the different device types that were used in the different datasets and what percentage of AGTs we find in each type. For example, 65% of all tweets in the Swedish dataset were posted using device type 1 (mobile) and only 0.1% of these tweets are labelled as AGTs. A noteworthy feature is that device types 1 (mobile) and 2 (web) dominate. They are used to post about 85% of all tweets. Notice also that the percentage of AGTs for these tweets (especially for type 1) is very low (0.01-3.3% depending on language). Hence, a tweet of type 1 or 2 is very likely to be a HGT. However, device type 5 (bot) is a clear (93.5-99.9%) indication of AGTs. This leaves three device types to be problematic. They are 3 (app), 4 (smm), and 6 (unknown), all of which can be either AGTs or HGTs. For these types, additional information drawn from the other properties is needed to make a classification.

### 2.5 Methodology in AGT classification

All classifiers are developed and evaluated using the Weka machine learning toolkit [23]. Rather than evaluating all applicable machine learning algorithms, we tested a few models but soon realized that tree-based models outperform the other available algorithms. Similar results (i.e. that tree-based models are suitable for tweet classification tasks) are reported in [9, 11, 21, 22]. The results presented below are generated using two different tree-based models: J48 and Random Forest.

J48 is an open-source Java implementation of the C4.5 algorithm by Quinlan [24]. Given a set of training data, the algorithm returns a decision tree in which

each split is made to increase the information gain by use of entropy. Random Forest [25] is an ensemble method, which utilizes de-correlated decision trees to produce a consensus model.

The metrics we use to evaluate the results of the classifiers are the error rate (%), and precision and recall [20]. Precision is the positive predictive value, i.e. the proportion of correctly classified instances among the total number of instances classified as AGTs. Recall or true positive rate is the proportion of correctly identified AGTs among the total number of AGTs.

### 2.6 Single language results

The number of tweets labelled as AGT varies between the languages: English (22.5%), Swedish (6.4%), and Finnish (11.4%). Table 2 shows brief classification results for classifiers trained (using 4,000 tweets) and tested (using 1,000) on the same language. The results are not as accurate as other more elaborated approaches for monolingual AGT detection (e.g. [9]), but still surprisingly good considering that only ten easily extracted metadata properties were used. These results show that English AGTs are rather easy to detect (accuracy 99.2%). A more detailed study of the English dataset shows that a large portion of the English AGTs are posted by bot accounts (e.g. weather bots) that are easily identified by the device type (5, bot) used to generate them.

| | **Error Rate (%)** | **Precision** | **Recall** |
|---|---|---|---|
| **English** | 0.82 | 0.991 | 0.973 |
| **Swedish** | 2.80 | 0.790 | 0.767 |
| **Finnish** | 4.75 | 0.865 | 0.701 |

**Table 2.** Error Rate (%), Precision, and Recall for RF-SP classifiers trained and tested using monolingual datasets.

As shown in the previous paragraph, the Swedish tweet dataset has the lowest AGT ratio (6.4%) and fewer AGTs generated by pure bot accounts than the English and the Finnish datasets. The most significant characteristic of the Swedish AGTs is that many are posted by SMM tools (type 4) that companies/organizations use to promote news published on their own websites. This type of AGTs are harder to detect but accuracy can be increased by including information of their device type (4) and combined with other properties (e.g. *accountReputation* or *statusesPerDay*)

Detecting AGTs in the Finnish dataset turns out to be the most difficult (i.e. recall being the lowest). Whereas Swedish newspapers often use SMM tools to promote news on Twitter, it looks like that many Finnish newspapers in our sample often take a more hands-on approach and manually share their newspaper web content on their Twitter account. Therefore, detecting this behavior automatically is difficult since the used device type often is of types (1 or 2) that we usually associate with HGTs, such as Twitter for Iphone or Twitter Web Client.

The fact that each dataset (language) has its own characteristics makes a classifier trained on a certain language less accurate when used to detect AGTs in another language. For example, Table 3 shows the result of applying the Finnish and Swedish classifiers on the English tweets. Notice that both classifiers are less accurate then the classifier (denoted English in Table 2) that was trained using English tweets.

| | **Error Rate (%)** | **Precision** | **Recall** |
|---|---|---|---|
| **Swedish** | 3.88 | 0.992 | 0.836 |
| **Finnish** | 1.72 | 0.987 | 0.936 |

**Table 3.** Error Rate (%), Precision, and Recall for RF-SP classifiers trained on Swedish and Finnish datasets and tested using the English dataset.

## 3 Multilingual AGT classifier

In this part, Section 3.1 presents the results of training an AGT classifier using a bilingual training set (Swedish and Finnish). Section 3.2 tests the classifier on a third unseen language, English, which is the dominant language in the NTS.

### 3.1 Bilingual training

In the training phase, we use the Swedish and Finnish datasets, a total of 10,000 tweets. The training results in terms of error rate, precision, and recall for each evaluated classifier (approximated by a separate 10-fold cross-validation) are shown in Table 4.

| | **RF-SP** | **J48-SP** | **J48-HP** |
|---|---|---|---|
| **Error Rate (%)** | 3.73 | 4.35 | 5.23 |
| **Precision** | 0.832 | 0.794 | 0.842 |
| **Recall** | 0.716 | 0.674 | 0.490 |

| | | predicted | |
|---|---|---|---|
| | | **AGT** | **HGT** |
| actual | **AGT** | 622 | 126 |
| | **HGT** | 247 | 9005 |

**Table 4.** Error Rate (%), Precision, and Recall for the Classifier models (left) and the Confusion Matrix for the best model (RF-SP).

The models RF-SP and J48-SP utilize the default pruning algorithms in Weka (SP = Standard Pruning) whereas J48-HP uses a hard pruning (low $\alpha$ value) that produces a smaller decision tree. The number of nodes in J48-HP is 39 compared to 160 nodes in the J48-SP model.

Random Forest (RF-SP) has the best training results (error rate 3.73%), followed by J48-SP (4.35%) and J48-HP (5.23%). Similar results for monolingual datasets (best model is Random Forest) are presented in [9, 11]. Note that the

| | RF-SP | J48-SP | J48-HP |
|---|---|---|---|
| **Error Rate (%)** | 2.84 | 3.98 | 1.84 |
| **Precision** | 0.992 | 0.936 | 0.998 |
| **Recall** | 0.882 | 0.884 | 0.920 |

| | | predicted | |
|---|---|---|---|
| | | **AGT** | **HGT** |
| actual | **AGT** | 1041 | 2 |
| | **HGT** | 90 | 3867 |

**Table 5.** Error Rate (%), Precision, Recall, and $F$-measure for the Classifier models (left) and the Confusion Matrix for the best model (J48-HP).

results presented here are not as accurate as the monolingual results presented for Swedish and Finnish in Table 2. Hence, while working with multilingual data streams may have benefits for sociolinguistic research, adding new languages to the classifier comes with a price. The classifier accuracy suffers even when it has been trained on a set that contains the same languages as the test set.

### 3.2 Testing on an unseen language

Our objective is to develop a light language independent application for AGT detection. The application should work on any language no matter if they are used in the training phase. Here we look into the results using the three most frequent languages in the NTS data. The classifiers were trained on a dataset with two languages (Swedish and Finnish), and Table 5 below shows the results of applying the classifiers on the English dataset.

The first thing to notice is that the standard pruning (RF-SP) results using a bilingual training set are better than the results using monolingual training sets in Table 3. Hence, a classifier trained on two languages outperforms classifiers trained on a single language when applied on tweets in an unseen language. While the result only applies to English tweets, the finding is encouraging since it indicates that by adding a few more languages to the training set we can expect a better result for yet unseen languages.

The second thing to notice, and a major surprise, is that contrary to the training results in Table 4, the simple model of J48-HP outperforms the more complex models (RF-SP and J48-SP). This indicates over-fitting, so that the complex models produced by RF-SP and J48-SP are much too specialized on the training languages Swedish and Finnish, and that the more coarse-grained J48-HP model focusing on only the essential information better adopts to handling a new language.

### 3.3 A closer look at J48-HP

Another criterion for our light language independent application for AGT detection is its suitability for digital humanities infrastructures, such as handling high-velocity data of the NTS. Our empirical findings suggest that the coarse-grained J48-HP model is the best model for classifying tweets in an unseen language by avoiding over-fitting to the languages used in the training data. Another advantage from the digital infrastructure perspective is that it is rather

```
favoritesPerDay <= 0.068
|  statusesPerDay <= 10.4
|  |  urlDensity <= 0.05: HGT
|  |  urlDensity > 0.05
|  |  |  isRetweet = 0
|  |  |  |  deviceType = 1: HGT
|  |  |  |  deviceType = 2
|  |  |  |  |  mentionDensity <= 0.026: HGT
|  |  |  |  |  mentionDensity > 0.026
|  |  |  |  |  |  isReply = 0
|  |  |  |  |  |  |  accountRep <= 0.48: AGT
|  |  |  |  |  |  |  accountRep > 0.48
|  |  |  |  |  |  |  |  accountRep <= 0.95: HGT
|  |  |  |  |  |  |  |  accountRep > 0.95: AGT
|  |  |  |  |  |  isReply = 1: HGT
|  |  |  |  deviceType = 3: HGT
|  |  |  |  deviceType = 4
|  |  |  |  |  urlDensity <= 0.11: HGT
|  |  |  |  |  urlDensity > 0.11: AGT
|  |  |  |  deviceType = 5: AGT
|  |  |  |  deviceType = 6
|  |  |  |  |  statusesPerDay <= 0.37
|  |  |  |  |  |  accountRep <= 0.27: AGT
|  |  |  |  |  |  accountRep > 0.27: HGT
|  |  |  |  |  statusesPerDay > 0.37: AGT
|  |  |  isRetweet = 1: HGT
|  statusesPerDay > 10.4
|  |  deviceType = 1: HGT
|  |  deviceType = 2: HGT
|  |  deviceType = 3: HGT
|  |  deviceType = 4
|  |  |  urlDensity <= 0.052
|  |  |  |  statusesPerDay <= 42.7: HGT
|  |  |  |  statusesPerDay > 42.7: AGT
|  |  |  urlDensity > 0.052: AGT
|  |  deviceType = 5: AGT
|  |  deviceType = 6: AGT
favoritesPerDay > 0.068: HGT
```

**Fig. 2.** The entire J48-HP decision tree.

easy to comprehend and therefore probably more robust (having low variance). Figure 2 below shows the entire model.

The J48-HP decision tree uses `favoritesPerDay > 0.068` to identify accounts with a typical human behavior (they like other posts) and marks all tweets published by these accounts as HGTs. It then separates the accounts into active (`statusesPerDay > 10.4`) and not-so-active (`statusesPerDay <= 10.4`). Next, it uses the `deviceType` property as the base and the other proper-

ties mainly to identify non-human behavior for more problematic device types, as presented in Section 2.4 (i.e. device types 2, 4, and 6).

## 4 Summary and future work

Handling and processing heterogeneous and highly variable natural language data is rarely without problems. Additional complications are added when processing needs to be done in real-time for high-velocity data. Our goal has been to present ongoing efforts to build a language independent classifier for detecting autogenerated tweets (AGTs) written in any language.

Here we piloted one system and trained one AGT classifier using a training set consisting of tweets in two languages, Swedish and Finnish. We evaluated the classifier using a third dataset of English tweets. The classifier is in principle language independent since it does not use the actual Twitter text but only relies on language and country independent metadata that are available in each tweet.

The results, considering this straightforward approach, are surprisingly[7] accurate: Error rate (1.84%), Precision (0.998), and Recall (0.920). However, the results are not as accurate as monolingual Twitter classifiers using the Twitter text and user timeline information, which is downloaded separately (e.g. [9,11]), but we propose that they can increase data accuracy in many fields in the humanities. Moreover, they are most likely sufficiently accurate for many digital humanities research projects that would like to filter out AGTs from their datasets. Lastly, our approach is also useful for online AGT detection and handling large Twitter datasets where the time needed to download user timeline information might be problematic because of speed and/or volume.

The results also indicate that better results for unseen languages can be achieved by using a training set with several languages. Hence, as a part of future work we plan to add 2-3 more languages to the training set. This obviously requires (wo)manpower, and we acknowledge the help from the students Hanna Kernen and Irene Taipale at the University of Eastern Finland for their help with labeling the Finnish dataset used here.

The results show that the most simple model (J48-HP) outperformed more complex models, which had better training results, when applied to an unseen language. This finding indicates over-fitting with respect to the languages used in the training data. The fact that the test set is qualitatively different from the training set is not a standard scenario in machine learning and needs to be addressed in future studies.

In this paper we trained the classifier using Swedish and Finnish tweets, and evaluated the approach using English tweets. Exploring other combinations (e.g. train on Finnish and English, evaluate using Swedish) as well as including more languages is also future work.

[7] Compared with much more elaborated monolingual approaches, such as [9].

## References


1. D.G. Campbell, "Egypt Unsh@ckled - Using Social Media to@#:) the System: how 140 Characters Can Remove a Dictator in 18 Days", Cambria Books, United Kingdom, 2011
2. A. Tumasjan, T. Sprenger, P. Sandner, I. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", Int. AAAI Conference on Web and Social Media, pp.178–185, 2010
3. D. Gayo Avello, P.T. Metaxas, E. Mustafaraj, "Limits of electoral predictions using Twitter", International AAAI Conference on Weblogs and Social Media, pp.490–493, 2011
4. J.N. Sutton, L. Palen, I. Shklovski, "Backchannels on the front lines: Emergency uses of social media in the 2007 Southern California Wildfires", International Conference on Information Systems for Crisis Response and Management, 2008
5. T. Sakaki, M. Okazaki, Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors", International conference on World wide web (WWW'10), pp.851–860, 2010
6. J. Bollen, H. Mao, X. Zeng, "Twitter mood predicts the stock market", Journal of Computational Science, vol. 2, pp.1–8, 2011
7. T. Scheffler, "A German Twitter Snapshot", International Conference on Language Resources and Evaluation (LREC'14), 2014
8. M. Laitinen, M., J. Lundberg, M. Levin, A. Lakaw, "Creating the Nordic tweet stream: A real-time monitor corpus of rich and big language data". Journal of Universal Computer Science, vol. 23, pp.1038–1056, 2018.
9. J. Lundberg, J. Nordqvist, A. Matosevic "On-the-fly Detection of Autogenerated Tweets ", arXiv preprint 1802.01197.
10. N. Ljubešic, Nikola and D. Fišer, T. Erjavec, "TweetCaT: a tool for building Twitter corpora of smaller languages ", Proceedings of LREC, 2014.
11. Z. Chu, H. Wang, S. Gianvecchio, S. Jajodia, "Detecting Automation of Twitter Accounts: Are you a Human, Bot, or Cyborg?", Transactions on Dependable and Secure Computing, vol. 9, pp.811–824, 2012
12. C.M. Zhang, V. Paxson, "Detecting and analyzing automated activity on Twitter", International Conference on Passive and Active Network Measurement, pp.102–111, 2011,
13. N. Chavoshi, H. Hamooni, A. Abdullah, "Identifying correlated bots in Twitter", Int. Conference on Social Informatics, pp.14–21, 2016,
14. F. Morstatter, L. Wu, T. Nazer, K.M. Carley, H. Liu, "A new approach to bot detection: striking the balance between precision and recall", International Conference on Advances in Social Networks Analysis and Mining, pp. 533–540, 2016,
15. V.S. Subrahmanian et al, "The DARPA Twitter Bot Challenge", Computer, The IEEE Computer Society, pp.38–46, vol. 49, 2016,
16. Y. Huang, G. Diansheng, A Kasakoff, J. Grieve. 2016. "Understanding US regional linguistic variation with Twitter data analysis" Computers, Environment and Urban Systems, vol. 59, pp. 244255, 2016
17. M. Laitinen, J. Lundberg, M. Levin, R. Martins. 2018, "The Nordic Tweet Stream: A Dynamic Real-Time Monitor Corpus of Big and Rich Language Data", Proc. of Digital Humanities in the Nordic Countries 3rd Conference, Helsinki, Finland, March 7-9, 2018
18. M. Laitinen, J. Lundberg, M. Levin, A. Lakaw, "Revisiting weak ties: using present-day social media data in variationist studies". In Tanja Sily, Minna Palander-Collin,

Arja Nurmi, Anita Auer (eds.), Exploring Future Paths for Historical Sociolinguistics, Amsterdam: John Benjamins, pp 303325, 2017
19. M. Laitinen, J. Lundberg. “ELF and social networks: Evidence from a third-generation ELF corpus”. In Anna Mauranen and Svetlana Vetchinnikova (eds.), Language Change: The Impact of English as a Lingua Franca. Cambridge: Cambridge University Press, 2018, *Forthcoming*
20. S. Coats, ”Grammatical frequencies and gender in Nordic Twitter Englishes”, Michael Beiwenger (eds.). Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities, Ljubljana: U. of Ljubljana Academic Publishing, pp.12–66, 2016
21. K. Lee, B.D. Eoff, J. Caverlee, ”Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter”, International AAAI Conference on Weblogs and Social Media, 2011
22. M. Mccord, N. Chuah, ”Spam detection on Twitter using traditional classifiers”, International Conference on Autonomic and Trusted Computing, pp.175–186, 2011
23. M. Hall et al., The Weka Data Mining Software: An Update, SIGKDD Explorations Newsletter, vol. 11, pp.10–18, 2009
24. J.R. Quinlan ”C4.5 : Programs for machine learning”, Morgan Kaufmann Publishers, 1993
25. L. Breiman, ”Random Forests” Machine Learning, vol. 45, pp.5–32, 2001
26. D. M. W. Powers, ”Evaluation: From precision, recall and f-measure to ROC, informedness, markedness & correlation”, Journal of Machine Learning Technologies, vol. 2, pp.37–63, 2011