Spring 5-29-2015

Privacy on Reddit? Towards Large-scale User Classification

Benjamin Fabian
Humboldt University Berlin, bfabian@wiwi.hu-berlin.de

Annika Baumann
Humboldt University Berlin, Annika.Baumann@wiwi.hu-berlin.de

Marian Keil
Humboldt University Berlin, keil.marian@gmail.com

ISBN 978-3-00-050284-2

PRIVACY ON REDDIT? TOWARDS LARGE-SCALE USER CLASSIFICATION

Complete Research

Fabian, Benjamin, Humboldt University Berlin, Germany, bfabian@wiwi.hu-berlin.de

Baumann, Annika, Humboldt University Berlin, Germany, annika.baumann@wiwi.hu-berlin.de

Keil, Marian, Humboldt University Berlin, Germany, keil.marian@gmail.com

Abstract

Reddit is a social news website that aims to provide user privacy by encouraging them to use pseudonyms and refraining from any kind of personal data collection. However, users are often not aware of possibilities to indirectly gather a lot of information about them by analyzing their contributions and behaviour on this site. In order to investigate the feasibility of large-scale user classification with respect to the attributes social gender and citizenship this article provides and evaluates several data mining techniques. First, a large text corpus is collected from Reddit and annotations are derived using lexical rules. Then, a discriminative approach on classification using support vector machines is undertaken and extended by using topics generated by a latent Dirichlet allocation as features. Based on supervised latent Dirichlet allocation, a new generative model is drafted and implemented that captures Reddit's specific structure of organizing information exchange. Finally, the presented techniques for user classification are evaluated and compared in terms of classification performance as well as time efficiency. Our results indicate that large-scale user classification on Reddit is feasible, which may raise privacy concerns among its community.

Keywords: Privacy, Social Media, Reddit, User Classification, Machine Learning

1 Introduction

The popularity of social media and social networking has risen continuously since its first appearance. Users generate massive amounts of content on Facebook, Twitter, and similar social networking sites as well as on blogs, video sharing platforms, etc. Many companies are searching for opportunities in analyzing social media entries (Kaplan and Haenlein, 2010). This ranges from automatic processing of product reviews and opinions especially with the social components of online retailers (Oelke et al., 2009; Popescu and Etzioni, 2007) to predictions of the financial market (Ferguson et al., 2009) and even mining companies' market structures (Netzer et al., 2012). Also social sciences increasingly use the information contained in social websites. Especially political movements such as the Arab Spring (Lotan et al., 2011) or Occupy Wall Street (Tremayne, 2014) have been studied with the help of content analysis of social media.

Many of these analyses require knowledge of certain demographics such as the origin of the authors of messages. However, many users might begin to share less information about their profiles as anonymity and privacy are gaining more attention during the last years. In particular, the so called NSA leak in which highly confidential documents about online user surveillance by the US

government were unveiled by Edward Snowden in 2013, increased the awareness that any information shared on the Internet could be abused by third parties. Therefore, many users search for opportunities to express themselves on social platforms without revealing their identity.

Reddit[1] is a bulletin board system where users can share news and posts from funny to controversial (Koh, 2013). When founded in 2005 it only had content in form of links or text entries and an up-and-down voting system for submissions. Soon a comment function was added that increased social interaction of users. Reddit pays a lot of attention to users' privacy, granting them the choice of disclosing none of their information except for the posts and comments they provide. This complicates large-scale content analyses compared to other social media platforms. So far, no research has investigated how well user privacy on Reddit is protected against sophisticated approaches to identify latent personal attributes.

In order to address this research gap, this article is dedicated to classifying users on Reddit based on their comments regarding their social gender as well as the more comprehensive classification into citizenship. These two characteristics have been chosen for analysis since research has shown that both have an important influence on user behaviour and user perception in the online context. Gender is a highly relevant factor in the area of marketing and e-commerce since men and women tend to behave rather differently (e.g., Slyke et al., 2010; Rodgers and Harris, 2003). Furthermore, based on the famous cultural dimensions model of Hofstede (Hofstede, 1984) research detected that citizenship influences, for example, the trust level of a user towards e-commerce activities (Gefen and Heart, 2006). In particular, we focus on answering the question whether it is possible to classify users on Reddit simply based on the post and threads they have commented on.

Because of its supposed anonymity, using Reddit for content analysis is not yet widely applied. Two potential examples of usage possibility will be given in the following to motivate user classification in Reddit. First, Reddit contains many discussions on consumer products that can provide insights on the various segments of customers. Demographic information supports the customer analysis. Second, discussions on political issues are also quite frequent. Information on the origin of users might help to understand and follow the discourses and opinions presented.

The remainder of our article is structured as follows. First, we present a review of related work on classifying users and extracting latent information in social media. Then, our research method and a newly extracted text corpus are presented. Afterwards, a selection of state of the art classifiers is introduced and their performance is evaluated. Finally, we discuss limitations and present conclusions.

2 Related Work

User classification, and in particular identifying differences between genders in online and offline communication, has been a well-studied subject (Baumann et al., 2015). Already in the early 1970's, researchers made increasing use of statistics on speech patterns to find differences between the two genders (Lakoff, 1972; Trudgill, 1972) and to analyze women's position in society. With the emergence of electronic communication such as email and Internet chat, research also shifted focus to these areas. Herring (1996) analyzed gender differences in terms of emails and later on also in Internet chat conversations (Herring, 2000; Panyametheekul and Herring, 2003).

Schler et al. (2006) presented findings on gender and age differences between bloggers and developed a classifier which could predict gender at 80.1% and age of three classes at 43.8% accuracy. They used the (multi-class) Real Winnow classification algorithm on features of blogs such as word

[1] URL: http://www.reddit.com

frequencies, categories of words as well as writing style. Their corpus contained information of 37,000 bloggers. Gender classification on bloggers was also performed by Yan and Yan (2006) using a naive Bayes classifier. They relied on features of a bag-of-word approach based on the content and additionally included the choice of font, punctuation marks as well as emoticons. For their corpus of 3,000 bloggers, they reported an F-measure of 68% and showed that it increases monotonically with the corpus size. Mukherjee and Liu (2010) classified the gender of bloggers achieving a decent accuracy of 88.56% by using *support vector machines* (SVM) regression on a well-chosen set of features. They selected and graded them using an ensemble feature selection algorithm that ranks the shares of all features on the overall performance by applying measures such as cross-entropy or mutual information.

Rao et al. (2010) build a scientific Twitter corpus to execute user classification. The authors used binary SVM for classification of gender, age, regional origin, and political affinity. They applied it to three kinds of features: socio-linguistic characteristics, unigrams and bigrams of tweets and a combination of these. For gender, the classifier showed a prediction accuracy of 72.3%, for age of 74.1%, for regional origin of 77.1% and for political orientation of 82.8%. Pennacchiotti and Popescu (2011a,b) also focused on user classification on Twitter. They looked at political orientation, affinity for a specific business (Starbuck's), and African-American ethnicity. They applied Gradient Boost Decision Trees to features such as profile information, sentiment of tweets and user behaviour. Furthermore, they used topics of latent Dirichlet allocation (LDA) as features, applying two different versions: One was trained on the set of all users, the second one was only trained on a domain-specific training set (e.g., only users with labels of political affinity when classifying Democrats or Republicans). Their results for the classification case of ethnicity are rather imprecise at an F-measure at 65.5% whereas Democrats could be predicted at 91.5% and Republicans at 84.0%. Fans of Starbuck's were identified with an F-measure of 76.1%. Burger et al. (2011) managed to discriminate gender on Twitter at a comparably high accuracy of 92.0%. They chose Winnow2 as their classification algorithm. Their feature set consisted of word and character *ngrams* of the tweet contents, user profile information, user screen names, and user full names. Compared to a manual classification task using the Amazon Mechanical Turk, the automatic classification performed with a higher accuracy.

Facebook has also been the interest of gender discrimination in research as Wang et al. (2013) looked at one million random English status updates and applied LDA topic modeling, while identifying 25 topics which they labeled. With separation of users over 25 and under 25 years they could show differences in occurrence frequency of each topic for the different genders.

3 Research Method

Though English is the dominant language on Reddit, it would be useful to classify users independent of the language they are writing in. This certainly excludes the application of full text analysis to extract features for classification. But as a social network, Reddit also offers non-textual elements that can be exploited, in particular the structure of Subreddits and Reddits together with the comments and (dis)likes of users. Our main research question is therefore the following: Is it possible to automatically extract latent user attributes such as gender and citizenship from Reddit by classifying users based on their comments? Our corresponding method essentially consists of five process steps that are illustrated in Figure 1 which will be explained in more detail in the following.

Figure 1. Process of user classification on Reddit.

To gather a text corpus, we extracted all comments for several users. The selected usernames were provided by a publicly available list which was constructed by collecting 660,464 comments of users in November 2013 posted within nine days (Redditor twentythree-nineteen, 2013). Given this user list, the complete set of comments and votes for each user was then collected using the Reddit API (Reddit Inc., 2014). By this procedure we retrieved comments for 76,767 users. It has to be noted that with this approach the user names were not drawn independently from the list of all registered users on Reddit because there is no method to access random users without knowing their names. For further processing, the corpus was preprocessed to ensure good data quality. For this, we excluded all users without any comments. Although the corpus has originally been sampled from names of users who have commented, users are able to delete their comments and votes.

The next step consisted of pre-tagging the gathered corpus. Supervised learning requires a training set in order to make predictions. The input data for training needs to consist of the prediction's input (the information the prediction is based on) as well as the outcome (target) that the prediction should generate based on the input data. The outcome or label for each data entry is the user's classified attribute. For the purpose of this article, social gender with the label classes "female" and "male" and citizenship which is grouped by continent into "Africa", "Asia", "Australia and Pacific", "Europe", "North America", and "South and Middle America" are the labels of choice.

Since the Reddit corpus does not provide these annotations and we could not identify any available Reddit corpus that provides additional user information, we derive and use probable labels. They might be not a 100 % correct representation of real Reddit user attributes. To derive more accurate labels, manual tagging should be applied which can be based on asking users explicitly as well as looking individually at each of their comments. For the automatic derivation of labels, no machine learning should be used in order to preserve independence of labeling and prediction. Pennacchiotti and Popescu (2011a) introduced an algorithm for Twitter user classification based on regular expressions. They proposed the following pattern as a simple method to extract users' age and ethnicity from the user's comments: "`(I|i)(m|am|´m)[0-9]+(yo|year old)white(man|woman|boy|girl)`" (Pennacchiotti and Popescu, 2011a, p. 282).

For our purpose of extracting users' citizenships, the labeling is based on regular expressions and extracts country names and demonyms using the "List of adjectival and demonymic forms for countries and nations" of Wikipedia (Wikipedia, 2014). For gender, a similar approach is applied suggested by Rao et al. (2010) which focuses on finding the possessive "my" in comments together with a clearly identifying terms, e.g., "husband", "girlfriend", or "hubby".

Since quotations can easily distort the label procedure, we use a feature of the Reddit API which provides a recommended style for quoting inside comments and enables easy removal of the quotations. It might be the case that users that do not adhere to that rule may be potentially misclassified. In order to reduce the number of false positives (for example, a woman being classified as "male"), all ambiguous users have not been included in the pre-tagged corpus (8 for citizenship; 122 for gender).

The final numbers of users for each label can be found in Table 1. A survey conducted on Reddit users (Reddit Inc., 2011) suggests similar relative numbers for gender (female: 18.8%, male: 81.1%) although a more recent survey based on Reddit's Internet traffic does not confirm these numbers instead suggesting a less strong difference (female: 36.3%, male: 63.7%). Either the traffic is not indicative for active users or Reddit became more popular among women within the time between the survey and the Internet traffic acquisition.

Gender	Users	Share
Male	19,991	78.5 %
Female	5,474	21.5 %
Total	25,465	100.0 %

Citizenship	Users	Share
Europe	5,613	37.4 %
North America	5,584	37.2 %
Asia	1,669	11.1 %
Australia and Pacific	1,173	7.8 %
South and Middle America	713	4.8 %
Africa	260	1.7 %
Total	15,012	100.0 %

Table 1. Number and relative share of tagged users by gender and citizenship.

A comparison to the 2011 survey (Reddit Inc., 2011) does not reveal similar relative numbers in terms of citizenship. Instead, North America is with 73.6% the strongest continent and Europe is with 17.2% only on the second position far behind (Pacific: 4.6%, Asia: 2.5%, South America: 1.6%, Africa: 0.4%). Recent reports provided by Alexa.com support the dominance of North American users above two-thirds of overall site visitors (Alexa Internet, 2014). One reason for the difference could be that particularly non-North American users will state their origin to set themselves apart since Reddit is a North American dominated website. Therefore, the labeled corpus of citizenship cannot be regarded as representative as a random sample.

Once the users are labeled in terms of gender and citizenship concerning the test set, a bag-of-words is created which is a representation of a document that disregards sequential order of its words. It can be considered as a table of words and the count each word appears inside this document. For our classification, users are treated as documents and Reddits or Subreddits as words. Therefore, the bag-of-(Sub)Reddits considers the number of all (Sub)Reddits a user commented on.

The actual user classification is the final step. It is conducted by using classification algorithms and evaluating their performance on the given test corpus. All the previous steps are providing the necessary input of the bag-of-(Sub)Reddits and the users' attribute labels. The user classification only considers one attribute at a time and does not include any statistical dependencies between the two attributes of gender and citizenship.

4 Classifiers

In our study, two families of state-of-the art classifiers were adopted: weighted soft-margin SVM classifiers and supervised LDA. SVM are a group of regression and classification algorithms that base on the idea of introducing a margin around the decision boundary to achieve a higher amount of generalization (Bishop et al., 2006, p.362). The support vector classification is a linear binary classifier and can therefore only discriminate between two classes. Its classification boundary is a separating hyper plane. A soft-margin permits a classification error in the training data to a certain

degree and usually facilitates a better generalization (Bishop et al., 2006, p.331). When looking at the data of our corpus (Table 1), both attributes show strong imbalances among the classes. To tackle the disproportion, a weighted SVM was introduced (Osuna et al., 1997).

The second classifier family is related to generative models, which are statistical models that try to capture an underlying causal process (Bishop et al., 2006, p.366). This allows a modeling of real world problems and derivation of statistical solutions by inferring latent information. For the case of user classification, it is possible to model the process of generation of comments by users. For example, a user chooses a topic to write about and based on this she selects a (Sub)Reddit to comment. Thereby, the topic is bound to the author's personality or attributes such as gender or origin that influences the choice of (Sub)Reddits. Therefore, we are trying to utilize this causality to classify a user.

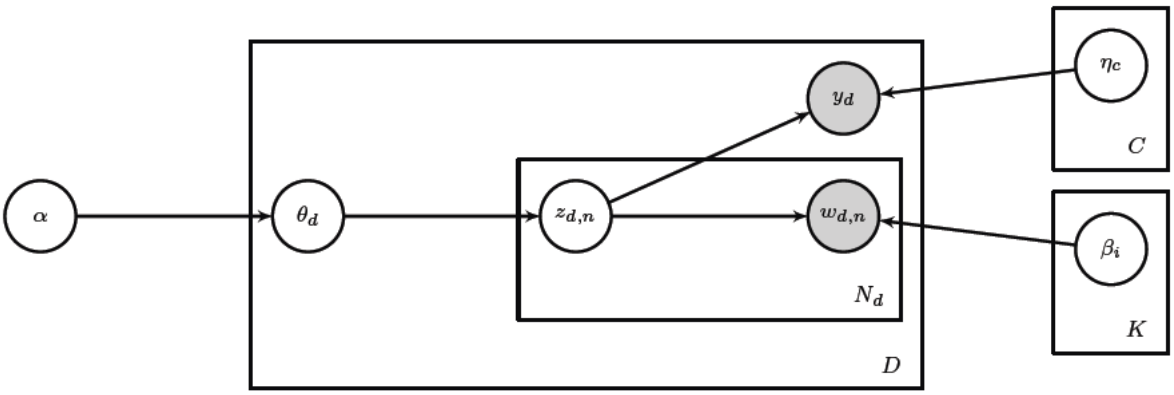

Figure 2. Plate notation of supervised latent Dirichlet allocation based on McAuliffe and Blei (2007). The outer plate (rectangular box) represents documents. The inner plate shows the choice of topics and words within a document.

The generation of topics to exploit hidden information of a document was first applied by Furnas et al. (1988) by introducing latent semantic indexing which is in machine learning usually called latent semantic analysis (LSA). Its algorithm is relying on the usage of singular value decomposition of the term-document matrix and reducing it to a term-topic matrix and a document-topic matrix. Each document therefore has a certain "amount" (including zero) of every topic which describes its intrinsic latent structure. This idea was later extended to a statistically grounded version: probabilistic latent semantic analysis (pLSA). Latent Dirichlet allocation (LDA) is an extension of pLSA that basically adds a Dirichlet prior to the topic choice and the topics' word emissions. As LDA is an unsupervised learning algorithm, the generated topics could be orthogonal to the user attributes that should be identified (e.g., topics that distinguish age groups while the classifier would discriminate gender). It would be possible to use the topics generated by LDA as features and learn them using, for example, an SVM classifier (Pennacchiotti and Popescu, 2011a,b).

But due to LDA's unsupervised nature it withholds the possibility to control the topic creation and therefore the ability to generate topics that – when learned – minimize the prediction error. Therefore it is useful to have an algorithm that supports supervision.

We adopt a model called *supervised latent Dirichlet allocation* (sLDA) (McAuliffe and Blei, 2007; Figure 2) which works as follows: Each topic is "emitting" certain terms as a multinomial distribution. Its prior is the Dirichlet-distributed random variable β_i which defines the probabilities for each term for each topic, and whose values are affected by its prior/parameter α_β. For each document, a mixture of topics is selected by sampling θ as a symmetric Dirichlet distribution parameterized by α_θ. The topic mix is then used to choose one topic for each word in the current document by sampling

$z_{d,n}$ from the multinomial distribution with prior θ. Based on the chosen topic, a term inside the vocabulary is selected by the multinomial distribution $w_{d,n}$. The sLDA's model is very similar to that of LDA; only the additional random variables *Y* for each document with an additional parameter η for each possible label class are added. The distinct number of classes is *C* as used in the generative process for sLDA (McAuliffe and Blei, 2007):

1. Draw word probabilities (for each topic *i*): $\beta_i|\alpha_\beta \sim Dir(\alpha_\beta)$
2. For each document *d* in [1, *D*]:
   a. Draw topic proportion $\theta_d|\alpha_\theta \sim Dir(\alpha_\theta)$.
   b. For each term *n* in [1, N_d]:
      i. Draw topic assignment: $z_{d,n}|\theta_d \sim Mult(\theta_d)$. Actually, this is a categorical random variable. But for simplicity all categorical distributions will be regarded as one-trial multinomial random variables. This has the advantage that vectors of the form (0, 0, ..., 1, ..., 0, 0), with one dimension set to 1 and the others to 0, can be used instead of a scalar value.
      ii. Draw word: $w_{d,n}|z_{d,n}, \beta \sim Mult(\beta_{z_{d,n}})$.
   c. Draw response variable $y_d|z_{d,1:N}, \eta \sim Mult(\eta^T \overline{z_d})$ where $\overline{z_d} = \frac{1}{N}\sum_{n=1}^{N} z_n$ . .

In the original paper (McAuliffe and Blei, 2007) the response type is only restricted to a generalized linear model (GLM) and they based specific implementation on a normal distribution. But for the Reddit user classification a discrete response type is more appropriate. Therefore we focus on a multinomial variable y_d that regresses on $\overline{z_d}$.

5 Results

As performance indicators for classification of gender, *receiver operating characteristic* (ROC) and *area under the curve* (AUC) are chosen since they are advantageous over F-measure and accuracy for imbalanced data classes (Fawcett, 2006). For ROC analysis, a graph is plotted as point pairs of the false positive (FP) rate and true positive (TP) rate over a varying cut-off threshold of class probability for each data point of the test dataset. Thus, the threshold walks from zero to one and, depending on the probability of each test data point belonging to a respective class, the FP rate and TP rates change. A FP in case of the gender attribute might be a male user who was falsely classified as female or the other way around, whereas a TP indicates a correctly classified user in terms of gender.

AUC is a good measure to summarize an ROC curve by a single number, the definite integral from zero to one. A perfect classifier has an AUC value of 1.0 (or 100%), a random algorithm would score with an AUC of approximately 0.5 (or 50%). Therefore the larger the area under the ROC curve is, the better the classifier performs in the sense that an increase in the number of TPs does not lead to a likewise increase of the FP rate.

ROC analysis has the drawback that it only regards two-class classification problems. But for citizenship, six classes need to be considered. One solution would be to create several ROC plots for each class against all other classes (Fawcett, 2006). But due to its high-dimensional surface, it is visually intractable. Therefore we will not consider multi-class ROC but will instead focus on multi-class AUC. Hand and Till (2001) suggested a multi-class AUC that is based on two-class AUC over all possible pairs of classes including their counterparts (class pair *(X, Y)* as well as *(Y, X)*). They calculate the overall AUC as:

$$A_{total} = \frac{1}{C(C-1)} \sum_{i \neq j} A_{i,j} \qquad (1)$$

Accordingly, performance of citizenship classification will be only depicted using AUC.

The following results were all acquired using a 6-fold cross-validation on the whole corpus of labeled gender and citizenship. Furthermore, as cross-validation is applied to prevent over-fitting, more than one ROC curve will be generated for each classification scenario. These will be combined using an algorithm called threshold averaging: A sample of the thresholds of all ROC curves is drawn. For each of these thresholds, one point from each ROC curve whose original threshold lied closest to the current threshold are taken and averaged to one ROC curve (Fawcett, 2006).

Overall, the two main classifiers we proposed, weighted soft-margin SVM classifier and supervised LDA, will be evaluated in the following. Furthermore, in order to gain better insights on the effect of topic models, the same SVM classifier will be applied to the output topics of unsupervised LDA. As performance indicators ROC (for binary classification), AUC, and also classification time duration will be considered to propose a model that can be used in future applications of user classification in Reddit.

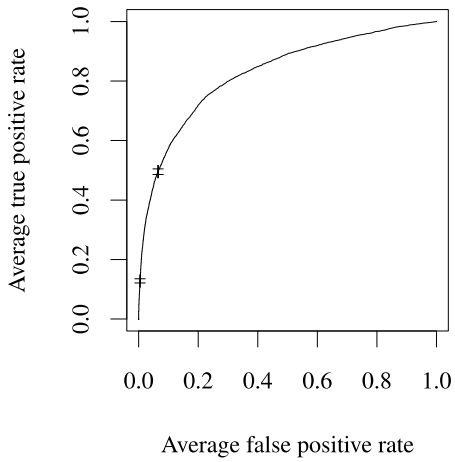

Figure 3. ROC curve of SVM classification of gender without use of a kernel, based on Subreddits as features. The AUC of 82.9% is at a high level. The plot contains intervals on the graph to indicate the maximum deviation from the average across the folds.

5.1 SVM Classifier

For the weighted soft-margin SVM classifier, one version with a radial base function (RBF) kernel and one without any kernel were chosen for comparison. The classification using the RBF kernel did not perform any better than a random classifier, as its AUC did not surpass the 50% threshold. This is consistent over the various versions of feature sets such as Reddits and Subreddits. Even the "simpler" binary classification of gender performed only at 50% AUC. Without the use of a kernel, classification of gender using Subreddits showed useful results with AUC of 82.9% (Figure 3).

5.2 SVM on LDA Topics

Although the SVM classifier already provided reliable results for the classification of gender, it might be possible to increase the performance even further by using other classifiers or different features. Instead of solely relying on the information provided by Subreddits, one can create topics with LDA and learn these with SVM classification. This approach is similar to the one Pennacchiotti and Popescu (2011a,b) applied for classification of political affiliation. Again, SVM classification using

an RBF kernel did not perform better than random choice. Instead, the linear kernel provided better results.

As LDA creates an explicitly specified number of topics, we varied the number of topics to see how the AUC measure develops. For gender classification the performance depends heavily on the number of topics (Figure 4). With the use of topics it seemed to be even possible to discriminate citizenship inside the data. But these results were not promising as the highest average AUC (across the six folds) of only 53.8% was recorded at 350 topics. For gender, the highpoint was reached also at 350 topics but with a promising AUC of 87.3%. Therefore, an increase in performance was gained by moving from plain Subreddits to topics as features. Additionally the amount of time for training was reduced to a third over the method of SVM classification on Subreddits, making it more time-efficient.

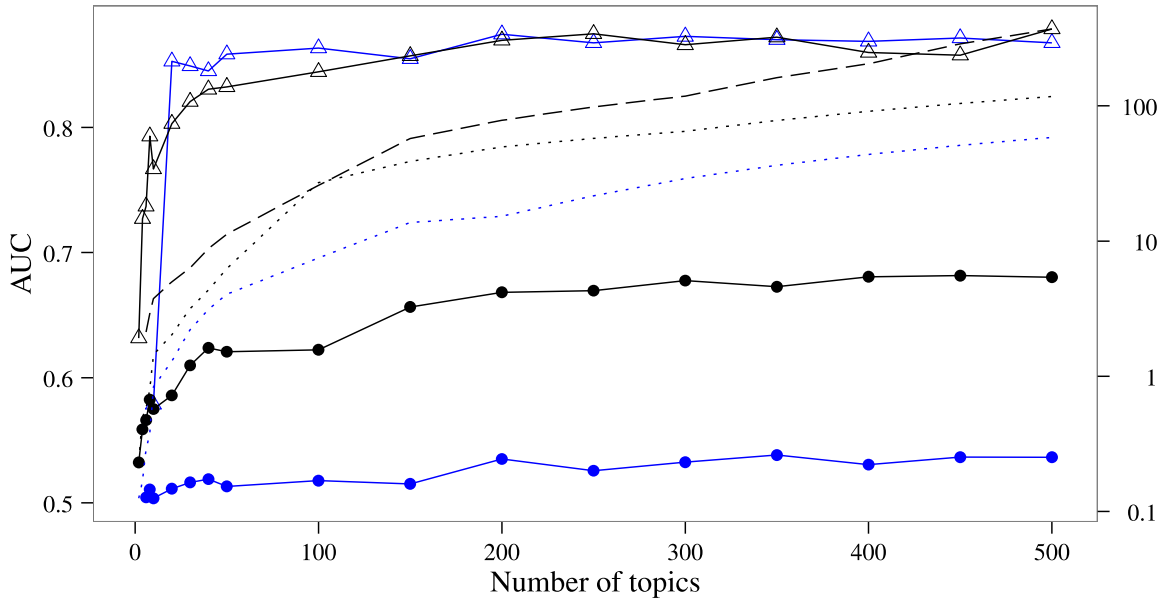

Figure 4. *AUC performance (continuous lines) for SVM-LDA (blue) and sLDA (black) for gender (Δ) and citizenship (•) with dependency on the number of topics. The dotted lines indicate the runtime in hours (with the y-axis on the right side) for gender and the dashed line for citizenship.*

5.3 sLDA

The performance of sLDA also relies on the number of topics. As Figure 4 indicates, the larger the number of topics, the higher the AUC. Due to computational limitations, 500 was the maximum possible number of topics in this evaluation. Future work on this might be able to show whether the graph converges to a certain threshold (as Figure 4 suggests) or whether it reaches a peak and then declines from that point. Therefore, for classification of gender using Subreddits as features, the best performance was observed at 500 topics with an AUC of 87.9% which is also the best result of all previously presented classifiers. Similarly, classification of citizenship with Subreddits as features had its best AUC performance (68.2%) at 450 topics, indicating the best performance of any method to classify citizenship. Classification using Reddits as features did not perform better than a random classifier.

6 Discussion and Outlook

The evaluation showed a slight advantage for the algorithm of sLDA when classifying social gender, This advantage became very obvious when classifying citizenship (Figure 5). However, it should be noted that each of the proposed algorithms depends on several parameters that can influence the performance heavily as shown with the number of topics for sLDA. It may be that a better parameter configuration for the SVM classifier might result in a useful classification algorithm.

For comparison of the individual performances of the algorithms, Figure 4 also provides information on time duration. In this respect, the advantages of sLDA for gender classification become negligible. The LDA-SVM combination reaches a high level of AUC (85.3%) already at 20 topics while being 56% faster as its counterpart of sLDA at the same number of topics. But a comparable number of topics for sLDA with the same AUC is only achieved at 150 topics (85.7%) while taking 6,300% longer. For citizenship classification, the difference in AUC between LDA-SVM and sLDA are too big to consider any training and prediction times. Classification of citizenship using sLDA takes approximately three times longer than classification of gender using sLDA.

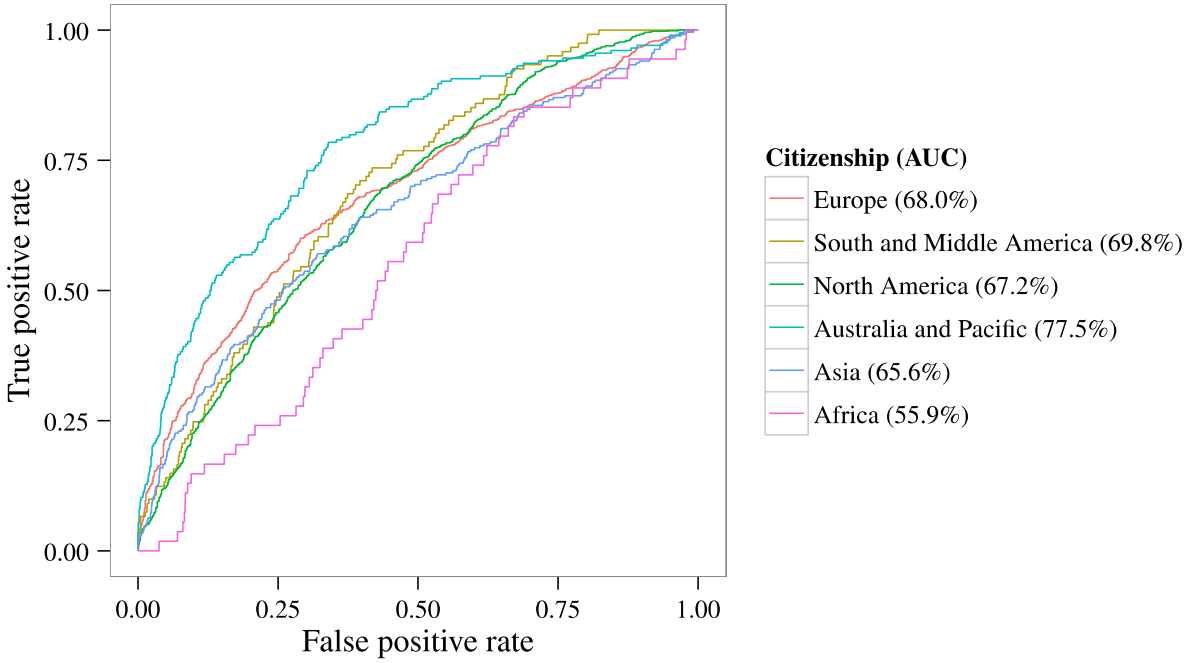

Figure 5. *ROC graphs for all classes of citizenship classified using sLDA with 500 topics on Subreddits. The ROC curves were acquired by plotting each class with the one-against-the-others method. This illustrates to what extent it is possible to identify continental citizenship of a user.*

There are some limitations of the Reddit corpus used in our study. Especially pre-tagging could involve some inconsistencies as firstly it only captures information on users that disclosed such information like being married to a "woman"/"man" or being "French" citizen. Secondly, this information – opposed to the approach of language-independent user classification of this article – only considers English comments. Creating a bigger amount of regular expressions in several languages could be one method that would decrease this inconsistency. Still the best approach might be to use manual tagging done by humans, although (Burger et al., 2011) showed that their classification algorithm could even classify with higher accuracy than humans.

Concerning limitations of classification models, the time-efficiency of sLDA is quite low compared to SVM classification or LDA combined with SVM classification. But for classifying citizenship it is advisable to use sLDA as the performance was significantly higher than with SVM or the LDA-SVM combination. One more time-efficient solution could be another Bayesian inference method called belief propagation which is supposed to be faster while also generating more accurate results than variational inference and also Gibbs sampling as claimed by (Zeng et al. 2013).

As support vector classification showed quite a high reliability for the prediction of gender, a combination of sLDA and SVM in one model could be a useful solution. It would – similar to sLDA – direct the topic generation and – similar to SVM – try to create a maximum margin between the classes. But this approach seems to be only helpful for classification of gender. For citizenship maximum-margin classification alone did not prove to be useful. One reason might be that it only "emulates" multi-class classification by using several binary classifications.

7 Conclusion

Our results show that it is possible to automatically classify gender and also citizenship of users on Reddit and therefore infringe their privacy, most probably without their awareness. This classification was performed based on users' comments but without using full text analysis, thereby keeping the procedure language-independent at least to some extent. We also introduced a process of data extraction and pre-tagging to acquire a Reddit corpus which can be utilized for supervised learning. The presented methods conferred findings that approximately one third of the original extracted users could be annotated with the respective gender, and one fourth with citizenship divided into continental groups. Subsequently, SVM classification was applied. To improve classification performance, LDA was used as a generative model that attempts to extract intrinsic information from the user corpus by summarizing Subreddits or Reddits into topics. These topics could then be applied as features for SVM classification. Furthermore, sLDA was suggested as a version of LDA that directs the topic generation and also allows label prediction based on these topics.

Performance of these classifiers was evaluated using the previously acquired corpora of annotated gender and citizenship. As both corpora contained (heavily) skewed class distributions (e.g., less female than male users), the ROC together with the AUC have been chosen as measures for performance evaluation. Our experiments showed that a plain SVM without a kernel could classify gender with a high AUC performance. But for classification of citizenship, SVM did not prove to be an acceptable model. When SVM classification was based on topics that have been generated by LDA, the performance increased slightly, however, time-efficiency decreased drastically. Overall, sLDA performed best for gender and citizenship compared to the other models. Future work could include the investigation of further latent attributes and possibly also finer granularity of the citizenship attribute. But already our current results should increase user awareness of privacy risks when posting on apparently "safe" social media platforms such as Reddit.

References

Alexa Internet (2014). *reddit.com: Site Overview*. URL: http://www.alexa.com/siteinfo/reddit.com/ (visited 11/02/2014).

Baumann, A., Krasnova, H., Veltri, N. F., Ye, Y. (2015): "Men, Women, Microblogging: Where Do We Stand?" In: Thomas. O.; Teuteberg, F. (Hrsg.): *Proceedings der 12. Internationalen Tagung Wirtschaftsinformatik (WI 2015)*, Osnabrück, pp. 857–871

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Burger, J. D. and J. C. Henderson (2006). "An Exploration of Observable Features Related to Blogger Age." In: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 15–20.

Fawcett, T. (2006). "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27 (8), pp. 861–874.

Ferguson, P., N. O'Hare, M. Davy, A. Bermingham, P. Sheridan, C. Gurrin, and A. F. Smeaton (2009). "Exploring the Use of Paragraph-Level Annotations for Sentiment Analysis of Financial Blogs." In: *Workshop on Opinion Mining and Sentiment Analysis (WOMAS 2009)*.

Furnas, G. W., S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum (1988). "Information Retrieval Using a Singular Value Decomposition Model of Latent Semantic Structure." In: *Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 465– 480.

Gefen, D. and Heart, T. H. (2006). "On the Need to Include National Culture as a Central Issue in E-commerce Trust Beliefs." *Journal of Global Information Management* 14 (4), pp. 1–30.

Hand, D. J. and R. J. Till (2001). "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems." *Machine Learning* 45 (2), 171–186.

Herring, S. (1996). "Two Variants of an Electronic Message Schema." *Pragmatics and Beyond New Series*, pp. 81–108.

Herring, S. C. (2000). "Gender Differences in CMC: Findings and Implications." *Computer Professionals for Social Responsibility Journal* 18 (1).

Hofstede, G. (1984). "Cultural Dimensions in Management and Planning." *Asia Pacific Journal of Management* 1 (2), pp. 81–99.

Kaplan, A. M. and M. Haenlein (2010). "Users of the World, Unite! The challenges and opportunities of Social Media." *Business Horizons* 53 (1), pp. 59–68.

Koh, M. (2013). *The 15 Best Discussions on Reddit Ever*. Thought Catalog.

Lakoff, R. T. (1972). *Language and Woman's Place*. Cambridge Univ Press.

Lotan, G., E. Graeff, M. Ananny, D. Gaffney, I. Pearce and D. Boyd (2011). "The Revolutions were Tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions." *International Journal of Communication* 5, pp. 1375–1405.

MacKinnon, I. and R. H. Warren (2007). "Age and Geographic Inferences of the LiveJournal Social Network." In: *Statistical Network Analysis: Models, Issues, and New Directions*. Springer, pp. 176–178.

McAuliffe, J. D. and D. M. Blei (2007). "Supervised Topic Models." In: *Advances in Neural Information Processing Systems* 20, pp. 121–128.

Mukherjee, A. and B. Liu (2010). "Improving Gender Classification of Blog Authors." In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 207–217.

Netzer, O., R. Feldman, J. Goldenberg, and M. Fresko (2012). "Mine Your Own Business: Market-structure surveillance through text mining." *Marketing Science* 31 (3), pp. 521–543.

Oelke, D., M. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L Haug, and H. Janetzko (2009). "Visual Opinion Analysis of Customer Feedback Data." In: *IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*. IEEE, pp. 187–194.

Osuna, E., R. Freund, and F. Girosi (1997). *Support Vector Machines: Training and Applications*. Technical Report. Massachusetts Institute of Technology.

Panyametheekul, S. and S. C. Herring (2003). "Gender and Turn Allocation in a Thai Chat Room." *Journal of Computer-Mediated Communication* 9 (1).

Pennacchiotti, M. and A.-M. Popescu (2011a). "A Machine Learning Approach to Twitter User Classification." *ICWSM* 11, pp. 281–288.

Pennacchiotti, M. and A.-M. Popescu (2011b). "Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter." In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 430–438.

Popescu, A.-M. and O. Etzioni (2007). "Extracting Product Features and Opinions from Reviews." In: *Natural Language Processing and Text Mining*. Springer, pp. 9–28.

Rao, D., D. Yarowsky, A. Shreevats, and M. Gupta (2010). "Classifying Latent User Attributes in Twitter." In: *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*. ACM, pp. 37–44.

Reddit Inc. (2011). *Who in the World is reddit? Results are in...* URL: http://www.redditblog.com/2011/09/who-in-world-is-reddit-results-are-in.html/ (visited on 11/02/2014)

Reddit Inc. (2014). *Reddit API Documentation*. URL: http://www.reddit.com/dev/api (visited on 11/02/2014).

Redditor twentythree-nineteen (2013). *I downloaded 600,000 Reddit comments over a week*. URL: http://www.reddit.com/r/datasets/comments/1r76wp/i_downloaded_600000_reddit_comments_over_a_week/ (visited on 11/02/2014).

Rodgers, S. and Harris, M. A. (2003). "Gender and E-commerce: An exploratory study." *Journal of Advertising Research* 43 (03), pp. 322–329.

Schler, J., M. Koppel, S. Argamon, and J. W. Pennebaker (2006). "Effects of Age and Gender on Blogging." In: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* 6, pp. 199–205.

Slyke, C. V., Bélanger, F., Johnson, R. D. and Hightower, R. (2010). "Gender-based Differences in Consumer E-commerce Adoption." *Communications of the Association for Information Systems*, Vol. 26, Article 2.

Tremayne, M. (2014). "Anatomy of Protest in the Digital Era: A network analysis of Twitter and Occupy Wall Street." *Social Movement Studies* 13 (1), pp. 110–126.

Trudgill, P. (1972). "Sex, Covert Prestige and Linguistic Change in the Urban British English of Norwich." *Language in Society* 1 (02), pp. 179–195.

Wang, Y.-C., M. Burke, and R. E. Kraut (2013). "Gender, Topic, and Audience Response: An analysis of user-generated content on Facebook." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 31–34.

Wikipedia (2014). *List of adjectival and demonymic forms for countries and nations*. URL: http://en.wikipedia.org/wiki/Adjectivals_and_demonyms_for_countries_and_nations/ (visited on 11/02/2014).

Yan, X. and L. Yan (2006). "Gender Classification of Weblog Authors." In: *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 228–230.

Zeng, J., W. K. Cheung, and J. Liu (2013). "Learning Topic Models by Belief Propagation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (5), pp. 1121–1134.