

Private or Corporate? Predicting User Types on Twitter

Nikola Ljubešić

Dept. of Knowledge Technologies
Jožef Stefan Institute
nikola.ljubesic@ijs.si

Darja Fišer

Department of Translation
Faculty of Arts, University of Ljubljana
darja.fiser@ff.uni-lj.si

Abstract

In this paper we present a series of experiments on discriminating between private and corporate accounts on Twitter. We define features based on Twitter metadata, morphosyntactic tags and surface forms, showing that the simple bag-of-words model achieves single best results that can, however, be improved by building a weighted soft ensemble of classifiers based on each feature type. Investigating the time and language dependence of each feature type delivers quite unexpected results showing that features based on metadata are neither time- nor language-insensitive as the way the two user groups use the social network varies heavily through time and space.

1 Introduction

Due to popularity of Twitter and its proactive content harvesting policy, automatic identification of latent user attributes has become a hot research topic in the past couple of years. Tasks range from discriminating between different types of users (Mislove et al., 2011), behaviour (Pennacchiotti and Popescu, 2011; Rao et al., 2010), location (Hecht et al., 2011), gender (Burger et al., 2011), age (Nguyen et al., 2013), occupation (Hu et al., 2016), social class (Borges et al., 2014) and personality type (Quercia et al., 2011).

In this paper we present a series of experiments on discriminating between corporate and private accounts on Twitter. We investigate language-independent features, features extracted from morphosyntactic annotations and simple bag-of-words features. We additionally investigate the time and space dependence of specific feature sets by measuring how well specific features handle training on one time span or closely-related language and testing on another time span or other closely-related language.

2 Related work

Automatic classification of different types of users is an active area of research and has been attempted at various degrees of granularity and for different communicative goals. Naaman et al. (2010), for example, distinguish two types of users: those who disseminate information (informers) and those who share what they are doing and how they are feeling (meformers). De Choudhury et al. (2012), on the other hand, distinguish between the accounts of individual persons, organizations, news media and other miscellaneous categories. Kim et al. (2010) use the same set, only adding the celebrities category, while (Wu et al., 2011), suggest a two-stage approach where they first classify users into two broad categories of elite and ordinary users, and then further classify the elite users into more fine-grained groups of media, celebrities, organizations and bloggers.

Similarly different both in terms of number and complexity are the features used for classification. Rao et al. (2010) use 2 categories of over 20 features that are easy to obtain from the Twitter timeline: the network structure (follower-following ratio, follower frequency, following frequency) and the communication behavior of individuals (response frequency, retweet frequency, tweet frequency), showing that tweeting behavior information is not useful for most classification tasks. Therefore, in addition to profile features, tweeting behaviour and social network, (Pennacchiotti and Popescu, 2011), show that,

This work is licenced under a Creative Commons Attribution 4.0 International License. License details:
<http://creativecommons.org/licenses/by/4.0/>

rich linguistic features (prototypical words and hashtags, generic and domain-specific LDA, sentiment words) yields consistently better results. Kucukyilmaz et al. (2008) employ a more traditional approach to author characterization by relying on a set of term-based features in users’ messages and a set of stylistic features (character usage, message length, word length, punctuation usage, punctuation marks, stopword usage, stopwords, smiley usage, smileys, vocabulary richness).

While our work has noticeable overlap with existing research that was carried out on English, it extends that research, beyond the language shift, by (1) performing a more detailed analysis of a broader list of features, and inspecting (2) the stability of those features in different time periods and (3) their portability to a closely-related language.

3 Dataset

The dataset used in this paper consists of 7.5 million Slovene tweets collected from June 2013 to January 2016 with the TweetCaT tool (Ljubešić et al., 2014). Each of the 7778 users in the collection was manually annotated as private or corporate. We performed our experiments on 5842 users for whom we had at least 100 tweets at our disposal. Out of these 5842 users, 4382 were annotated as private and 1460 as corporate.

4 Features

We discriminate between two basic types of features: language-independent (LI) and language-dependent (LD) ones. We calculate our language-independent features both from the text of a tweet and its additional metadata. We use two types of language-dependent features: focused (FLD) and bag-of-words (BoW) features. While the bag-of-words features are simply types found in the text of all the training tweets, the focused ones are mostly based on morphosyntactic annotation of the text. It should be stressed here that we do not presuppose the language-independent features to work well across languages for our task, but just that these features can be found in different languages.

All the features are calculated on user level as we perform our predictions on each Twitter user.

4.1 Language independent features

Among the language-independent features we discriminate between the following feature types:

- `avg` – average number of tweets satisfying a condition
- `mean` – mean of a continuous tweet-level variable
- `med` – median of a continuous tweet-level variable
- `var` – variance of a continuous tweet-level variable
- `user` – user-level metadata

In Table 1 we give an overview of all language-independent features, sorted in descending order by the p-value obtained with the two-sided Mann-Whitney U test (Mann and Whitney, 1947) on the variable in question regarding the distribution in private and corporate accounts. The null hypothesis of this test is that the two samples (variable measurements on private and corporate accounts) come from the same distribution. Beside the feature description and its univariate p-value, we report the effect size as the difference in the median between private and corporate accounts. The variables reporting a positive effect size, like the `average_inreply` variable, therefore have a higher median in private tweets while the variables with a negative effect size, like the `average_http` variable, have a higher median among corporate tweets.

Our univariate feature analysis shows, among many other things, that private accounts are most specific for replying to tweets, mentioning other users, favoring other user’s tweets, tweeting in various hours of the day, producing tweets of variable length and posting more frequently during weekends. On the other

Variable	Description	p-value	Effect size
avg_http	average number of tweets containing URLs	0.0	-0.61704
avg_inreply	average number of tweets that are replies	0.0	0.38549
avg_mention	average number of tweets containing mentions	10^{-306}	0.45676
avg_work_hrs	average number of tweets published in working hours	10^{-235}	-0.15252
user_favor	number of tweets the user has favorited	10^{-209}	302
var_hour	variance of the posting hour	10^{-208}	5.96132
var_len_text	variance of the tweet length	10^{-203}	279.092
mean_hour	mean of the posting hour	10^{-188}	1.71806
median_hour	median of the posting hour	10^{-185}	2
mean_len_text	mean of the tweet length	10^{-113}	-12.8003
med_len_text	median of the tweet length	10^{-99}	-17.25
avg_weekend	average number of tweets posted on weekends	10^{-97}	0.03847
mean_day	mean of the posting weekday	10^{-95}	0.25134
var_day	variance of the posting weekday	10^{-78}	0.36336
user_ff_ratio	followers / friends ratio of a user	10^{-63}	-0.61667
var_favor	variance of number of favorites	10^{-61}	0.28596
avg_favor	average number of favorited tweets	10^{-55}	0.10024
mean_favor	mean of number of favorites	10^{-55}	0.15367
med_day	median of the posting day	10^{-53}	0
avg_comma	average number of tweets containing a comma	10^{-49}	0.08569
avg_quest	average number of tweets containing a questionmark	10^{-49}	0.04413
avg_emoji	average number of tweets containing an emoji	10^{-40}	0.0
avg_is_quote	average number of tweets that are quotes	10^{-37}	0.00196
avg_excl	average number of tweets with an exclamation mark	10^{-31}	-0.05624
user_status	number of statuses the user produced	10^{-26}	518
avg_client	average number of tweets sent through the web client	10^{-16}	0.12875
user_friend	number of friends the user has	10^{-16}	69
avg_retweet	average number of retweeted tweets	10^{-14}	-0.03827
user_per_day	number of statuses a users sends per day	10^{-11}	0.18937
avg_ellipsis	average number of tweets containing an ellipsis	10^{-8}	0.01381
avg_hash	average number of tweets containing a hashtag	10^{-5}	0.03624
user_follow	number of followers the user has	10^{-5}	-40.5
mean_retweet	mean of number of retweets	10^{-5}	-0.03772
med_retweet	median of number of retweets	0.002	0.0
var_retweet	variance of number of retweets	0.011	0.01123
user_listed	number of lists the user is on	0.041	0.0
med_favor	median of number of favorites	0.287	0.0
avg_trunc	average number of tweets that are truncated	0.999	0.0

Table 1: An overview of the language-independent features, sorted by their p-value,

Variable	Description	p-value	Effect size
adjpos	adjective in positive degree	0.0	-0.02052
prondemsgdl	demonstrative pronoun in singular / dual	0.0	0.0059
verb12sgdl	verb in first / second person singular / dual	0.0	0.01406
pronneg	negative pronoun	10^{-257}	0.00088
supine	supine	10^{-251}	0.00146
verbmainpl	main verb in plural	10^{-245}	-0.00963
verbimppl	verb in imperative plural	10^{-238}	-0.00198
private_words	list of auxiliary verb and personal pronoun forms	10^{-212}	0.00641
num	number	10^{-204}	-0.00942
pronind	indefinitive pronoun	10^{-203}	0.00097
lower	lowercased word	10^{-181}	0.0721
adjpart	adjectival participle	10^{-153}	-0.00219
pronperspl	personal pronoun plural	10^{-151}	-0.00235
normalised	word automatically normalised	10^{-137}	-0.05837
interjection	interjection	10^{-130}	0.00105
adjsup	adjective in superlative degree	10^{-119}	-0.00054
content	content word	10^{-108}	-0.04989
nounprop	proper noun	10^{-105}	-0.01867
title	titlecase word	10^{-92}	-0.02505
pronpers	personal pronoun	10^{-79}	0.00453
double	token repetition	10^{-62}	0.00014
foreign	foreign word	10^{-48}	0.04383
upper	uppercase word	10^{-44}	-0.00862
pronposs1sgdl	possesive pronoun in singular / dual	10^{-44}	-0.00022
adjposs	possesive adjective	10^{-42}	-0.00049
pronposspl	possesive pronoun in plural	10^{-37}	-0.00016
emoticon	emoticon	10^{-13}	0.00595

Table 2: An overview of the language-dependent focused features, sorted by their p-value

hand, corporate accounts use more URLs, post more during working hours, produce longer tweets and have a higher followers / friends ratio.

All the engineered features, except for the median of number of favorites (because it is most frequently 0) and the average number of truncated tweets (most frequently 0 as well) have shown to be statistically significant with $p < .05$.

4.2 Language-dependent features

In Table 2 we give an overview of the focused language-dependent features, with a statistical analysis identical to the one of the language-independent features. All the variables are user-level probabilities of specific linguistic phenomena. Again, the effect size is the difference in the median between private and corporate accounts, i.e. a positive effect size points at higher values among private users.

This univariate analysis shows, among other things, for private accounts to use more singular / dual demonstrative pronouns, singular / dual verbs, negative pronouns and supines (short verb infinitive forms). On the other hand, corporate accounts use more adjectives, verbs in plural, numbers, superlatives and content words.

In Table 3 we give an overview of the 20 strongest word features identified through the coefficient value of the linear discriminative classifier that will be described in the Experiments section. As the table shows, private users prefer pronouns and verbs in 1st person singular (I, me, my, I did, I wish, I hope), nouns and verbs depicting thoughts and activities carried out in their spare time (wish, grill, slim), while 3rd and 2nd person plural forms (we, our, we did, we can, you check, we congratulate), nouns describing

sem(did-1st prs sg) i moj(my) cerar http://t.co/h7jwauc0rv me my bravo želim(wish-1st prs sg) zalogi(stock) mesto(city) upam(hope-1st prs sg) zoya #delajvitko(#workslim) ko(when) http://t.co/qcfz6ii342 pod(under) salesforce žaru(grill) @delo tv
zaradi(because) smo(did-3rd prs pl) teden(week) klik(click) čestitamo(congratulations) evrov(euro) tf2 piše(writes) festival preveri(check-2nd prs sg) we people ste(did-2nd prs pl) naš(our) kamala #volitve14(#election14) lahko(can) novo(new) akcija(sale) facebook

Table 3: List of the strongest language-dependent bag-of-word features, for private and for corporate users

Feature type	Corporate	Private	All	FLD	BoW	ENS
Most frequent class (MFC)	0.0000	0.8572	0.6430			
Language independent (LI)	0.7944	0.9335	0.8987	< 0.001	< 0.001	< 0.001
Focused lang. dependent (FLD)	0.8520	0.9510	0.9263		< 0.001	< 0.001
Bag of words (BoW)	0.8742	0.9577	0.9368			< 0.01
Ensemble (ENS)	0.8864	0.9620	0.9431			

Table 4: Evaluation results via weighted F1 on the per-feature-type classifiers and the ensemble classifier. Statistical significance of the differences is calculated via approximate randomisation.

public events (festival, sales, election, Euro) and named entities (Kamala, Facebook), as well as verbs lexicalizing professional activities (he writes, you check) are characteristic of the corporate users.

5 Experiments

We run three batches of experiments. In the first batch we experiment with the three types of features presented in the previous section. Additionally, we build an ensemble classifier which uses the output of the three feature-type classifiers. In the second batch we investigate the time sensitivity of the presented feature types by training on data from one time span and testing on a later one. Finally, in the third batch we investigate the portability of two feature types, namely language-independent features and bag-of-words features, to a closely-related language.

As our weak baseline we use the most-frequent-class (MFC) baseline. Our classifier of choice on all feature types are support vector machines (SVM). In case where the number of features is smaller than the number of instances we use the RBF kernel while in the opposite case we use a linear kernel.

In all the experiments we 5-fold with stratification over all our data, optimising in each iteration the classifier via grid search on the development data and evaluating it on the test data via weighted F1.

In case of the RBF kernel we optimise the C ($2^n, n \in \{-5, -3, -1, \dots, 15\}$) and γ ($2^n, n \in \{-15, -13, -11, \dots, 3\}$) hyperparameters while for the linear kernel we optimise the C parameter in the same range as with the RBF kernel.

5.1 Feature type comparison

We present the results of training a classifier on each type of features, the language independent (LI), focused language dependent (FLD) and bag-of-words (BoW), in Table 4. We additionally compare these classifiers to our weak most-frequent-class (MFC) baseline and an weighted soft ensemble of all three classifier outputs. The weighted soft ensemble classifier uses the per-class probability output of each classifier, weights each class distribution with the weighted F1 of that classifier estimated on the development data, averages over all the distributions and picks the class with maximum probability.

We report F1 on each class and an overall weighted F1. We additionally report a p-value obtained through the approximate randomisation test (Yeh, 2000) with $R = 1000$ for each classifier combination. Approximate randomisation estimates the p-value as $p = \frac{r}{R}$ where r is the number of iterations where randomly switching responses of system 1 and system 2 has produced a greater or equal difference in the evaluation metric than the original difference between system 1 and system 2, while R is the number of iterations over the responses.

Feature type	All	Time dependence		Close language	
		≥ 2015	Loss (%)	Croatian	Loss (%)
Most frequent class (MFC)	0.6430	0.6430	-	0.3388	-
Language independent (LI)	0.8987	0.8370	7%	0.6270	30%
Focused language dependent (FLD)	0.9263	0.9149	1%	-	-
Bag of words (BoW)	0.9368	0.9280	1%	0.7304	22%

Table 5: Results on the time and language sensitivity of specific feature types

The results show that each of the feature types improves over the weak MFC baseline by a large margin. While all the three feature types come quite close to each other, the weakest one are the language-independent features, mostly relying on tweet and user metadata. We find it quite surprising that the bag-of-words model improves over the focused language-dependent model which uses a language abstraction in form of morphosyntactic descriptions.¹ Finally, the weighted soft ensemble improves over each of the classifiers.² The difference between each classifier combination has proven to be statistically significant.

The performance of each classifier on specific classes did not yield any surprises. The private class, which is more represented, performs regularly better. The best-performing ensemble classifier yields weighted F1 of 0.89 on the corporate and 0.96 on the private class.

Regarding the number of features, the single best performing classifier is also the heaviest one with 3,978,030 features. On the other hand the LI classifier works with 38 features, while the FLD classifier works with 27 features.

5.2 Time sensitivity

In this set of experiments we investigate the impact of training on one time span and evaluating on another one. We again experiment with our three types of features. In each iteration of cross-validation, the development set consists only of tweets published before 2015, removing all users for which after filtering we do not have 100 tweets available. We repeat the same process on the test set by keeping only tweets published 2015 onward.

Along with our new time dependence evaluation results, we repeat the results from the previous batch of experiments which presents the performance of the feature type in the same time span.

The results of this experimental batch are presented in the left hand side of Table 5.

The results are somewhat surprising showing that the LI features are actually the least time insensitive. Probably due to the way the Twitter social network was used before and after 2015 the loss reaches 7% of weighted F1. On the other hand, the FLD and BoW classifiers measure a minor loss of 1%.

5.3 Portability to a similar language

We perform a final set of experiments investigating the portability of the language independent (LI) and the bag-of-words (BoW) features to a closely related language. As our closely related language we chose Croatian as the two countries are strongly interconnected, expecting therefore that Twitter is used in a comparable manner and that a usable overlap in surface forms among the two closely-related languages should occur. This lexical overlap has already been used in previous work on bilingual lexicon extraction (Fišer and Ljubešić, 2011) and identification of false friends (Ljubešić and Fišer, 2013).

We discard the focused language dependent (FLD) features from these experiments as they require the texts of the tweets to be morphosyntactically annotated with the same tagset which would require a tagset mapping.

In these experiments we use the full Slovene data for development and evaluate on a set of 101 Croatian users that were annotated by hand, using the same criteria applied while annotating Slovene data. In the Croatian dataset 50 users were annotated as private and 51 as corporate.

¹We have investigated expanding the BoW model with character and word n-grams, but it did not improve over the simple word unigram model.

²Beside the weighted soft ensemble we also experimented with a stacked classifier, but this classifier, being more complex than our ensemble, did not yield any significant improvement.

The results are presented in the right hand side of Table 5. While both classifiers outperform the weak baseline by a wide margin, interestingly, the language independent features prove to be the less stable feature type when a shift in the target domain / language occurs. While, naturally, the BoW feature set experiences a heavy loss of 22% when evaluating on Croatian data, the loss of 30% on the LI features is even greater, in addition to the initial lower results.

5.4 Error analysis

We conclude the experiments with an error analysis of the instances misclassified by the ensemble classifier. Among those instances we specifically analyse the cases where the classifier based on language-independent features (LI) and the bag-of-words classifier (BoW) disagree.

We performed a detailed qualitative analysis of the tweeting behaviour and language of 10 random accounts per class misclassified by each system in order to gain more insight into why the misclassification might have occurred. In total, 4 scenarios were analysed: (1a) corporate and (1b) private accounts misclassified by LI but correctly classified by BoW, and (2a) corporate and (2b) private accounts misclassified by BoW but correctly classified by LI.

As far as the corporate accounts misclassified by the LI approach but correctly classified by the BoW approach (1a) are concerned, they were smaller / early-stage societies / NGOs / civil initiatives, popular radio stations, records companies, event agencies and bloggers who (probably deliberately) share many tweeting characteristics of private accounts in terms of tweeting amount, time, retweets, mentions, favourites and hashtags. Among the private accounts misclassified by the LI approach (1b), 9 out of 10 examined accounts belong to journalists, politicians and professionals who do not explicitly disclose their professional status or affiliation in their account profiles but nevertheless tweet for professional purposes and demonstrate all the behaviour of the corporate accounts. This suggests that in addition to straightforward binary categories, there is a grey zone of borderline, difficult to define account types which call for more detailed annotation guidelines.

Interestingly, there was only one corporate account that was misclassified by the BoW approach but correctly classified by the LI one (2a), which is a fair-trade store that raises awareness for fair-trade home products, clothes and food in a communicative, conversational, friendly way. The topics covered and vocabulary used is therefore closely related to our every-day lives, not unlike the topics and vocabulary present in the tweets published by private accounts. Similar results were obtained with the qualitative analysis of 10 private accounts erroneously classified by BoW but correctly by LI (2b). They show that 9 of them are borderline because they are accounts of journalists, managers and small business owners whose tweets, published with the typical corporate dynamics, cover topics conveyed by the linguistic features that are typically associated with people's private lives, such as health and lifestyle, food, cosmetics, hairstyle, sports and charity.

The error analysis clearly shows that the erroneously classified accounts are indeed outliers, either by tweeting behaviour, or by tweeting content and / or language style. This suggests that manual annotation guidelines should be refined to include advice how to treat such potentially difficult cases and / or extend the classification beyond a binary approach.

6 Conclusion

In this paper we have presented a series of experiments on discriminating between private and corporate Twitter accounts among Slovene Twitter users.

In the first part of the paper we have analysed a series of features, casting some light on the way the two groups use our medium of interest.

In our classification experiments we have shown that the simple bag-of-words model outperforms the language-independent and focused language-dependent features. Building a classifier ensemble yielded a statistically significant improvement, showing that these feature sets are at least partially complementary.

We have shown that language-independent features based mostly on Twitter metadata are actually very unreliable, showing bad performance both when different time spans or different languages are used for training and evaluation. A big surprise was that the bag-of-words model was more portable to a closely

related language than the feature set based on metadata. These results point at significant differences in the way the social network is used across time and space.

Overall we have achieved very good results when combining all classifiers in a weighted soft ensemble. Given that most researchers are interested in removing corporate accounts from their Twitter collections, we can expect that our classifier will be very useful for Slovene as F1 on the private class is more than 96%.

References

- Guilherme R Borges, Jussara M Almeida, Gisele L Pappa, et al. 2014. Inferring user social class in online social networks. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, page 10. ACM.
- John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*, pages 1301–1309, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Munmun De Choudhury, Nicholas Diakopoulos, and Mor Naaman. 2012. Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 241–244. ACM.
- Darja Fišer and Nikola Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 125–131, Hissar, Bulgaria. RANLP 2011 Organising Committee.
- Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. 2011. Tweets from Justin Bieber’s Heart: The Dynamics of the Location Field in User Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11*, pages 237–246, New York, NY, USA. ACM.
- Tianran Hu, Haoyuan Xiao, Jiebo Luo, and Thuy vy Thi Nguyen. 2016. What the Language You Tweet Says About Your Occupation.
- Dongwoo Kim, Yohan Jo, Il-Chul Moon, and Alice Oh. 2010. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *ACM CHI workshop on microblogging*. Citeseer.
- Tayfun Kucukyilmaz, B Barla Cambazoglu, Cevdet Aykanat, and Fazli Can. 2008. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, 44(4):1448–1466.
- Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2014. TweetCaT: a Tool for Building Twitter Corpora of Smaller Languages. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Nikola Ljubešić and Darja Fišer. 2013. Identifying false friends between closely related languages. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 69–77, Sofia, Bulgaria, August. Association for Computational Linguistics.
- H.B. Mann and D.R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.
- Alan Mislove, Sune Lehmann Jørgensen, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist, 2011. *Understanding the Demographics of Twitter Users*, pages 554–557. AAAI Press.
- Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM.
- Dong-Phuong Nguyen, Rilana Gravel, RB Trieschnigg, and Theo Meder. 2013. ”how old do you think i am?” a study of language and age in twitter.
- Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification.

- Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. 2011. Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 180–185. IEEE.
- Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents, SMUC '10*, pages 37–44, New York, NY, USA. ACM.
- Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 705–714. ACM.
- Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics - Volume 2, COLING '00*, pages 947–953, Stroudsburg, PA, USA. Association for Computational Linguistics.