Evaluating topic-based representations for author profiling in social media

Miguel A. Álvarez-Carmona[1], A. Pastor López-Monroy[1],
Manuel Montes-y-Gómez[1], Luis Villaseñor-Pineda[1], and Ivan Meza[2]

Instituto Nacional de Astrofísica, Óptica y Eletrónica (INAOE)[1]
LabTL, Computer Science Department
Luis Enrique Erro No. 1, C.P. 72840, Tonantzintla, Puebla, México
Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas (IIMAS)[2]
Universidad Nacional Autonoma de Mexico (UNAM)
Escolar 3000, Ciudad Universitaria, Ciudad de México, D.F.

Abstract. The Author Profiling (AP) task aims to determine specific demographic characteristics such as gender and age, by analyzing the language usage in groups of authors. Notwithstanding the recent advances in AP, this is still an unsolved problem, especially in the case of social media domains. According to the literature most of the work has been devoted to the analysis of useful textual features. The most prominent ones are those related with *content* and *style*. In spite of the success of using jointly both kinds of features, most of the authors agree in that content features are much more relevant than style, which suggest that some profiling aspects, like age or gender could be determined only by observing the thematic interests, concerns, moods, or others words related to events of daily life. Additionally, most of the research only uses traditional representations such as the BoW, rather than other more sophisticated representations to harness the content features. In this regard, this paper aims at evaluating the usefulness of some *topic-based representations* for the AP task. We mainly consider a representation based on Latent Semantic Analysis (LSA), which automatically discovers the topics from a given document collection, and a simplified version of the Linguistic Inquiry and Word Count (LIWC), which consists of 41 features representing manually predefined thematic categories. We report promising results in several corpora showing the effectiveness of the evaluated topic-based representations for AP in social media.

1 Introduction

The Author Profiling (AP) task aims to analyze written documents to extract relevant demographic information from their authors [14]. The following problems have gained interest recently: gender prediction [2, 31], age estimation [23, 24], personality detection [33], native language identification [2], and political orientation detection [25]. The AP task has a wide range of practical applications. For example, in marketing, companies may leverage online reviews to improve

targeted advertising, and in forensics, the linguistic profile of authors could be used as valuable additional evidence. In this paper we are interested in profiling age and gender from authors of social media domains. Social media documents are difficult to analyze by standard text mining methods because of several challenging characteristics such as spelling-grammar errors and out-of-vocabulary terms[1].

The AP task has mainly approached as a single-labeled classification problem, where the different profiles (e.g., *males* vs. *females*, or *teenager* vs. *young* vs. *old*) stand for the target classes. The common processing pipeline is as follows: i) extracting textual features, ii) representing documents by these features, and iii) learning a classification model of documents. The extraction of textual features is the stage that has received more attention. In this direction, two kind of attributes stand out from others: content features (i.e., nouns, verbs and adjectives), and style features (i.e., function words, punctuation marks, emoticons and POS tags) [23, 31]. In AP tasks, content and style features are extracted by observing words usage to reveal people interests and writing style. In spite of the success of using jointly both kind of attributes, a number of authors have reported results suggesting that content features are the most valuable for AP [19, 27]. This can be explained by the fact that people from the same demographic group tend to share interests, concerns, hobbies and opinions [22, 29].

In this work, rather than define a suitable set of features for AP, we focus on studying the informative value of content features. More importantly, unlike other works using standard representations like BoW, in this work we propose using topic-based representations to better exploit the content information. Our hypothesis is that by using content features in conjunction with topic-based representations, it is possible to obtain comparable results than other more elaborated strategies from the state of the art. A second contribution of this paper is the evaluation of two different approaches for computing the topic-based representations. The first approach consists in automatically compute topic-based features by means of Latent Semantic Analysis (LSA) [4]. Although LSA has been preciously used in several text mining problems, to the best of our knowledge this is the first time it is fully evaluated on pure content features for the AP task [2]. The second approach builds the topic-based representation by considering a set of hand-crafted content features. For this, we devise a simplified version of Linguistic Inquiry and Word Count (LIWC) [34], which consists of 41 predefined topic categories. Each LIWC category contain a number of associated words, which were defined by a group of socio-linguistic experts. In particular, the main contribution of this study consists in exposing the strengths and weaknesses of each topic-based approach over different social media domains.

[1] It is very hard to accurately apply typical procedures like stemming or extract specific syntactic information from informal documents.

[2] In AP tasks, several authors have used LSA as part of elaborated strategies involving different kinds of features, for example: ensemble strategies, or fusion strategies [21]. Nevertheless, they have not reported experimental results to show the real contribution of LSA features.

The evaluation was done using the data sets from PAN14 [27]. The obtained results showed that the two kinds of topic-based representations outperformed the standard BoW in most social media domains. Furthermore, using only 41 features, manually or automatically defined, they obtained competitive results to state of the art methods.

This paper is organized as follows: Section 2 presents some relevant work for this research. Section 3 explains the textual features we used and the considered topic-based representations. Section 4 explains the experimental settings, and then, Section 5 shows the evaluation results. Finally, Section 6 presents our conclusions and some future work directions.

2 Related work

The AP task has been approached from different areas, including psychology [26], linguistics [11], socio-linguistics [5], and natural language processing (NLP) [14, 31]. In this section we review the related work from the NLP perspective. Mainly, we focus on describing the *content* and *stylistic* features that have been employed.

According to the literature, a wide range of different approaches have been proposed for the AP task. The different methods for learning specific textual patterns range from simple lexical approaches to elaborated strategies requiring syntactic/semantic analysis of the documents. For example, the bag of words (BoW) [14] have been successfully used for gender prediction in formal documents. Another example are Probabilistic Context-Free Grammars (PCFG) [30] and language models, which have been designed for gender detection in scientific articles [3]. Likewise, other authors have gone beyond by exploiting latent biographic attributes (e.g., gender, native language), with the aim of analyzing the discourse style between people of the same/different age-gender [9]. Notwithstanding the usefulness of these features for profile prediction, most of them are only relevant for domains having formal documents (i.e., books, articles, etc.), and they remain unexplored in informal domains, such as the case of social media sources. For example, the building process of a PCFG involves the extraction of part-of-speech (POS) tags, which are difficult to accurately extract from social media texts.

In the case of social media, the majority of the works have focused on using *content* and *stylistic* features [18, 27, 28]. Moreover, several works suggest that content words usually are much more relevant than style features. For example, an analysis of information gain presented in [31], showed that the most relevant attributes for gender prediction are those related with content words, for example: *linux* and *office* for discriminating males, whereas *love* and *shopping* for discriminating females. Furthermore, Schler et al. (2006) also concluded that syntactic features are less useful than very basic lexical thematic features when analyzing blogs. Other works have also considered interesting stylistic features, namely slang vocabulary and the average sentence length, but in all the cases

these features have been used in combination –as a complement– of content features [1, 10].

In this work, we attempt to evaluate the relevance of content features for the task of AP in social media. Our main hypothesis is that content features, which capture the topics of interests of users, are the cornerstone to reveal profiling cues in social media domains. In particular, we propose modeling this content features by means of two different topic-based representations: LSA [15], which automatically extracts the topics from the given document collection, and LIWC [34], which is a set of manually defined topics. These two topic-based representations have been previously used in AP [12, 20, 36], but always in combination with other features and strategies, making it impossible to observe its real relevance to the AP task.

3 Features

The main idea behind this paper is that topic-based representations are effective in capturing the content –thematic– information of documents, and therefore that they could be appropriate for the task of AP in social media domains. As mentioned before, we consider two ways of representing the topics from social media profiles. First, we use a set of automatically extracted topics discovered by means of the LSA algorithm [6], and secondly, a set of manually defined topics obtained from the LIWC resource [34]. In the following subsections we describe both approaches.

3.1 LSA

Latent Semantic Analysis (LSA) is a method for representing the contextual-usage meaning of words. It assumes that words close in meaning tend to occur in similar contexts [16], and therefore, uses occurrence and co-occurrence information to associate words and to measure their contribution to automatically generated concepts (topics) [15].

LSA is a method to extract and represent the meaning of the words and documents. LSA is built from a matrix $\mathbf{M}$ where m_{ij} is typically represented by the TFIDF [35] of the word i in document j. LSA uses the Singular Value Decomposition (SVD) to decompose $\mathbf{M}$ as follows:

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{1}$$

Where the $\mathbf{\Sigma}$ values are called the singular values and $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors respectively. $\mathbf{U}$ and $\mathbf{V}$ contain a reduced dimensional representation of words and documents respectively. $\mathbf{U}$ and $\mathbf{V}$ emphasize the strongest relationships and remove the noise [16]. In other words, it makes the best possible reconstruction of the $\mathbf{M}$ matrix with the less possible information [17]. In this work we compute $\mathbf{U}$ and $\mathbf{V}$ from the given training documents as described in [37].

3.2 LIWC

The way that the Linguistic Inquiry and Word Count (LIWC) works is fairly intuitive. Basically, it reads a given text and counts the percentage of words associated with a set of manually defined categories. Given that LIWC categories were developed by researchers from cognitive psychology, they were created with the aim of capturing people's social and psychological states [13], which have proved to be useful in the AP task [8, 24, 32].

LIWC has two types of categories; the first kind captures the style of the author by considering features like the POS frequency or the length of the used words. The second group captures content information by counting the frequency of words related with some thematic categories such as family, work, friends and others. In this research we focused on the content information, and consequently we decided ignoring the style categories. In particular, we considered the 41 thematic categories, each of them described by a name and a set related words. Table 1 lists the 41 LIWC categories, and Table 2 shows some example words associated to the categories of family, work, body, religion and friends.

Table 1: The 41 LIWC content categories

relativity	feel	money	causation	insight
humans	discrepancy	sad	anger	see
affect	home	work	sexual	negative emotion
death	family	tentative	religion	verbs
quant	achievement	health	body	perception
assent	positive emotion	time	leisure	inhibition
hear	friends	anxiety	cognitive	certainty
space	motion	swear	social	biological
ingestion				

Table 2: Examples of five LIWC categories: name of categories and a subset of associated words

Category	Subset of associated words			
Family	uncle	granddad	mommy	son
Work	sector	commerce	feedback	corps
Body	thigh	flesh	cornea	hands
Religion	amish	pope	rabbi	zen
Friends	comrad	sweetheart	mates	roomate

3.3 Corpora

For the experiments we used the datasets from the PAN 2014 AP task. These corpora were especially built to study the AP in social media domains. They consist of two gender profiles (female vs. male) and five non-overlapping age profiles (18–24, 25–34, 35–49, 50–64, 65-plus). All document collections are in English and they belong to four different domains: Blogs, Social Media, Hotel Reviews, Twitter [27]. Tables 3 and 4 describe the distribution of profiles for the different domains for the gender and age classes respectively. It is important to notice that gender classes are balanced, whereas age classes are highly unbalanced.

Table 3: Distribution of the gender classes across the different domains.

Class	Blogs	Reviews	Social Media	Twitter
Female	73	2080	3873	153
Male	74	2080	3873	153
Σ	147	4160	7746	306

Table 4: Distribution of the age classes across the different domains.

Class	Blogs	Reviews	Social Media	Twitter
18-24	6	360	1550	20
25-34	60	1000	2098	88
35-49	54	1000	2246	130
50-64	23	1000	1838	60
65+	4	800	14	8
Σ	147	4160	7746	306

4 Experimental settings

In this section we describe the configuration used in all the experiments.

Preprocessing: First we removed stop words, then we extracted content words and applied stemming on them. Finally, we considered the 5000 most frequent terms for each domain.

Text representation: For building the LIWC representation we considered the 41 thematic categories shown in Table 1. For the LSA representation we set the parameter k to 41 in order to be able to compare its results against those using the LIWC topics.

Classification: In all the experiments we used the LibLINEAR classifier [7] and performed a stratified 10 cross fold validation (10-CFV). As a baseline we used the results from the BoW representation considering the 5000 selected words.

5 Results

The goal of the following experiments is two fold: first, to determine the effectiveness of topic-based representations, namely LSA and LIWC, for AP in social media, and second, to compare their performance with the traditional BoW representation as well as with one state of the art (BSoA) approach. In particular, we used the results reported in [19] as BSoA results. This work uses a combination of content and style features and representation based on automatically discovered subprofiles.

5.1 Age results

Table 5 shows the obtained results. They indicate that the LSA and LIWC based approaches outperform the BoW results in all social media domains. These results allows to conclude that applying a topic-based reprentation is useful for the task of age predecition.

In these experiments LSA obtained the best results for blogs, reviews and social media domains, whereas LIWC obtained the best result for the twitter collection. We presume this may be explained by the great vaiability of topics communicated by a user in their different tweets, which difficults LSA to discover word relations and to extract discriminative topics. On the contrary, LIWC is based on manually defined topics and it is independent from the data. Summarizing, the experimental results show that for highly diverse domais, such as Twitter, it seems a better option to defined the topic representation based on external knowledge.

Table 5: Accuracy results for age classification in four social media domains.

Approach	Blogs	Reviews	Social Media	Twitter
BoW	0.34(±0.10)	0.28(±0.02)	0.32(±0.01)	0.42(±0.05)
LSA	**0.48(±0.09)**	**0.34 (±0.02)**	**0.36(±0.01)**	0.39(±0.06)
LIWC	0.42(±0.26)	0.29(±0.02)	0.34(±0.02)	**0.47(±0.05)**
BSoA	0.48	0.34	0.37	0.48

The results from Table 5 also show that the best results from the topic-based representations are comparable to those from the BoSA method. Given that the BoSA method captures both content and style information, these results allows to observe the importance of content features (thematic interests) for the subtask of age classification in social media domains. Table 6 shows the three topics with the greatest information gain for both, LSA and LIWC. In the case of LSA we list the four most important words associated to each topic. It is interesting to notice that for the blogs collection there are only 2 topics and for Twitter only one. As we explained before, the Twitter collection has a wide range of subjects, and it was difficult for LSA to find relations between the words and to build relevant topics for the AP task.

Table 6: The topics with more information gain for age classification.

Domain	LSA 1	LSA 2	LSA 3	LIWC 1	LIWC 2	LIWC 3
Blogs	thesis memory technology education	tutorial bank market company	_	religion	-	-
Reviews	fantastic wonderful great view	amazing balcony excellent lobby	beach resort lovely pool	affect	cognitive	positive emotion
Social media	boot coach handbag shoes	vuitton louis shoes handbag	smoke cigarett dog nike	cognitive	work	quant
Twitter	fb ow sigir gamif	_	_	assent	swear	certainty

5.2 Gender results

In this Section we show the results for gender classification on four different social domains. Table 7 shows the obtained accuracy results.

As we can see, the BoW representation obtained the best result for the blogs collection; LSA outperformed the BoW in the reviews and social media domains, and LIWC was the best approach in the Twitter corpus. In all domains, the BoSA method obtained the best results, and, furthermore, it considerably outperformed the results from the topic-based representations. We consider this is because style information is possible more relevant for gender classification than for age prediction.

Table 7: Accuracy results for gender classification in four social media domains.

Approach	Blogs	Reviews	Social Media	Twitter
BoW	**0.72(±0.13)**	0.62(±0.02)	**0.52(±0.02)**	0.70(±0.08)
LSA	0.70(±0.10)	**0.65(±0.01)**	**0.52(±0.02)**	0.66(±0.11)
LIWC	0.60(±0.13)	0.62(±0.01)	0.50(±0.01)	**0.71(±0.07)**
BSoA	0.78	0.69	0.55	0.71

Table 8 shows the three topics with the greatest information gain for LSA and LIWC. It is interesting to notice that, such as some previous works have pointed out, the some of the topics that helped mostly to distinguish between women and men are those related to work, home and leisure.

Table 8: The topics with more information gain for gender classification.

Domain	LSA 1	LSA 2	LSA 3	LIWC 1	LIWC 2	LIWC 3
Blogs	love holiday conference system	women diet food eat	tutorial media social inventor	insight	cognitive	work
Reviews	lovely wonder great nice	beach park place york	pool bathroom bed resort	sexual	biological	social
Social media	handbag vuitton louis bag	jersey outlet jordan replica	_	sad	tentative	negative emotion
Twitter	wp seo beso swim	instagram sigir cikm trec	tumblr instagram linkedin vine	work	home	leisure

6 Conclusions

This paper studied the relevance of content features for the author profiling task. It proposed using *topic-based representations* to better capture and exploit the thematic information from the documents. The described experiments mainly focused on evaluating the effectiveness of two topic-based representations, LSA and LIWC, to predict gender and age of users from four different social media domains.

The obtained results provide evidence that topic-based representations outperform the traditional BoW representation. Also, these results are comparable to those from a current state of the art approach, which considers content and style information, indicating that content information is highly informative for the AP task. In particular, content information was very important to predict the age of users from social media domains; in the case of gender classification the results were not as conclusive as in the age classification, showing that style information is possible more relevant for discriminating between men and women.

Regarding the use of LSA and LIWC, the results indicate that topics automatically discovered from the training set are, in most of the cases, a better representation for AP than using a set of manually defined topics. However, for the collections having a small number of training examples and high vocabulary richness, such as Twitter, the best results were obtained using the manually defined topics from LIWC.

Acknowledgment This work was partially supported by CONACYT under scholarships 401887 and 243957, project 247870, and the Thematic Network in Language Technologies, projects 260178 and 271622.

References

1. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Mining the blogosphere: Age, gender and the varieties of self-expression. First Monday 12(9) (2007)
2. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Communications of the ACM 52(2), 119–123 (2009)
3. Bergsma, S., Post, M., Yarowsky, D.: Stylometric analysis of scientific articles. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 327–337. Association for Computational Linguistics (2012)
4. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science 41(6), 391 (1990)
5. Eckert, P.: Age as a sociolinguistic variable. The handbook of sociolinguistics 151, 67 (1997)
6. Evangelopoulos, N.E.: Latent semantic analysis. Wiley Interdisciplinary Reviews: Cognitive Science 4(6), 683–692 (2013)
7. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. The Journal of Machine Learning Research 9, 1871–1874 (2008)
8. Fink, C., Kopecky, J., Morawski, M.: Inferring gender from the content of tweets: A region specific example. In: ICWSM (2012)
9. Garera, N., Yarowsky, D.: Modeling latent biographic attributes in conversational genres. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. vol. 2, pp. 710–718. Association for Computational Linguistics (2009)
10. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers age and gender. In: Third International AAAI Conference on Weblogs and Social Media (2009)
11. Holmes, J., Meyerhoff, M.: The handbook of language and gender, vol. 25. John Wiley & Sons (2008)
12. Iqbal, H.R., Ashraf, M.A., Nawab, R.M.A.: Predicting an author's demographics from text using topic modeling approach (2015)
13. Kahn, J.H., Tobin, R.M., Massey, A.E., Anderson, J.A.: Measuring emotional expression with the linguistic inquiry and word count. The American journal of psychology pp. 263–286 (2007)
14. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing 17(4), 401–412 (2002)
15. Landauer, T.K., Dumais, S.T.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review 104(2), 211 (1997)
16. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes 25(2-3), 259–284 (1998)

17. Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W.: Handbook of latent semantic analysis. Psychology Press (2013)
18. López-Monroy, A.P., Montes-y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L.: Using intra-profile information for author profiling. In: CLEF (Working Notes) (2014)
19. López-Monroy, A.P., y Gómez, M.M., Escalante, H.J., Villaseñor-Pineda, L., Stamatatos, E.: Discriminative subprofile-specific representations for author profiling in social media. Knowledge-Based Systems 89, 134 – 147 (2015)
20. McCollister, C., Huang, S., Luo, B.: Building topic models to predict author attributes from twitter messages (2015)
21. Meina, M., Brodzinska, K., Celmer, B., Czokow, M., Patera, M., Pezacki, J., Wilk, M.: Ensemble-based classification for author profiling using various features notebook for pan at clef 2013. In: CLEF (Working Notes) (2013)
22. Newman, M.L., Groom, C.J., Handelman, L.D., Pennebaker, J.W.: Gender differences in language use: An analysis of 14,000 text samples. Discourse Processes 45(3), 211–236 (2008)
23. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: How old do you think i am?: A study of language and age in twitter. In: Seventh International AAAI Conference on Weblogs and Social Media (2013)
24. Nguyen, D., Smith, N.A., Rosé, C.P.: Author age prediction from text using linear regression. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 115–123. Association for Computational Linguistics (2011)
25. Pennacchiotti, M., Popescu, A.M.: Democrats, republicans and starbucks afficionados: User classification in twitter. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 430–438. KDD '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/2020408.2020477
26. Pennebaker, J.W., Stone, L.D.: Words of wisdom: language use over the life span. Journal of personality and social psychology 85(2), 291 (2003)
27. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the author profiling task at PAN 2014. In: CLEF (Online Working Notes/Labs/Workshop). pp. 898–927 (2014)
28. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: Notebook Papers of CLEF 2013 LABs and Workshops, CLEF-2013, Valencia, Spain, September. pp. 23–26 (2013)
29. Rude, S., Gortner, E.M., Pennebaker, J.: Language use of depressed and depression-vulnerable college students. Cognition & Emotion 18(8), 1121–1133 (2004)
30. Sarawgi, R., Gajulapalli, K., Choi, Y.: Gender attribution: tracing stylometric evidence beyond topic and genre. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning. pp. 78–86. Association for Computational Linguistics (2011)
31. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of age and gender on blogging. In: Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. pp. 199–205 (2006)
32. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one 8(9), e73791 (2013)

33. Schwartz, H.A., Eichstaedt, J.C., Dziurzynski, L., Kern, M.L., Blanco, E., Kosinski, M., Stillwell, D., Seligman, M.E., Ungar, L.H.: Toward personality insights from language exploration in social media. In: AAAI Spring Symposium: Analyzing Microtext (2013)
34. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology 29(1), 24–54 (2010)
35. Turney, P.: Mining the web for synonyms: Pmi-ir versus lsa on toefl (2001)
36. Weren, E.R., Kauer, A.U., Mizusaki, L., Moreira, V.P., de Oliveira, J.P.M., Wives, L.K.: Examining multiple features for author profiling. Journal of Information and Data Management 5(3), 266 (2014)
37. Wiemer-Hastings, P., Wiemer-Hastings, K., Graesser, A.: Latent semantic analysis. In: Proceedings of the 16th international joint conference on Artificial intelligence. pp. 1–14. Citeseer (2004)