Identifying Tweets from the Economic Domain

Alexander Böhm
University of Passau
Innstraße 41
Passau, Germany
boehm@computer.org

Christin Seifert
University of Passau
Innstraße 41
Passau, Germany
christin.seifert@uni-passau.de

Jörg Schlötterer
University of Passau
Innstraße 41
Passau, Germany
joerg.schloetterer@uni-passau.de

Michael Granitzer
University of Passau
Innstraße 41
Passau, Germany
michael.granitzer@uni-passau.de

ABSTRACT

The rapid spread of information on the microblogging platform Twitter bears great potential for scientists to share their ideas and keep track on recent research results. However, Twitter lacks the ability to identify (and hence to connect to) other users in a particular domain. We adapt an approach to identify computer scientists on Twitter to the domain of economics. We learn a model, which classifies users as economist or non-economist, based on ground-truth data obtained by means of an initial set of Twitter accounts of economic journals. We further conduct an analysis of content-based and network features. Our evaluation shows, that economists can be detected rather accurately on twitter based on tweet terms ($F1 = 0.7$ with Naive Bayes). Neither adding hash-tags nor followers could further improve the classification model.

1. INTRODUCTION

With over 280 million monthly active users, Twitter is one of the biggest online social networks in the world[1], empowering its users to share information and news between their peers and the community. By the sheer amount of users, a lot of different topics of interest are shared and communicated by using Twitter. The rapid spread of information bears great potential for researchers to share recent results, but the information reaches only a small amount of the potential target group, due to the lack of domain-specific search capabilities: There is no comprehensive directory for those researchers. Even though Twitter provides recommendations of potentially interesting users to follow [3], it does not provide a dedicated tool for users to connect and interact in a particular topic of interest. Also, Twitter directories[2] which categorize users, cannot fully satisfy this task, since users need to register to such a directory themselves and manually update the information. Manually curated lists, such as the list of the Saturday Economist[3], can only capture a small percentage of scientists, i.e. the most popular ones. Publication databases would provide the possibility to find scientists of a particular domain, but in turn, lack the connection to social networks.

A successful approach to identify computer scientists on Twitter has been proposed by Hadgu and Jäschke [5]. Our approach differs in two ways: First, our application domain is economics. Second, we classify on the level of single tweets (instead of tweet history and user profile data), which makes it applicable for classifying based on the twitter streaming API, and identifying economists based on individual tweets. Our evaluation shows, that even though economists are less active on Twitter than computer scientists, it is still possible to identify the former. Specifically, the contributions of this paper are the following:

- We show that the approach for computer scientist detection [5] is adaptable to the domain of economists.
- We provide a feature analysis of content-based and network features and evaluate the extent to which each feature contributes to classifier performance.

The rest of the paper is structured as follows: In section 2 related work is reviewed. Section 3 provides details of the conceptual approach and implementation. Evaluation results are provided in section 4. Section 5 provides a summary and outlook on future work.

2. RELATED WORK

Twitter itself offers a suggestion service, which recommends accounts a user may be interested in. These sug-

[1] http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[2] e.g., http://justtweetit.com or http://www.tweetfind.com/

[3] https://www.thesaturdayeconomist.com/top-economists-on-twitter.html

gestions are based on "shared interests, common connections and a number of other factors" [3] and Twitter does not provide detailed insights on the exact features and algorithms used. Kywe et al. [7] provide an overview on recommendation techniques on Twitter, which contains (amongst others) approaches to recommend interesting users to follow. The limitation of these recommendations in the context of identifying scientists in a particular domain is that the suggestions cannot be filtered for a certain domain, but account for a broader user context and similarity.

Several approaches exist, that aim to identify user interests on Twitter. For example, Kapanipathi et al. [6] construct hierarchical interest profiles from a user's tweets. In their work the interest hierarchy is based on the Wikipedia category graph and the profiles are constructed on the basis of Wikipedia entities detected in the tweets and spreading activation in the Wikipedia graph. Similarly, Michelson and Macskassy [10] also use the Wikipedia category system to derive and represent a user's interests, but omit a taxonomy. These approaches result in a distribution over different topics of interest and hence are not suitable to identify researchers in a particular domain without further effort. Another approach to construct interest profiles from tweets has been presented by Tao et al. [11]. However, their profiles only exhibit a limited set of topics, which renders the identification of economists infeasible.

Hadgu and Jäschke [5] presented an approach to identify computer scientists on Twitter, based on machine learning. They collect ground truth data through expansion of a set of initial seeds (i.e., scientific conference accounts) by followers and validate the real names of users in this set against a publication database to acquire positive examples. Negative examples are obtained by collecting random users and removing those, which are already contained in the set of positive examples. Based on this dataset, a classifier is trained on features extracted from the users' profiles (e.g., number of followers, keywords) and tweets (e.g., fraction of tweets with URL(s)). We adapt the processing pipeline and ground-truth data collection and present our adapted approach in the subsequent section.

3. APPROACH

We adopt the basic processing pipeline from the just presented approach: First, we collect ground truth data by means of initial seeds and afterwards, we train a classifier on the collected data.

An overview of the ground-truth data collection process is depicted in figure 1. Our initial seed accounts are formed by economic journal[4] accounts instead of conference accounts as reported in [5]. We opted for the former, since in a first probe, only 10 out of 42 conferences maintained a suitable Twitter account. The set of seed accounts is expanded by followers and followees of the seed accounts (users that follow the seed accounts and users that are followed by the seed accounts).

From this candidate set, we extract valid economist accounts by matching their real names against an economic publication database[5]. Accounts (and their tweets respectively) that feature a real name consisting of at least two words and providing a match with at least one publication in the database are considered valid and hence make up the positive examples in our ground-truth dataset. To obtain negative examples also, we manually select Twitter accounts tweeting about preferably diverse topics and in turn expand them by their followers/followees. Again, we match the candidate set against the publication database, while this time, those accounts (and corresponding tweets) that do not match are considered as valid negative examples. Further, we extract follower/following relationships of the candidate accounts, to be used as classification feature.

We collect the users and tweets with *Tweepy*[6] via the *Twitter REST-API*[7]. In addition to our validation filter, we filter out non-English tweets with *langid.py*, a language identification tool, that provides high accuracy even on microblog messages [8]. In a preprocessing step to classification, we remove URLs, user mentions and stopwords and lemmatize the remaining terms with the help of NLTK [1]. For constructing the feature space we use the bag of words approach. The vector space model for tweet terms is built using TF-IDF weighting, for hashtags and followers we use term occurrence. Terms occurring in more than 50% of the documents are removed.

4. EXPERIMENTS

In our experiments, we compared the three different features types, namely tweet content, hash tags and followers and their combinations using three different classifiers, linear Support Vector Machines (SVM)[4], multinomial Naive Bayes (NB) [9], and Decision Trees (DT) [2]. For the evaluation, we randomly selected 15k tweets from each class (economist and non-economist). This sampled data set was split randomly into training and test data for each class, resulting in 11.250 samples for training and 3.750 for test. All classifiers were trained on the training data set, evaluation measures are reported for the test data set.

4.1 Data Set

The Twitter data was retrieved in summer 2014 and consists of around 1.3 million tweets from March 2006 to August 2014, crawled from almost 870k user accounts. Table 1 provides an overview of the data set.[8]

Table 1: Data set overview.

Total	
Accounts	868.534
Tweets	1.296.651
Users with real name	309.004
After Validation	
Positive accounts	4.300
Negative accounts	4.340
Positive tweets	401.392
Negative tweets	285.283

[4] we based our search for journal accounts on the list available at `http://en.wikipedia.org/w/index.php?title=List_of_economics_journals&oldid=596742762`, resulting in 29 accounts

[5] `https://www.econbiz.de/`

[6] `http://code.google.com/p/tweepy`

[7] `https://dev.twitter.com/docs/api/1.1`

[8] The dataset is available from `http://purl.org/eexcess/datasets/tugd`

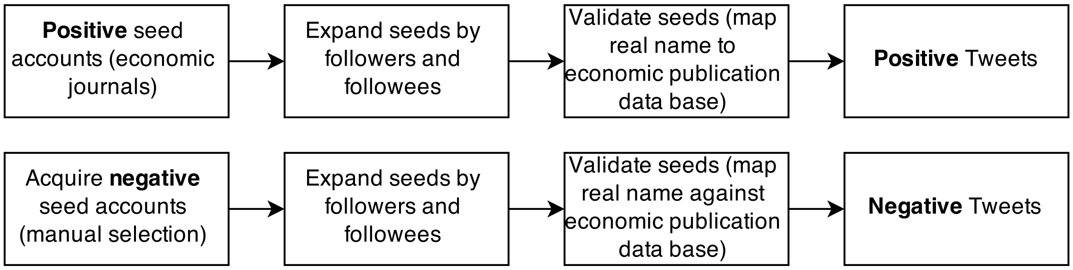
Positive seed accounts (economic journals)
Expand seeds by followers and followees
Validate seeds (map real name to economic publication data base)
Positive Tweets
Acquire **negative** seed accounts (manual selection)
Expand seeds by followers and followees
Validate seeds (map real name against economic publication data base)
Negative Tweets

Figure 1: Overview of the ground-truth data collection process

About 87.25% of all tweets are in English language, indicating that the seeds were adequately chosen for retrieving tweets in mainly one language. Only 20% of the user accounts (174.627 out of 868.534) had at least one tweet, the remaining user accounts were followers with no collected tweets. 30.1% of every user with a valid real name was not found on EconBiz (and resulted in being in the non-economist group). But 99.6% of all users within the group of economists had a valid real name field. This verifies the assumption from [5] that a specified real name is a good indicator for a scientist.

4.2 Results

Table 2 shows an overview of the results for different classifiers and feature combinations. For single features (tweet terms, hash tags or followers), the tweet terms provide most information for classification, with the NB and SVM both outperforming the DT classifier in terms of F1 value (0.705 for NB and SVM). NB completely fails to classify based on hash tags alone, precision is 1.0 while recall is ≈ 0 meaning that nearly no user was classified as economist independent of whether it should have been or not. Hash tags provide no additional information to the tweet terms, the $F1$ measures are the same as with tweet terms alone ($F1 = 0.705$ for NB and SVM). Combining followers and tweet terms leads to a slight decrease in accuracy ($F1 = 0.697$ for NB) indicating that followers introduce noise rather than more information for the classification task. Also the combination of all three features results in lower accuracy ($F1 = 0.699$ for NB) than classification on tweet terms only ($F1 = 0.705$ for NB).

From the experiments we can conclude that the best features for classifying economists are TF-IDF weighted tweet terms. In terms of classifiers, NB and SVM yield the same accuracy and outperform the DT on the best feature (tweet terms). The results suggest that classification of economists can be done on the basis of a single tweet ("was this tweet written by an economist?") without the need to crawl additional information like tweet history or user profile information, which makes it applicable for usage with the Twitter streaming API.

5. CONCLUSION AND FUTURE WORK

In this paper we propose an approach to detect economists on Twitter, based on individual tweets with high accuracy ($F1 = 0.705$). Ground truth data was acquired using manually curated lists of seed accounts, expanding the seeds to their followers and validating the accounts against a database of economic publications. Our feature analysis shows that the best accuracy is already attained by using tweet content only, neither hash tags nor follower information improved the classification model.

Although the accuracy is lower as reported for computer scientists ($F1 = 0.94$) [5] economist classification can be performed on the basis of individual tweets removing the need to obtain tweet history data and user profile information. This can be used on the Twitter streaming API to detect candidate tweets (and therefore user accounts) from the economic domain, which can then be further verified with an extended feature set.

6. ACKNOWLEDGMENTS

The presented work was developed within the EEXCESS project funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement number 600601.

7. REFERENCES

[1] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* O'Reilly, Beijing, 2009.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth International Group, Belmont, CA, 1984.

[3] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh. Wtf: The who to follow service at twitter. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 505–514, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[4] I. Guyon, B. E. Boser, and V. Vapnik. Automatic capacity tuning of very large vc-dimension classifiers. In *Advances in Neural Information Processing Systems 5, [NIPS Conference]*, pages 147–155, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[5] A. T. Hadgu and R. Jäschke. Identifying and analyzing researchers on twitter. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 23–32, New York, NY, USA, 2014. ACM.

[6] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth. User interests identification on twitter using a hierarchical knowledge base. In V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, and A. Tordai, editors, *The Semantic Web: Trends and Challenges*, volume 8465 of *Lecture Notes in Computer Science*, pages 99–113. Springer International Publishing, 2014.

Table 2: Overview of classifier performance for single features and feature combinations. Showing precision, recall and F1 measure on test data set. Best measures are marked bold.

Feature	Classifier	Precision	Recall	F1
Tweet terms	Naive Bayes	0.718	0.693	**0.705**
	Linear SVM	0.722	0.688	**0.705**
	Decision Tree	0.654	0.629	0.641
Hash tags	Naive Bayes	**1.000**	0.001	0.002
	Linear SVM	0.501	1.000	0.667
	Decision Tree	0.501	1.000	0.667
Followers	Naive Bayes	0.992	0.190	0.320
	Linear SVM	0.992	0.190	0.320
	Decision Tree	0.992	0.190	0.320
Tweet terms + Hash tags	Naive Bayes	0.718	0.639	**0.705**
	Linear SVM	0.723	0.687	**0.705**
	Decision Tree	0.653	0.595	0.623
Tweet terms + Followers	Naive Bayes	0.552	0.945	0.697
	Linear SVM	0.816	0.327	0.511
	Decision Tree	0.723	0.303	0.427
Tweet terms	Naive Bayes	0.551	**0.955**	0.699
+ Followers	Linear SVM	0.815	0.374	0.512
+ Hash tags	Decision Tree	0.651	0.278	0.390

[7] S. M. Kywe, E.-P. Lim, and F. Zhu. A survey of recommender systems in twitter. In *Proceedings of the 4th International Conference on Social Informatics*, SocInfo'12, pages 420–433, Berlin, Heidelberg, 2012. Springer-Verlag.

[8] M. Lui and T. Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[9] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[10] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: A first look. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, AND '10, pages 73–80, New York, NY, USA, 2010. ACM.

[11] K. Tao, F. Abel, Q. Gao, and G.-J. Houben. Tums: Twitter-based user modeling service. In R. García-Castro, D. Fensel, and G. Antoniou, editors, *The Semantic Web: ESWC 2011 Workshops*, volume 7117 of *Lecture Notes in Computer Science*, pages 269–283. Springer Berlin Heidelberg, 2012.