Exploring the Use of Psycholinguistic Information in Author Profiling

Delia Irazú Hernández Farías(✉), Rosa María Ortega-Mendoza, and Manuel Montes-y-Gómez

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE),
Santa María Tonantzintla, Puebla, Mexico
{dirazuherfa,mortega,mmontesg}@inaoep.mx

Abstract. Identifying profile characteristics of the author of a given text is the aim of the *Author Profiling* (AP) task. In this paper, we explore the use of two well known psycholinguistic dictionaries, the *Linguistic Inquirer and Word Count* and the *General Inquirer*, with the objective to capture relevant information for recognizing the age and gender of the author of a given text. The contribution of this paper is two-fold. Firstly, we introduce the use of General Inquirer in the AP task. Secondly, we propose different text representations based on these dictionaries, which help to analyze their relevance and complementariness to accomplish author profiling. We experiment with benchmark corpora on AP. The obtained results are competitive with state-of-the-art, validating the usefulness of psycholinguistic information for recognizing profile attributes of authors.

Keywords: Author Profiling · Psycholinguistic dictionaries · LIWC · General Inquirer

1 Introduction

Identifying the gender, age, personality, or native language of the people based on their writings is the aim of *Author Profiling* (AP) [6]. This task attempts to analyze texts in order to predict various attributes related to its author. AP has attracted the attention of the research community due to the many applications that can benefit from it, ranging from forensic to marketing methods and tools.

From a computational linguistics perspective, AP has been addressed as a text classification problem. There are many approaches attempting to tackle this task. Some of them use stylistic features such as the bag of words, presence of URLs, punctuation marks, POS-tags labels, etc. [3,10]. Others take advantage of more sophisticated techniques such as topic-based representations [1] and word embeddings [4]. Furthermore, since 2013 each year a shared task[1] dedicated to identify different aspects of author profiling has been organized.

[1] https://pan.webis.de/tasks.html.

J. A. Carrasco-Ochoa et al. (Eds.): MCPR 2019, LNCS 11524, pp. 411–421, 2019.
https://doi.org/10.1007/978-3-030-21077-9_38

From a different perspective, research in related areas such as Sentiment Analysis, Personality Recognition, and Emotion Detection has been taken advantage of lexical resources. For AP, where the use of particular linguistic aspects could shed light on the differences among distinct types of authors, the use of such resources has also shown to be beneficial. We believed that the use of language and also psychological aspects (psycholinguistic characteristics) of people are involved in their writings, which can be studied to distinguish traits of authors. For example, the way of authors reflect basic emotional and cognitive dimensions reveal cues for recognizing classes of authors. In this context, there is one psycholinguistic resource that has been widely exploited: the *Linguistic Inquirer and Word Count* (hereafter LIWC) [9].

LIWC is a dictionary of words labeled according to different categories covering grammatical and psycholinguistic aspects. It includes more than four thousand words belonging to at least one of 64 categories, which consider, among others, *social processes* (words related to family, friends, etc.), *effective processes* (words associated to positive and negative emotions), *personal concerns* (words related to work, home, leisure, etc.), and *biological processes* (words associated with body, health, ingest, etc.). In AP, information from LIWC categories is commonly used to generate feature vectors [1]. Also, representations based on LIWC have been combined with other lexical resources [5] and with stylistic features [2].

There are other psycholinguistic resources considering different kinds of categories such as the *General Inquirer* [11] (hereafter GI). GI has been already used in various NLP tasks, but never in Author Profiling. It is a dictionary composed by 182 categories[2] developed for analyzing language considering several aspects, ranging from cognitive to emotion-laden words. The categories in this dictionary cover words associated to pleasure and pain, regarding roles and forms of interpersonal relations, and associated to places and locations, among others.

In this paper, we aim to evaluate the performance of both dictionaries when they are used to characterize aspects related to age and gender identification. Thus, the main contributions of this work can be summarized as follows: (*i*) it proposes three representations based on psycholinguistic information for the AP task; (*ii*) it uses for the first time –to the best of our knowledge– the *General Inquirer* lexicon in AP; and (*iii*) it presents a qualitative and quantitative analysis of the kind of information relevant for AP that is captured by these two dictionaries, paying special attention to their differences and similarities.

2 Psycholinguistic-Based Representations for AP

The AP task has been traditionally tackled as a supervised text classification problem, where a classifier is trained to assign predefined author classes to a collection of documents. Recently, the use of psycholinguistic dictionaries, such

[2] A complete list of categories and their description is found in http://www.wjh.harvard.edu/~inquirer/homecat.htm.

as LIWC, has been explored. In this paper, we consider information from LIWC and GI by means of three different representations, as described below.

Let $D = \{d_1, \ldots, d_{|D|}\}$ denote the collection of documents, and $V = \{t_1, \ldots, t_{|V|}\}$ its term vocabulary, where the terms correspond to word n-grams of different sizes. Also, let $C = \{C_1, \ldots, C_{|C|}\}$ represents the set of categories in a given dictionary (e.g. LIWC or GI), where each category is a set of words (lexical unigrams) denoted by $C_f = \{w_1, \ldots, w_{|C_f|}\}$.

Traditional Term-Based Representation. In this representation, each document d_i is modeled by a vector $\mathbf{d_i^w}$:

$$\mathbf{d_i^w} = <v_{i,1}, ..., v_{i,|V|}> \tag{1}$$

where $v_{i,j} = f(d_i, t_j)$ represents the number of occurrences of the term t_j in the document d_i.

Rep 1. Category-Based Representation. This representation exclusively relies on the information provided by the dictionary. Therefore, each document d_i is represented by a vector $\mathbf{d_i^c}$, whose feature space is determined by the categories compressed in the resource:

$$\mathbf{d_i^c} = <v_{i,1}, \ldots, v_{i,|C|}> \tag{2}$$

where $v_{i,j} = \sum_{s=1}^{|C_j|} f(d_i, w_s)$ represents the sum of occurrences of words belonging to category C_j of the dictionary in the document d_i.

Rep 2. Term-Category Based Representation. Term and category based representations are quite different, the former has good coverage but it is ambiguous and imprecise, whereas the latter is the opposite. For taking as much benefit as possible from both of them, we decide to combine them. Let $\mathbf{d_i^w}$ and $\mathbf{d_i^c}$ be the vector representations for a document d_i based in terms and categories respectively, the enriched vector $\mathbf{d_i^e}$ is the result of their concatenation.

$$\mathbf{d_i^e} = \mathbf{d_i^w} \parallel \mathbf{d_i^c} \tag{3}$$

where $\parallel$ indicates the vector concatenation operation. Therefore the dimensionality of the enriched vector $\mathbf{d_i^e}$ corresponds to $|\mathbf{d_i^e}| = |\mathbf{d_i^w}| + |\mathbf{d_i^c}|$.

Rep 3. Category-Masked Term-Based Representation. It consists in transforming the original text by "masking" the words that belong to a certain category in the resource. The masking process is done as follows: each word in the text is replaced by its corresponding category(ies) in a given dictionary. Words out of the dictionary's vocabulary are kept in their same position. Therefore, this representation avoids having redundant information by including the same knowledge more than once in the feature space (i.e., terms and their respective category, as in the previous representation). Following we present an example of a sentence and its masked version.

- Original text: "*Lovely hotel, comfortable room*"
- Masked text[3]: "*social-affect-posemo hotel, affect-posemo space-relativ-home*"

Once the texts are masked, we build their term-based representation. However, in this case there is a new vocabulary $V' = \{t'_1 \dots t'_k\}$, where each t'_j represents a n-gram that may include words and categories. For instance, from our example, the vocabulary will include the unigrams "social-affect-posemo" and "hotel", and also the bigram "social-affect-posemo hotel".

Formally, a document d_i is represented by the enriched vector, $\mathbf{d_i^m}$:

$$\mathbf{d_i^m} = <v_{i,1}, \dots, v_{i,|V'|}> \tag{4}$$

where $v_{i,j} = f(d_i, t'_j)$ represents the number of occurrences of the new term t'_j in the document d_i.

3 Experiments

3.1 Evaluation Datasets

For evaluation purposes, we used the corpora from the *2nd* and *5th International Competitions on Author Profiling*, hereafter *PAN2014* and *PAN2017*, respectively. The PAN2014 corpus includes collections of blogs (*Blogs*), hotel reviews (*Reviews*), tweets (*Tw14*), and social media posts (*SMedia*), which are different kinds of social media data allowing us to assess the proposed approach over distinct domains. On the other hand, the PAN2017 corpus only includes a collection of tweets written in different languages and annotated according to gender. In this paper we only consider the English partition of this dataset (*Tw17*). For the sake of the comparison, we used the same training and test data partitions than in the aforementioned competitions. Table 1 shows the distribution for each label in the used corpora.

3.2 Experimental Settings

We applied a preprocessing process consisting in replacing all urls, Twitter marks (mentions and hashtags), emoticons, and emojis, by a corresponding label. We also coverted all texts to lowercase. Additionally, we lemmatized all words from texts and psycholinguistic dictionaries (LIWC and GI). Once built the representations described in the previous section, we normalized them by applying the L2 norm. Finally, we addressed the AP task as a classification problem by means of a Support Vector Machine. In line with the shared tasks on AP, as well as with most work in the state-of-the-art, we evaluated our approach using the *accuracy* measure.

[3] The sentence in the example was processed with LIWC. The word *Lovely* belongs to three categories: social, affect, and posemo. The word *comfortable* belongs to two categories: affect and posemo. The word *room* belongs to three categories: space, relativ, and home.

Table 1. Data distribution of the Author Profiling corpora.

	Blogs		*Reviews*		*Tw14*		*SMedia*		*Tw17*	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Female	73	39	2,080	821	153	77	3873	1,688	1,800	1,200
Male	74	39	2,080	821	153	77	3873	1,688	1,800	1,200
18–24	6	10	360	148	20	12	1,550	680	-	-
25–34	60	24	1,000	400	88	56	2,098	900	-	-
35–49	54	32	1,000	400	130	58	2,246	980	-	-
50–64	23	10	1,000	400	60	26	1,838	790	-	-
65-xx	4	2	800	294	8	2	14	26	-	-
Total	147	78	4,160	1,642	306	154	7746	3376	3,600	2,400

3.3 Results

Comparing LIWC and GI

The purpose of this experiment is to evaluate the relevance of using psycholinguistic information in the AP task. We decided to take advantage of the *Category-based* representation by exploiting two settings: each dictionary individually (denoted as ***GI*** and ***LIWC***, respectively) and by combining both resources into a single one (denoted as ***GI+LIWC***). The first one allows to evaluate the performance of each resource at its own, while the second one also serves to analyze how complementary the dictionaries are. Table 2 shows the obtained results.

Table 2. Results from the *Category-based representation* (Rep 1).

	Gender					Age			
	Blogs	*Reviews*	*Tw14*	*SMedia*	*Tw17*	*Blogs*	*Reviews*	*Tw14*	*SMedia*
GI	0.602	**0.641**	**0.681**	0.499	0.608	0.333	**0.295**	0.389	0.263
LIWC	0.538	0.632	0.558	0.505	0.623	**0.384**	0.258	**0.402**	0.219
GI+LIWC	**0.628**	0.64	0.668	**0.514**	**0.715**	0.307	**0.295**	0.324	**0.285**

In general, results show that the categories of each dictionary contain words that help to reveal the profile of authors. Regarding the gender classification, GI slightly outperforms LIWC, whereas, for age classification, results indicate that both resources obtained the best performance in two collections. From these results, we can infer that these resources capture psycholinguistic information in a different way, which is highly related to the traits of profiles. For example, several categories of LIWC correspond to popular topics mentioned by people of a certain age range, such as *work*, *past*, and *home*. On the other hand, GI has a greater number of categories than LIWC, thus different dimensions are captured benefiting to the binary problem on gender identification.

Regarding the combination of the dictionaries, our results show that when both resources are used together, there is no a clear advantage with respect to using each dictionary on its own. This indicates that both resources are not complementary, maybe due to the redundancy (or overlap) of the words belonging to their categories. One example of this is the high overlap between the positive and negative effective categories from both dictionaries.

Combining Lexical and Psycholinguistic Information

As shown in the previous experiment, using only information from the dictionaries increases the probability of missing important clues for identifying users' profiles. On the other hand, it has been recognized that lexical features, such as word n-grams, are good discriminators of profiles. Nevertheless, many of them are not covered by the psycholinguistic dictionaries. One example are slang terms, which are very popular is social media texts. In order to take advantage of both kinds of information, the following experiments consider their combination by means of the *Term-category based representation* (referenced as **Rep2**), and the *Category-masked term-based representation* (denoted as **Rep3**). Both representations were instantiated with information from the GI and LIWC dictionaries. Table 3 shows the obtained results. It also shows two baseline results, namely, the results from the *Traditional term-based representation* (**Traditional**), as well as the best result from the *category-based representation* (**Rep1**), when using a single dictionary.

Table 3. Obtained results when combining lexical and psycholinguistic information, using the proposed representations.

	Gender					Age			
	Blogs	*Reviews*	*Tw14*	*SMedia*	*Tw17*	*Blogs*	*Reviews*	*Tw14*	*SMedia*
Traditional	0.576	0.704	0.623	0.530	0.768	0.346	0.315	0.357	0.308
Rep1	0.602	0.641	0.681	0.505	0.623	0.384	0.295	0.402	0.263
Rep2-GI	**0.705**	0.697	**0.681**	0.529	0.729	0.358	0.308	**0.402**	0.325
Rep2-LIWC	0.653	**0.708**	0.597	0.520	0.767	0.333	**0.316**	0.376	0.239
Rep3-GI	0.666	0.660	0.675	0.507	0.726	0.371	0.294	0.402	0.242
Rep3-LIWC	0.602	0.635	0.675	0.522	0.615	0.358	0.283	0.324	**0.335**

The results from Table 3 indicate that the combination of lexical and psycholinguistic information works. In 7 out of 9 collections, this combination outperformed the baseline results. It is also possible to notice that GI obtained slightly better results than LIWC, demonstrating its usefulness for the AP task. This advantage could be caused by its broader coverage of terms used in formal communications such as the ones from social media. Finally, these results show a clear disadvantage of the Rep 3 with respect to Rep 2, confirming the relevant role of lexical information for the task of AP in social media.

Comparison with State of the Art

As mentioned before, for comparison purposes we used the same datasets than in the PAN2014 and PAN2017 shared tasks. In Table 4 we present the obtained results[4]. Concerning to *Blogs* collection, we improved the best performing approach for gender classification. This is an encouraging result because of size of the collection, which represents a great challenge. Overall, the obtained results at the PAN2014 collections are very competitive against those from the shared task, particularly if we consider that the proposed approach is quite simple and straightforward. With respect to the PAN2017 collection (*Tw17*), we ranked on the 12th position, but our result is higher than the average performance of the share task participants. Furthermore, despite the simplicity of our approach, it showed a similar performance than other methods based on novel techniques such as word embeddings and deep learning[5]. For further details on the best ranked systems in the shared task, see [7] and [8] for the *2014* and *2017* editions, respectively.

Table 4. Comparison of the obtained results with the state of the art

Dataset	Gender				Age			
	BestTeam	AvgPerf	OurRes	Rank	BestTeam	AvgPerf	OurRes	Rank
Blogs	0.679	0.567	**0.705**	1	0.461	0.332	0.384	2
Reviews	0.725	0.606	0.708	2	0.35	0.274	0.316	5
Tw14	0.733	0.586	0.681	3	0.506	0.378	0.402	5
SMedia	0.542	0.524	0.529	5	0.365	0.307	0.335	6
Tw17	0.823	0.757	0.767	12	-	-	-	-

4 Analysis

Content Analysis. The purpose of this analysis is to explore the use of words from the different dictionaries' categories regarding to each profile trait. Specifically, we investigated what are the categories mostly used according to a profile group. For each dataset, we grouped the texts according to gender and age. Then,

[4] Table 4 shows the best performing approach in each dataset (in the 2nd column), the average performance of all the participating teams (3rd column), the higher accuracy achieved by our proposal (4th column), as well as the ranking corresponding to our best result (5th column).

[5] Just to mention, three of the methods exploiting such kind of approaches have accuracy rates around 0.78, only 0.02 above the best score achieved in our experiments.

we calculated the frequency of the words included in each category. Finally, we manually selected a subset of the most frequent categories and analyzed their content with respect to the each class.

In general, as it was expected, the categories most frequently used in each dataset comprise words referring to prepositions, pronouns, articles, adverbs, verbs, etc. We also observed that by using either of the dictionaries, it is possible to catch clues related to the use of personal information, that have been recognized as a key feature for AP [6]. Particularly, we observed several categories associated to some particular profiles. Table 5 summarizes the most frequent categories for each of the profile traits in the used datasets, showing some

Table 5. A subset of the most frequent categories used in the AP corpora.

Task	Resource	Category and some words included on it
Female	GI	Afill: *love, son, side, thank, friend, care, team, helpful, friendly*, etc.
		ABS: *need, think, right, idea, learn, reason, holiday, cause, pace*, etc.
	LIWC	percept: *ear, thin, hot, see, look, hear, feel, view, feeling, listen*, etc.
Male	GI	Means: *war, say, make, live, free, ready, hand, job, order, build*, etc.
		Strong: *king, own, win, able, great, make, gain, love, show, game*, etc.
		ECON: *job, business, price, money, tax, project, custom, company*, etc.
	LIWC	quant: *some, every, each, most, best, many, enough, worst, plenty*, etc.
18–24	GI	Overst: *just, great, right, last, always, amazing, full, quite, high*, etc.
		Afill
	LIWC	work: *staff, working, office, report, meeting, publish, interview*, etc.
		percept
25–34	GI	Afill, ECON, and Overst
	LIWC	quant
		achieve: *win, top, goal, lost, effect, gain, success, challenge, effort*, etc.
35–49	GI	ECON
		ECON@: *own, fee, rent, market, shop, social, fill, serve, import*, etc.
	LIWC	work
50–64	LIWC	home: *room, bath, family, garden, kitchen, studio, garage, door*, etc.
		motion: *walk, visit, trip, travel, went, move, arrived, walking, drive*, etc.
65–xx	GI	Afill and ABS
		Underst: *only, small, never, something, suggest, care, nothing*, etc.
	LIWC	home and achieve
		past: *ate, was, did, met, been, stayed, won, made, loved, went, told*, etc.

intuitive and interesting aspects. For example, regarding LIWC, words related to perceptual processes ("*percept*" category), such as *ear*, *thin*, *hair*, *look*, *feel*, and *eye*, are more used by female than by men. Instead, men use more quantifiers. According to GI, female use more words related to supportive ("*Afill*" category) than males. Similarly, terms related to economy (*rent*, *earn*, *shop*, etc.) tend to characterize people within 25–49 age range.

Discriminative Analysis. To deeply understand the contribution of the evaluated dictionaries, the most discriminative attributes were identified. For achieving it, information gain was calculated on the *Term-Category based representation* for each problem in each dataset. Table 6 shows some of the features with the highest information gain per dataset.

As it can be observed, word unigrams emerged as more relevant than bigrams or trigrams. There are some intuitive categories from GI appearing among the most discriminative for gender identification: "*Female*" and "*Male*", both contain words[6] referring to women/male and social roles associated to them. Some categories including words related to negation and negative feelings ("*negate*",

Table 6. Some of features with the highest information gain rate per dataset according to gender and age traits. Words in italic font represent lexical n-grams from Rep 2. Category tags are listed per dictionary.

Dataset	Top discriminative features	
	Gender	Age
Blogs	TERMS: *publish*, *sinc*, and *internet* GI: EnlOth, EnlTot, and Know LIWC: work	TERMS: *wife*, *husband*, and *love* GI: Goal
Reviews	TERMS: *wife*, *husband*, and *love* GI: SklAsth, Our LIWC: sexual, we	TERMS: *amaz* GI: Self, Ovrst, and Strong LIWC: i, pronoun, funct
Tw14	TERMS: *play*, *beat* GI: Know, Male, and Ovrst LIWC: negate, negemo, tentav	TERMS: *me*, *emoticon*, and *haha* GI: Self, NegAff LIWC: swear
SMedia	TERMS: *here*, *live*, and *2012* GI: Tool, IAV LIWC: funct, incl	TERMS: *me*, *individu*, and *repost* GI: WltTot, MeansLw, Econ@ LIWC: quant, achieve
Tw17	TERMS: *love*, *my*, and *emoji* GI: Female, AffOth LIWC: i	

[6] Some terms included in these categories are: *aunt*, *girl*, *lady*, *bride*, *boy*, *dad*, *gentleman*, *son*, etc.

"*negemo*", and "*NegAff*") were identified as very discriminative for age identification. Furthermore, it is possible to observe that there are various categories ("*our*", "*self*", "*i*", and "*we*") reflecting personal pronouns found among the most relevant ones. It is also important to mention that there are some onomatopoeic expressions as well as non verbal elements used in social media for enriching written communication; ("*haha*", "*emoticon*", and "*emoji*") emerged as very discriminant (maybe for identifying young people). Such kinds of terms are hard to be found in dictionaries like LIWC or GI. This points out the relevance of combining lexical and psycholinguistic information for AP.

5 Conclusions

In this paper we assessed the performance of two psycholinguistic dictionaries in the AP task: Linguistic Inquirer and Word Count (LIWC) and General Inquirer (GI). The knowledge in such resources was exploited by three novel text representations attempting to capture psycholinguistic information for distinguishing the age and gender of a given user, by considering only her/his written texts. Several experiments were carried out, demonstrating the usefulness of taking advantage of psycholinguistic dictionaries as well as the viability of the proposed representations for AP. Particularly, this paper introduces the use of GI in AP. The results provide evidence that the categories in this resource allow to wrap peculiarities of users which help to profile classification.

The experimental evaluation showed that the categories from both dictionaries, LIWC and GI, incorporate relevant discriminative information for the AP task. However, we observed that there is not a clear evidence allowing to state than one is better than the other. Besides, it seems that they are not complementary resources. Each one captures information associated to specific traits of profiles (for example, GI outperformed LIWC in the gender problem, whereas the opposite happens in the age case). Finally, according to our findings, it can be stated that the combination of lexical and psycholinguistic information is very relevant for AP.

As future work, it could be interesting to incorporate the information coming from psycholinguistic dictionaries into systems considering other kinds of techniques, such as with deep learning and word embeddings. Furthermore, evaluating the performance of lexical resources available in different languages in a cross-lingual setting for Author Profiling is also matter of future work.

Acknowledgments. This research was funded by CONACYT (project FC 2016-2410 and postdoctoral fellowship CVU-174410).

References

1. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-y-Gómez, M., Villaseñor-Pineda, L., Meza, I.: Evaluating topic-based representations for author profiling in social media. In: Montes-y-Gómez, M., Escalante, H.J., Segura, A., Murillo, J.D. (eds.) IBERAMIA 2016. LNCS (LNAI), vol. 10022, pp. 151–162. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47955-2_13
2. Bartoli, A., De Lorenzo, A., Laderchi, A., Medvet, E., Tarlao, F.: An author profiling approach based on language-dependent content and stylometric features. In: 2015 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (2015)
3. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-GrAM: new groningen author-profiling model. CoRR abs/1707.03764 (2017)
4. Bayot, R.K., Gonçalves, T.: Author profiling using SVMs and word embedding averages. In: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5–8 September 2016, pp. 815–823 (2016)
5. Marquardt, J., et al.: Age and gender identification in social media. In: 2014 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (2014)
6. Ortega-Mendoza, R.M., López-Monroy, A.P., Franco-Arcega, A., Montes-y-Gómez, M.: Emphasizing personal information for author profiling: new approaches for term selection and weighting. Knowl.-Based Syst. **145**, 169–181 (2018)
7. Pardo, F.M.R., et al.: Overview of the author profiling task at PAN 2014. In: Working Notes for CLEF 2014 Conference, pp. 898–927 (2014)
8. Pardo, F.M.R., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at PAN 2017: gender and language variety identification in Twitter. In: Working Notes of CLEF - Conference and Labs of the Evaluation Forum (2017)
9. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic Inquiry and Word Count: LIWC 2001, vol. 71. Lawrence Erlbaum Associates, Mahway (2001)
10. Posadas-Durán, J.P., et al.: Syntactic N-grams as features for the author profiling task: notebook for PAN at CLEF 2015. In: CLEF (2015)
11. Stone, P.J., Hunt, E.B.: A computer approach to content analysis: studies using the general inquirer system. In: Proceedings of the May 21–23, 1963, Spring Joint Computer Conference, AFIPS 1963 (Spring), pp. 241–256. ACM (1963)