Identifying the Authors' National Variety of English in Social Media Texts

Vasiliki Simaki[1,2], Panagiotis Simakis[3], Carita Paradis[1] and Andreas Kerren[2]
[1]Centre for Languages and Literature, Lund University, 221 00 Lund, Sweden
{vasiliki.simaki, carita.paradis}@englund.lu.se
[2]Department of Computer Science, Linnaeus University, 351 95 Växjö, Sweden
andreas.kerren@lnu.se
[3]XPLAIN, 153 42 Athens, Greece
simakis@xplain.com

Abstract

In this paper, we present a study for the identification of authors' national variety of English in texts from social media. In data from Facebook and Twitter, information about the author's social profile is annotated, and the national English variety (US, UK, AUS, CAN, NNS) that each author uses is attributed. We tested four feature types: formal linguistic features, POS features, lexicon-based features related to the different varieties, and data-based features from each English variety. We used various machine learning algorithms for the classification experiments, and we implemented a feature selection process. The classification accuracy achieved, when the 31 highest ranked features were used, was up to 77.32%. The experimental results are evaluated, and the efficacy of the ranked features discussed.

1 Introduction

The spread of social media has been rapid and impressive during the past decade. More and more people use social media on a daily basis and they often choose this channel to express their opinions about various topics such as politics, music, lifestyle, environment, or personal matters. This activity produces a massive number of sound data, images, and text data everyday that needs to be further analysed and grouped according to different criteria that we set in each case. Text data from social media can provide important information about social media users, their preferences, habits and the trends they follow. The identification of authors' sociodemographic and personality information has attracted a great deal of attention in the research community, and numerous studies and methodologies about this task have been proposed.

The identification of sociodemographic information about the social media authors is an interesting task for a number of reasons and contributes to the monitoring of the users' opinions on various topics. This information provides an important input to sociological studies, and at the same time it is indispensable for Market Analysis and e-commerce services. Text Mining and Natural Language Processing are among the scientific fields that benefit from this development. New methods and tools have been proposed, and significant results have been observed in Author Profiling, Language Variety Identification, and other similar tasks. The research activity in these domains is also a result of the significant expansion the past few years of the available resources due to the data and information flow.

The present study can provide useful information to the field of dialectology as well. Studies in this field have observed the different linguistic choices that speakers of different English varieties make at various language levels (morphology, phonology, lexicon, syntax, etc.). The varieties of the English language that we investigate are used by 315 million speakers approximately[1] (225 million speaker in the USA, 55 in the UK, 19.4 in Canada, and 15.6 in Australia). Previous studies in this topic (Schneider, 2007) that have observed different linguistic choices among the various varieties can be evaluated in new data, and new clues about the linguistic attitude of speakers that use different English varieties can be detected.

In this paper, we present a study for the identification of the authors' national variety of the English. The annotation labels used for this study is

[1]According to the information provided on wikipedia: https://en.wikipedia.org/wiki/Varieties_of_English

Proceedings of Recent Advances in Natural Language Processing, pages 671–678,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_086

US for the American English speakers, *UK* for the British English speakers, *AUS* for the Australian English variety and *CAN* for the Canadian English variety. The non-native speakers of the English are annotated with the *NNS* label. This label is attributed according to the information that the authors provide about themselves on their profile pages on social media or other internet sources (their place of birth, and/or the place they were raised). For this study, we used data from both Facebook and Twitter that are annotated with various information about the authors (gender, age, profession) additionally to the authors' national English variety. We extracted four different feature sets: formal linguistic features, Part-of-speech features, lexicon-based features that are related to the different linguistc variety, and data-based features from each English variety. We performed classification experiments by using a set of various machine learning algorithms, and we implemented a feature selection process. After the experimental results, we achieved classification accuracy of 77.32% with the 31 most infromative features and the NaiveBayesMultinomial classifier. The efficacy of the different features used in this study is an interesting finding, which is evaluated and discussed.

2 Related Work

Identifying information about the author of a text has been the subject of various studies in the fields of Text Mining and Natural Language Processing. Researchers approached the problem of the automatic identification of authors' identity and personality information from different angles.

The first studies in this field were about the Authorship Attribution (Stamatatos, 2009; Koppel et al., 2009; Grieve, 2007; Zheng et al., 2006), where researchers used linguistic features to detect authors' identity in texts from literary works, journalism, and other sources. These studies set the research basis in the identification of a text's author, and motivated the investigation of more refined characteristics. Their methodological approach motivated our work, and many features used in these studies, especially in Zheng et al. (2006) were used in this study.

The detection of gender, age, and other clues of the author's personality and language has also attracted a great deal of attention (Argamon et al., 2007a; Cheng et al., 2011; Schler et al., 2006; Argamon et al., 2007b; Peersman et al., 2011; Rangel and Rosso, 2013; Simaki et al., 2015a,b, 2016, 2017; Sboev et al., 2016; Lins and Gonçalves, 2004). These studies investigate one or more sociodemographic factors, and many of them use data from social media. The profiling of the author (Wright and Chin, 2014; Stamatatos et al., 2015; Rangel et al., 2016) is a recent task, and the findings are important for Forensic Linguistics among other disciplines (van de Loo et al., 2016; Zaeem et al., 2017).

Studies in the field of Native Language Identification (NLI) can be considered as relevant to ours, with Koppel et al. (2005) being the first to infer the native language of an author based on texts written in a second language by using various NLP and Second Language Acquisition features. The studies in this topic that followed implemented different methods and characteristics for the identification of the author's native language by using various feature types like syntactic clues and grammars (Wong and Dras, 2011; Wong et al., 2012) or different resources and evaluation techniques (Tetreault et al., 2012). In his doctoral thesis, Malmasi (2016) offers an extensive presentation of the field's literature, and describes his numerous studies, application and evaluative tasks.

Our task is part of the *Language Variety Identification* research topic. Studies in this field aim at labeling texts in a native language with their specific variation. This topic has become quite popular within the NLP community and numerous events have been organized to this end, with the 5th Author Profiling Task at PAN 2017 as the most recent one (Rangel et al., 2017b). In some of the investigations with different languages, the problem of identifying between pairs of similar languages and language variants on sentences from newspaper corpora is adressed (Zampieri et al., 2014; Tan et al., 2014). Lui and Cook (2013) evaluate various approaches to classify documents into Australian, British and Canadian English, including a corpus of tweets. For Spanish, there are various studies in this task, and researchers achieve good results in terms of classification accuracy mostly by using character and word n-gram models as well as POS and morphological information (Maier and Gómez-Rodrıguez, 2014; Franco-Salvador et al., 2015; Rangel et al., 2017a). Other languages, as for instance the Portuguese (Zampieri and Gebre, 2012) and the Ara-

bic (Sadat et al., 2014), have also been investigated in term of their different varieties.

Most of the studies in the above domains share common methodologies and similar features, and tackle the search task mainly as a classification problem, which usually involves machine learning algorithms and classification experiments.

3 Data Description and Methodology

3.1 Data description

For this study, we used a data set of 712,033 posts (13,424,523 words and 89,347,103 characters in total). The posts were extracted from the official Facebook and Twitter profiles of public figures like actors, authors, singers, athletes, politicians, and they were annotated with the author's sociodemographic clues. To extract the data, we used the *Facepager* software (Keyling and Jünger, 2013). The average size of the corpus posts is 125 characters per post, and the topics discussed vary from personal branding, opinions about social and political matters, nature, etc. The corpus was compiled from September to December 2015, and data from 838 different users (535 male and 302 female users) were manually annotated with information about the author's gender, age, professional activity, national variety of the English and any other additional information available such as his/her educational background or professional details. Concerning the author's national English variety, 584 different users are native speakers of the American English (US), 117 of the British English (UK), 21 of the Australian English (AUS), 31 of the Canadian English (CAN), and 84 of the authors are not native speakers of the English language (NNS). The annotation labels were given according to the information that the users provide about themselves in their social media accounts, and in some cases according to the information that Wikipedia[2] entries or other internet sources provide (as most authors are well-known personalities).

3.2 Methodology

In this study, a text classification methodology was followed for the identification of the authors' national variety of English in our data set. For the experiments four feature sets were extracted:

[2] https://en.wikipedia.org/wiki/Main_Page

Formal Features
Frequency of all symbols
Frequency of all punctuation
Frequency of spaces
Frequency of upper case characters
Frequency of alphabetical characters
Frequency of digit characters
Frequency of short words (less than 3 characters)
Total number of word characters
Average word length
Average sentence length/word
Average sentence length /characters
Number of different words
Hapax legomena
Hapax dislegomena
Frequency of each symbol (˜ , @, /, $, %, ˆ , &, *, -, =, +, >, <, _)
Frequency of each punctuation ((,), [,]. —, ,, ;, ?, ., !, :, ', “, ”)

Table 1: The formal features extracted in the data set.

- *formal features*, which are general linguistic characteristics used in a wide set of studies in Text Mining and Author Profiling. This feature set contains basic counts of character frequencies,word and sentence metrics, as Table 1 presents. The formal features are 41 in total.

- *Part-of-Speech (POS) features*, which count the following basic grammatical categories (according to NLTK' POS tags) in the data set: nouns, prepositions, pronouns, adjectives, determinants, verbs, adverbs, conjunctions, interjections and particles. The number of the POS features is 10.

- *lexicon-based features* from slang and national varieties lexicons for the English language (4 features in total). We extracted idiomatic terms from slang and geographical lexicons[3] that had at least one hit in the data set. Some examples are presented in Table 2.

- *data-based features* based in the different forms used by each national variety of English, which are frequent in the data set (5

[3] http://www.manythings.org/slang/
https://www.anglotopia.net/
http://aussie-slang.com/
https://www.fluentland.com/

US	UK	AUS	CAN
cool	taking	mate	click
call	brilliant	legit	rad
eat	fit	togs	flat
kick	throw	grit	hoodie
clip	pants	footy	hosed
cut	wicked	barbie	pissed
con	bloody	arvo	frog
dope	chips	dag	grit
vibes	ace	slab	tad
chicken	sorted	prawn	emo
grand	uni	goon	hammered
jam	chap	wuss	puck
joint	bangers	hydro	randy
cop	gutted	aboriginal	beaver

Table 2: Some of the lexicon-based features extracted in the data set.

features in total). We kept only the forms that were frequent and unique in each national group by eliminating the frequent forms that appeared in more than one national class. We show some examples in Table 3. We observe in Table 3 that many of these characteristics are related to the trends and popular subjects of each national group during the data collection period, which means that these features are corpus-sentitive characteristics, and have to be re-extracted when different resources are used.

To extract these features, we used the NLTK[4] toolkit. For the classification stage, we used a number of different machine learning algorithms, which are well studied and have been used extensively in several text classification tasks. All classifiers are implemented using the WEKA[5] toolkit (Witten et al., 2016). For all algorithms, the free parameters that are not reported were kept in their default values.

4 National Variety Classification Experiments

4.1 Experimental Setup

For the classification experiments of our study, we tested the performance of various machine learning techniques. In particular, we used the following algorithms:

- a multilayer perceptron neural network (MLP),
- a bayesian classifier (NaiveBayesMultinomial),
- a bagging algorithm using decision trees (Bagging),
- a simple decision table majority classifier (DecisionTable),
- a fast decision tree learner (RepTree),
- a tree algorithm that considers K randomly chosen attributes at each node (RandomTree),
- a classifier for building linear logistic regression models(SimpleLogistic),
- various support vector machine classifiers (SVM, SMO, SVM with radial kernel).

The classifiers are implemented using WEKA, and a 10-fold cross validation protocol for each algorithm was followed. The classification accuracy was evaluated in terms of percentages of correctly classified posts. The classification results when all features were used achieved an accuracy up to 73.86% with the Bagging algorithm, as Table 4 shows.

In Table 4, the results of the classification process are presented. The results are tabulated in descending order, from the highest accuracy percentage to the lowest one. We observe that four classifiers achieved classification accuracy above 70%.

4.2 Feature Selection

The large number of the features used in our preliminary experiments, as well as the promising results in terms of classification accuracy, led us to the investigation of the feature informativity.

We performed a feature selection process in order to highlight the most efficient features and/or feature types for the identification of the national variety of English. We used a Relief feature selection algorithm (Kira and Rendell, 1992), which is heuristics-independent, noise-tolerant, robust to feature interactions and it runs in low-order polynomial time. In our case, we used the updated ReliefF algorithm proposed by Koronenko (1994), which improves the reliability of the probability

[4] http://www.nltk.org/
[5] http://www.cs.waikato.ac.nz/ml/weka/

US	UK	AUS	CAN	NNS
trump	jamieol	ambrose	nelly	gric
msnbc	easytolove	trashed	celine	bieniek
pbs	maxipriest	rpmotorsports	btmontreal	jamaica
slumerican	paulmccartney	bala	furtado	fiberboard
fam	recipeoftheday	aussiecycling	celinedion	ineedyourlove
mypinkfriday	ziggy	cyclingaus	avril	nonfiction2015
bitly	gandy	stanleyracing	makeovers	usain
yall	stardust	aussie	abuse	charlize
cmt	whosay	athletics	gaza	reggae
xzibit	amg	keithurban	palestinian	protocol
jukebox	mercedes	canberra	getinspired	iriesocial
gat	wuss	itsstephrice	beerscontemporary	lama
postmodern	itv	tires	sarahstyle	un
hillary	labour	dymocks	adespatie	por

Table 3: The most salient data-based features extracted in the data set.

Classifier	Accuracy
Bagging	**73.86%**
DecisionTable	73.07%
MLP	73.05%
RepTree	72.93%
RandomTree	59.44%
NaiveBayes	30.65%

Table 4: The classification results when all features are tested.

approximation, it is robust to incomplete data, and generalized to multi-class problems. Our dataset was processed by the ReliefF algorithm, implemented using the WEKA machine learning toolkit, and feature ranking scores were estimated. The 31 highest ranked features are presented in Table 5.

In Table 5, the 31 highest ranked features are presented. We observe that all data-based features and three lexicon-based features (only the US lexicon-based feature is not among the most informative ones) are among them. The POS features appear to be particularly important (nine from a set of ten features). From the set of formal features, the characteristics that are related to word and sentence length, punctuation use, and other lexical clues (e.g., hapax and dis legomena, number of different words, number of short words, etc.) that authors of a different English national variety use, appear to be very informative. This list highlights that the main differences among speakers of a different national variety of English are primarily found at lexical and syntactic levels. In a future study, a more descriptive and qualitative analysis of these findings can be an interesting task.

The results of the feature selection process are evaluated and presented in the Subsection below.

4.3 Second Round of Classification Experiments

We performed a second round of classification experiments where only the most informative features were used. The best results achieved are presented in Table 6.

We observed that the best results were achieved when the Bayesian (NaiveBayesMultinomial) algorithm was used. The Bagging algorithm, which achieved the highest classification accuracy when all features were used, is not that effective and achieved a low accuracy (32.74%). The results which show that the feature selection process improved the performance of the classification algorithms are promising. One interesting finding is that the best results with the reduced feature set are achieved with a Bayesian classifier (the same classifier that performed the worst in the first round of experiments). This confirms the fact that Bayesian models suffer from the curse of dimensionality, and that dimensionality reduction helps improving their performance.

5 Conclusion

In this paper, our study of the identification of the author's national variety of English from social media texts is presented. In our data set, which

Ranking	ReliefF Score	Feature
1	0.00231732	NNS data-based
2	0.00230444	AUS data-based
3	0.00229047	upper case char.
4	0.0021488	spaces
5	0.00174613	symbol char.
6	0.00163237	word length
7	0.00127163	alphabetical char.
8	0.00115416	short words
9	0.00113051	punctuation char.
10	0.00106681	CAN data-based
11	0.00106118	UK data-based
12	0.00096111	char. in words
13	0.00083135	digit char.
14	0.00075679	sent. length/char.
15	0.00070453	nouns
16	0.00052172	prepositions
17	0.00047358	pronouns
18	0.00040154	AUS lexicon-based
18	0.00034311	adjectives
20	0.00034107	determinants
21	0.00033963	verbs
22	0.00022508	hapax legomena
23	0.00020172	adverbs
24	0.00019028	different words
25	0.00013848	US data-based
26	0.00012652	conjunctions
27	0.00010855	hapax dislegomena
28	0.00003396	interjections
29	0.00001187	sent. length/words
30	0.00000442	UK lexicon-based
31	0.00000197	CAN lexicon-based

Table 5: The 31 highest ranked features.

is annotated with various sociodemographic variables, we searched for the national variety of English of each author based on the labels US, UK, AUS, CAN, NNS that were attributed to each author/post. For this task, we tested various linguistic, lexicon- and data-based features and we performed a number of classification experiments by implementing various algorithms. We also tested the informativity of the features and we showed that the lexicon- and data-based features, as well as lexical and syntactic-related features can improve the classification accuracy of our experiments. For the 31 most informative features we achieved 77.32% accuracy.

This preliminary work is among the recent studies in Language Variety Identification field that approaches the identification of the author's national variety of English from a NLP perspective. Our results confirm the theoretical work in dialectology, stating that basic differences among English national varieties (in written discourse) can be detected at the level of lexical choices and syntactic patterns. This study can be further expanded, more resources from different sources can be tested, and new methods can be implemented. Also, the feature selection findings can be analysed and used for further qualitative studies. Additionally, the thematic patterns and the trending subjects found in the data of each variety can be analysed for sociological and cultural purposes.

Classifier	Accuracy
NaiveBayesMultinomial	**77.32%**
SVM(radial kernel)	76.02%
SMO	73.45%
MLP	54.92%
SimpleLogistic	41.43%
Bagging	32.74%

Table 6: The classification results for our data set, when the highest ranked features are tested.

Acknowledgments

This research is funded by the StaViCTA project[6], supported by the Swedish Research Council (framework grant the Digitized Society Past, Present, and Future, No. 2012-5659).

References

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2007a. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday* 12(9).

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007b. Stylistic text classification using functional lexical features. *Journal of the Association for Information Science and Technology* 58(6):802–822.

Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. 2011. Author gender identification from text. *Digital Investigation* 8(1):78–88.

Marc Franco-Salvador, Francisco Rangel, Paolo Rosso, Mariona Taulé, and M Antònia Martít. 2015. Language variety identification using distributed representations of words and documents. In *International*

[6] http://cs.lnu.se/stavicta/

Conference of the Cross-Language Evaluation Forum for European Languages. Springer, pages 28–40.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing* 22(3):251–270.

R Keyling and Jakob Jünger. 2013. Facepager (version, fe 3.3). *An application for generic data retrieval through APIs* .

Kenji Kira and Larry A Rendell. 1992. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*. pages 249–256.

Igor Kononenko. 1994. Estimating attributes: analysis and extensions of relief. In *European conference on machine learning*. Springer, pages 171–182.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the Association for Information Science and Technology* 60(1):9–26.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous authors native language. *Intelligence and Security Informatics* pages 41–76.

Rafael Dueire Lins and Paulo Gonçalves. 2004. Automatic language identification of written texts. In *Proceedings of the 2004 ACM symposium on Applied computing*. ACM, pages 1128–1133.

Marco Lui and Paul Cook. 2013. Classifying english documents by national dialect. In *Proceedings of the Australasian Language Technology Association Workshop*. pages 5–15.

Wolfgang Maier and Carlos Gómez-Rodrıguez. 2014. Language variety identification in spanish tweets. In *Proceedings of the EMNLP2014 Workshop: Language Technology for Closely Related Languages and Language Variants (LT4CloseLang 2014)*. pages 25–35.

Shervin Malmasi et al. 2016. Native language identification: explorations and applications .

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. ACM, pages 37–44.

Francisco Rangel, Marc Franco-Salvador, and Paolo Rosso. 2017a. A low dimensionality representation for language variety identification. *arXiv preprint arXiv:1705.10754* .

Francisco Rangel and Paolo Rosso. 2013. Use of language and author profiling: Identification of gender and age. *Natural Language Processing and Cognitive Science* 177.

Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017b. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working Notes Papers of the CLEF* .

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. *Working Notes Papers of the CLEF* .

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of arabic language varieties and dialects in social media. *Proceedings of SocialNLP* page 22.

Aleksandr Sboev, Tatiana Litvinova, Dmitry Gudovskikh, Roman Rybka, and Ivan Moloshnikov. 2016. Machine learning models of text categorization by author gender using topic-independent features. *Procedia Computer Science* 101:135–142.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*. volume 6, pages 199–205.

Edgar W Schneider. 2007. *Postcolonial English: Varieties around the world*. Cambridge University Press.

Vasiliki Simaki, Christina Aravantinou, Iosif Mporas, Marianna Kondyli, and Vasileios Megalooikonomou. 2017. Sociolinguistic features for author gender identification: From qualitative evidence to quantitative analysis. *Journal of Quantitative Linguistics* 24(1):65–84.

Vasiliki Simaki, Christina Aravantinou, Iosif Mporas, and Vasileios Megalooikonomou. 2015a. Automatic estimation of web bloggers age using regression models. In *International Conference on Speech and Computer*. Springer, pages 113–120.

Vasiliki Simaki, Christina Aravantinou, Iosif Mporas, and Vasileios Megalooikonomou. 2015b. Using sociolinguistic inspired features for gender classification of web authors. In *International Conference on Text, Speech, and Dialogue*. Springer, pages 587–594.

Vasiliki Simaki, Iosif Mporas, and Vasileios Megalooikonomou. 2016. Age identification of twitter users: Classification methods and sociolinguistic analysis. In *17th International Conference on Intelligent Text Processing and Computational Linguistics*. Springer Verlag.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology* 60(3):538–556.

Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. 2015. Overview of the pan/clef 2015 evaluation lab. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, pages 518–538.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*. pages 11–15.

Joel R Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *COLING*. pages 2585–2602.

Janneke van de Loo, Guy De Pauw, and Walter Daelemans. 2016. Text-based age and gender prediction for online safety monitoring. *International Journal of Cyber-Security and Digital Forensics (IJCSDF)* 5(1):46–60.

Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1600–1610.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2012. Exploring adaptor grammars for native language identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 699–709.

William R Wright and David N Chin. 2014. Personality profiling from text: introducing part-of-speech n-grams. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, pages 243–253.

Razieh Nokhbeh Zaeem, Monisha Manoharan, Yongpeng Yang, and K Suzanne Barber. 2017. Modeling and analysis of identity threat behaviors through text mining of identity theft stories. *Computers & Security* 65:50–63.

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of portuguese. In *KONVENS2012-The 11th Conference on Natural Language Processing*. Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI), pages 233–237.

Marcos Zampieri, Liling Tan, Nikola Ljubešic, Jörg Tiedemann, and Nikola Ljube. 2014. A report on the dsl shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*. pages 58–67.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the Association for Information Science and Technology* 57(3):378–393.