The Role of n-grams in Firstborns Identification

Gabriela Ramírez-de-la-Rosa[1], Verónica Reyes-Meza[2], Esaú Villatoro-Tello[1], Héctor Jiménez-Salazar[1], Manuel Montes-y-Gómez[3], and Luis Villaseñor-Pineda[3]

[1] Information Technologies Department,
Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa, México
{gramirez,evillatoro,hjimenez}@correo.cua.uam.mx
[2] Psychology Department,
Universidad Popular Autónoma del Estado de Puebla (UPAEP), México.
veronica.reyes@upaep.mx
[3] Language Technologies Lab., Computational Sciences Department,
Instituto Nacional de Astrofísica Óptica y Electrónica (INAOE), México.
{mmontesg,villasen}@ccc.inaoep.mx

Abstract. Psychologists have long theorized about the effects of birth order on intellectual development and verbal abilities. Several studies within the field of psychology have tried to prove such theories, however no concrete evidence has been found. Therefore, in this paper we present an empirical analysis on the pertinence of traditional Author Profiling techniques. Thus, we re-formulate the problem of identifying developed language abilities by firstborns as a classification problem. Particularly we measure the importance of *lexical* and *syntactic* features extracted from a set of 129 speech transcriptions, which were gathered from three minutes length videos. Obtained results indicate that both bag of words n-grams and bag of part-of-speech n-grams are able to provide useful information for accurately characterize the language properties employed by firstborns and later-borns.

Keywords: Lexical Features, Syntactic Information, Text Classification, Author Profiling, Natural Language Processing

1 Introduction

The study of how birth order influence on several aspects of our lives has been an attractive long-established research topic in the field of psychology. According to some recent publications [1,2], more than 2K studies have been performed during the last four decades, aiming at finding why firstborns are normally compared favorably to later-borns. Generally, it is said that birth order plays a key role in the development of a wide variety of acquired abilities, such as cognition and language advantages as well as with the presence of some personality traits.

As part of his personality theory, Adler [3] suggested that the situation into which a person is born, such the family size, sex of siblings and birth order, plays an important role in our personality development. He was able to identify

common characteristics among firstborns, such as their inclination to be more conservative, always follow the rules and submit to authority as well as more motivated to greater achievements than later-borns. However, more recent studies [1,2,4] state that such acquired abilities might be result of several other aspects in addition to the birth order, such as the social context in which a child is raised.

Regarding the birth order, in [1] it is mentioned that firstborns appear to have an early advantage in the development of vocabulary and syntax, but later-borns may have an advantage in the development of conversational skills. In addition, it has been found that later-borns tend to produce a higher number of personal pronouns than firstborns [5]. Nonetheless, most of the collected evidence of a general firstborn advantage in early vocabulary development comes from studies performed in children between 8 and 30 months old [1,5,6]. Consequently, there is no clear evidence with respect to longer term vocabulary differences associated with birth order, *i.e.,* there is no similar studies among teenagers neither in adults.

Traditionally, in order to correlate the acquisition of some (*language*) abilities among firstborns and later-borns, psychologists apply a set of well-defined questionnaires to the subjects that are being analysed [2,3,7]. By means of measuring certain factors such as attitudes and opinions, tastes and interests, studies or work information, personality, etc., it is possible for psychologists draw conclusions regarding the behaviour, abilities or even the personality of firstborns. Although there are some works [5,7,8,9] that measure some language related factors (*e.g.,* vocabulary length, use of pronouns and some morphological cues, as well as some syntactic aspects), these works rely on the answers provided by the subjects in a predefined questionnaire, which is afterwards manually reviewed.

In order to overcome some of the main drawbacks of the current work in the field of psychology, in this paper we present an empirical analysis on the pertinence of classical Author Profiling (AP) techniques for automatically identify firstborns. AP main goal is to obtain as much information as possible from authors by means of analyzing their written text [10]. For instance, in [10,11] authors have shown that it is possible to automatically detect age and gender from text documents. More recent works have also proposed AP methods for detecting personality [12] and leadership orientation [13]. Therefore, in order to apply AP techniques, we *re*-formulate the problem of identifying language developed abilities by firstborns as a classification problem, *i.e.,* once we compute a set of *lexical* features (word n-grams) and *syntactic* features (POS n-grams), we train a classification model to distinguish firstborns from later-borns. Performed experiments among 129 subjects, between 11 and 16 years old, demonstrate that our proposed methodology is able to accurately characterize the language properties employed by firstborns and later-borns.

In this paper we investigate which of these features are more discriminative in the posed task. Consequently, the research questions we aim to answer are: *(i) Is the firstborns vs. later-borns detection a special case of AP?*, *(ii) Are there*

any differences in the use of lexical and syntactic elements between firstborns and later-borns?

The rest of this document is organized as follows. Section 2 present some related work concerning to the author profiling task. Section 3 describes some important aspects of the employed data for performing our experiments, as well as some statistics regarding this corpus. Then, Section 4 describes our followed methodology, experimental setup and obtained results. Next, Section 5 provides a brief discussion over our principal findings. Finally, Section 6 depicts our conclusions and some future work directions.

2 Related work

As we have mentioned before, we face the problem of distinguish language characteristics developed by firstborns and later-borns as an Author Profiling task. In this context, our AP task is approached as a single-labeled classification problem, where the different profiles (*firstborns* vs. *later-borns*) stand for the target classes. It is important to mention that the posed problem is not addressed as a traditional thematic classification task neither as an Authorship Attribution task since we are not interested in the content of the documents nor in the specific writing style of each author. On the contrary, the AP methodology will allow us to model more general sociolinguistic features that apply to groups of authors, revealing how both type of authors use different linguistic features.

Most of the recent work on AP focuses in two particular problems that have gained interest during the last years, *i.e.,* identifying age an gender from writers [10,11]. However, there are more sophisticated problems that have been addressed using similar AP approaches, such as detecting personality traits [12] and leadership identification within social media [13]. At the end, the underlying idea behind AP techniques represent the likelihood of revealing as much as possible information from a given author's text (*e.g.,* age, gender, cultural background, native language, etc.). In general, the AP problem has been approached from different areas, including psychology where the work described in [14] might be considered amongst the most representative. Nevertheless, we will only explain some of the most representative approaches from the Natural Language Processing perspective.

Traditionally, the AP pipeline can be summarized in three steps: *i)* extracting a specific set of features from the text, *ii)* building an appropriate representation (*e.g.,* BoW-like) and, *iii)* building a classification model for the target profiles (classes). It is worth mentioning that second and third steps are usually addressed by means of standard representations (*e.g.,* BoW) and well defined classifiers (*e.g.,* Support Vector Machines, Naïve Bayes, etc.) respectively. Nonetheless, the research community has centered its efforts in the first step, *i.e.,* determining which are the most pertinent set of features that can be extracted from the documents. Accordingly, most of the employed textual features used in AP fall into two general categories: lexical (*e.g.*, content and functional words), and syntactical (*e.g.*, POS tags). One of the first published works in

demonstrating the pertinence of such feature is [10], where a combination of function words and POS based features achieved approximately 80% of accuracy for gender prediction in formally written texts from the National British Corpus.

In spite of the good results shown by [10], its proposed method it is not suitable for more dynamic environments, such as AP in social media (*e.g.,* blogs, chats, etc.). As shown by [11,13] social media represent a totally different matter, and traditional AP settings are no longer acceptable. In order to overcome such limitations, in [11] authors propose a representation for documents that capture discriminative and subprofile-specific information of terms. Under this representation, documents are represented in a low-dimensional (and discriminative) space which is non-sparse, allowing to obtain good results in the task of gender and age detection.

Although we are not proposing a new AP approach, our work differs from many of the previous research in that we want to determine if AP is possible within texts that represent literal transcriptions from spontaneous speech phenomena. Additionally, to the best of our knowledge, evaluating the firstborns advantage in the development of language through AP techniques has never been addressed. Therefore, the main contributions of this work are: *i)* a analysis on the pertinence of using *lexical* and *syntactic* features for the identification of firstborns, and *ii)* the relevance of employing Natural Language Processing techniques within the field of Psychology, particularly for characterizing language use by firstborns.

3 Data Set

The data set used in this research was gathered as part of a master's thesis on Biological Sciences at the Tlaxcala Biology Center of Conduct in the Autonomous University of Tlaxcala [15]. It is important to mention that all participants in the study are teenagers between 11 and 16 years old. Additionally, all participant subjects do not have any twins, half siblings. Their families are formed by more than one member, and none of the subjects reflected any psychiatric or neurological disorder. As part of the followed methodology for gathering the data, we needed an explicit authorization from the legal tutor of the teenagers in order to be part of this experiment, and upon approval by the ethics committee of the Ministry of Health and the Commission for Bioethics at the UNAM (National Autonomous University of Mexico).

As additional information from the analysed subjects, it is worth mentioning that they all belong to a small town in the state of Tlaxcala, Mexico; which is considered by the Secretary of Social Development of Mexico to be below the national average indicators of poverty and social vulnerability. Particularly, 54.7% of the total population are moderate poverty and 6.2% extreme poverty. This is an important indicator, since in the literature there are very few studies (if none) that have evaluated populations with these characteristics.

Even though determining birth order from a teenager might seem like an easy task for a psychologist, in practice it is not. Some factors such as being the child of a second marriage, or being a child that was born after five or more years from its last sibling, might alter the purpose of this type of studies. Therefore, after carefully interviewing all candidate subjects, we preserve the data from only 129 participants, from which 48 were labeled as firstborns and 81 as later-borns. In order to acquire the data we were interested in, *i.e.,* spontaneous speech, all participants were interviewed without previous warning. During the interview, subjects were asked to talk about themselves by a three minutes period. All interviews were recorded and manually transcribed. As we have mentioned before, we use the transcripts as the only input for our performed experiments. Is worth to mention that all the subjects are native Spanish speakers. Table 1 shows some statistics about the data set such as number of words and vocabulary's size.

Table 1. Statistics describing the employed data set. For each profile it is shown the number of words and vocabulary elements. Between parenthesis appears the standard deviation, whilst between brackets there is the minimum and maximum number of words and vocabulary terms respectively.

Profiles	**Subjects**	**Words** (SD) [min, max]	**Vocabulary** (SD) [min, max]
Firstborns	48	274.2 (± 123.9) [56, 521]	120.5 (± 42.4) [42, 205]
Later-borns	81	235.0 (± 114.0) [52, 474]	109.4 (± 42.8) [30, 205]

4 Followed methodology

As we have mentioned before, our main goal is finding out if there is a distinct use of language in teenagers, such that allows to categorize firstborns and later-borns. This hypothesis has been around in the psychological field for several years, therefore, our designed experiments are oriented to provide some evidence regarding such theory, *i.e.,* that language developed abilities by firstborns are different from those developed by later-borns. The hypothesis behind these experiments is that by means of employing AP techniques, using *lexical* and *syntactic* representations, we can generate good classification models for identifying firstborns subjects and later-borns, and therefore implicitly determine the pertinence of these attributes.

In the following subsections we first describe the general experimental setup and then, we describe in detail each of our performed experiments. In the first experiment we attempt to answer if lexical information by itself can separate subjects into our two classes (firstborns vs later-borns). In a second set of experiments we aim at determining if the use of some grammatical categories reveals patterns that differentiate firstborns language's use from later-borns language's use.

4.1 Experimental setup

For all performed experiments we used a BoW-like form of representation using boolean weights. We choose this representation because is appropriate when using few and short documents. In our case we have 129 transcripts with 250 words in average, approximately (see Table 1). To serve our purpose of analyzing the role of lexical and syntactic information, we use words and POS tags respectively, as atomic units for the BoW-like representation. For constructing the classification model we employed the Weka [16] implementation of the well-konwn Naïve Bayes algorithm and as a validation strategy a 10-*fold-cross-validation* technique.

The pre-processing done to each transcripts was simple and consisted on converting the text into lowercases words. We preserve all stopwords given that these tokens convey important structural information that can be captured through both lexical and syntactic features.

A special treatment was given to numbers or references to numbers in the transcripts. Particularly, for our first experiment (analysis of lexical information) we replace all number references with a single tag in order to limit the vocabulary's growth. For our second experiment (employing syntactic information) we intentionally kept the references to numbers since the grammatical category can be different if they refer to an ordinal or cardinal number.

4.2 Lexical features

This first experiment aims at considering only *lexical* information and determining if it is useful for identifying firstborns. To accomplish this goal, we use words as atomic units. As we mentioned before, we use classical author profiling techniques to infer if this set of attributes can be useful in identifying firstborns. Accordingly, we experimented with the traditional bag of words representation, and using words n-grams with $n = 2, ..., 5$. As previous research works has shown, using word n-grams up to $n = 5$ allows the inclusion of contextual information, which might result useful for some AP tasks.

Figure 1 shows the obtained performance by this experiment when using different sizes of word n-grams. Notice that when $n = 1$ represents the traditional bag of words configuration, *i.e.,* using single words as features. We report our experimental results in terms of precision, recall and F-score metrics for the positive class only (*i.e.,* firstborns profile).

As can be observed in Figure 1, adding a wider context to the documents' representation, *i.e.,* using a big value for n, affects the performance of the classification process. On the contrary, when small values of n are employed ($n = 2$) a better performance is obtained. These results are mainly due to the length of the transcripts, thus it is very difficult to find large n-grams in the texts.

It is worth remembering that our main goals were not directed to obtain a higher classification performance. Instead, by performing this experiment we want to validate the existence of some *lexical* features that allow to distinguish how firstborns are able to develop their language use in a different manner than later-borns. Accordingly, from the best configuration (*i.e.,* word $2-$grams) we

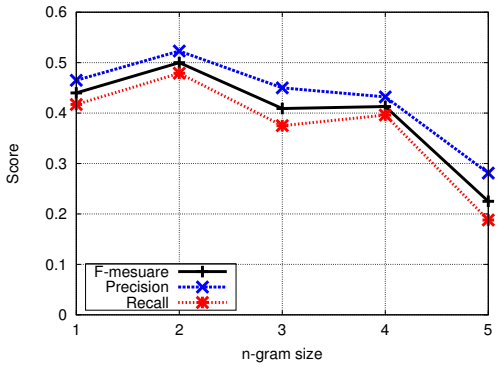

Fig. 1. Obtained performance of the classification model when word n-grams ($n = 1$ to $n = 5$) are employed as representation strategy.

compute a new representation considering only the most discriminative features. In order to construct this new representation we employed a widely used attribute selection strategy, namely Information Gain (IG).

Table 2. Results for the firstborns class obtained when IG is applied in order to select the most discriminative lexical features. First row depicts the performance of the best configuration, *i.e.,* word 2-grams. Second row shows obtained results when IG is applied to the best configuration, *i.e.,* word 2-grams+IG.

Features	**Vocabulary**	**Performance metrics**		
		Precision	*Recall*	*F-measure*
word 2-grams	14,086	0.52	0.48	0.50
word 2-grams+IG	367	1.00	0.88	0.93

Table 2 shows the obtained performance when this reduced form of representation is employed. As it is possible to observe, classification performance improves as a result of this dimensionality reduction. Intuitively, this results indicate that there are several *lexical* elements that are noisy features, and at the same time, support the intuition on the existence of certain *lexical* attributes that distinguish how firstborns and later-borns use their language abilities. In addition, we can also observe that the number of features is significantly reduced from 14,086 to 367 attributes, *i.e.,* only the 2% of the original feature set. In a following section (see Section 5) we present the list of top ten more useful word n-grams.

4.3 Syntactic features

The lexicon is usually attached to a particular thematic. Therefore, if we go up one level in the organization of the language production, we encounter the syntactic rules, which convey knowledge such as word's grammatical categories. Accordingly, grammatical categories generalize the use of the language and can enclose different lexical forms under a same category, for instance, the uses of *family*, *friend*, *dad*, *mom*, can be all be labelled as *noun*.

The goal of the experiments described in this section was to investigate if using grammatical categories for representing documents it would be possible to discriminate between firstborns and later-borns. Specifically, we used Part-Of-Speech (POS) tags to label each word on the transcripts with its grammatical category. Particularly, we employed the TreeTagger [17] tool with a set of 75 tags for the Spanish language[4].

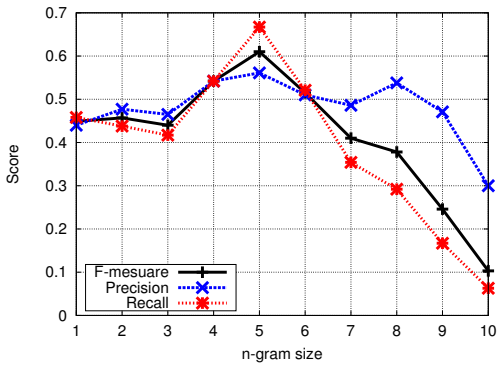

Fig. 2. Obtained performance of the classification model when POS n-grams ($n = 1$ to $n = 10$) are employed as representation strategy.

Similarly to the experiments described in previous section, we build a BoW-like representation using n-grams as main features. For the experiments described here, we employed POS n-grams of different sizes (from $n = 1$ to $n = 10$). Intuitively, larger values of n provide information about the structure (syntax) employed by authors.

Figure 2 shows the obtained performance by the classification model when POS tags are used for representing documents. Contrary to the experiments from previous section, we can notice that for this experiments the context information is important, thus the best result is achieved when $n = 5$.

In the same way we did with the *lexical* features, we applied the IG strategy to preserve the most discriminative POS n-grams features. Table 3 shows the

[4] The full set of POS tags and its descriptions can be found at `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/spanish-tagset.txt`

Table 3. Results for the firstborn class obtained when IG is applied in order to select the most discriminative POS features. First row depicts the performance of the best configuration, *i.e.,* POS 5-grams. Second row shows obtained results when IG is applied to the best configuration, *i.e.,* POS 5-grams+IG.

Features	**Vocabulary**	**Performance metrics**		
		Precision	*Recall*	*F-measure*
POS 5-gram	25,551	0.56	0.67	0.61
POS 5-gram+IG	464	1.00	0.79	0.88

obtained results after reducing the dimensionality of the POS 5-grams representation. Obtained results indicate that there is a particular subset of POS n-grams features that are useful describing how firstborns and later-borns compose their sentences, *i.e.,* the have particular syntactic rules when producing spontaneous speech. In Section 5 we will discuss about the list of n-grams present on the top ten attributes with more information gain.

5 Analysis and discussion

In order to contribute with the understanding of the features that are useful in distinguishing firstborns from later-borns, this section presents an initial analysis on the top ten lexical and syntactic patterns. Table 4 shows the top ten words 2-grams and Table 5 the most discriminative POS 5-grams. Both tables include some examples with their respective closest translation.

Table 4. Top ten word 2-grams with more information gain.

Top ten word 2-grams	
Literal n-gram	*Closest translation*
y-la	*(and the)*
papá-y	*(dad and)*
la-música	*(the music)*
va-en	*(goes in)*
porque-por	*(because)*
un-tiempo	*(some time)*
este-me	*(this me)*
que-ya	*(already)*
no-le	*(do not)*
o-lo	*(or it)*

Tabla 4 shows that the more discriminant words 2-grams are those that function as connectors of utterances. For instance, *y-la* (and the) is a connector

Table 5. Top ten POS 5-grams with more information gain. The POS tag labels in this example are ADV (Adverbs), ART (Articles), CC (Coordinating conjunction), CSUBX (Subordinating conjunction underspecified for subord-type), NC (Common nouns), NEG (Negation), PPO (Possessive pronouns), PREP (Preprosition), PPX (Clitics and personal pronouns), VEinf (Verb *estar*. Finite), VLfin (Lexical verb. Finite) and VLinf (Lexical verb. Infinitive)[4]

Top ten POS 5-grams	
Literal n-gram	*Example (closest translation)*
PPX-VLfin-VLinf-ART-NC	me gusta jugar el futbol *(I like to play the football)*
CSUBX-NEG-PPX-VLfin-ADV	porque no me gustaba mucho *(because I did not liked much)*
PREP-PPO-NC-CC-PPX	con mis amigos y yo *(with my friends and I)*
ART-NC-CC-PPX-VLfin	una persona y me desquito *(one person and I take it out)*
PPX-VLfin-ADV-VEinf-PREP	me gusta mucho estar en *(I like a lot to being in)*
PPO-NC-CC-PPX-VLfin	su trabajo o me pongo *(its work or I start)*
PREP-ART-NC-PREP-ART	en las calles con los *(in the streets with the)*
VLfin-PREP-VLinf-PREP-ART	voy a hacer en el *(I'm going to do in the)*
VLfin-ADV-VLinf-PREP-PPO	gusta mucho cantar con mis *(like a lot singing with my)*
NC-VLinf-PREP-ART-NC	este platicar con la gente *(ahhm talk with the people)*

that is used as link between clauses. A particular interesting word appearing in the list is *este* that is frequently used as filler. Moreover, the use of *porque por* might indicate hesitation in the speech, since both words in this case can be mean *because*).

Regarding the syntactic attributes (Table 5), our analysis revels the frequent use of personal pronouns. In addition, it is notorious the absence of nouns (NC) in the 5-grams, especially when the average length of a sentence in Spanish is around 5 words. Another interesting aspect to note is that some 5-grams do not have any verbs (*e.g.*, PREP-PPO-NC-CC-PPX, PREP-ART-NC-PREP-ART). We noticed that POS n-grams reveal some of the followed rules during the language production of teenagers. However, we intend to perform a deeper analysis on these patterns aiming at getting complete interpretation of these language phenomena.

6 Conclusions

In this paper we have addressed one of the long-established problems within the psychology field, *i.e.,* determining if there are long term vocabulary and language production differences associated with birth order among teenagers. Accordingly, we faced this problem as an Author Profiling task, where we modeled general sociolinguistic features that apply to our particular group of authors, namely *firstborns* and *later-borns*.

Our performed experiments demonstrate that there is a strong relation between a subset of *lexical* and *syntactic* features and the order of birth. These findings represent an important contribution for both, psychological and computational research fields. On the one hand, we provided empirical evidence on the differences of language use among firstborns and later-borns, and on the other hand, this work represents the first attempt in employing NLP techniques, particularly traditional author profiling techniques, to this concrete problem.

Although good results were achieved, our future work is directed to perform a deeper analysis on found *lexical* and *syntactical* features. A detailed analysis on the meaning of these language production rules will imply an additional research work.

Acknowledgments. This work was partially funded by CONACyT under the Thematic Networks program (Language Technologies Thematic Network project no. 260178). Additionally, authors would like to thank to INAOE, UAM-Cuajimalpa, UPAEP and SNI-CONACyT for their support.

References

1. E. Hoff, "How social contexts support and shape language development," *Developmental Review*, vol. 26, pp. 55–88, 2006.
2. M. D. Healey and B. J. Ellisb, "Birth order, conscientiousness, and openness to experience tests of the family-niche model of personality using a within-family methodology," *Evolution and Human Behavior*, vol. 28, pp. 55–59, 2007.
3. A. Adler, *The practice and Theory of Individual Psychology.* New York, 1927.
4. C. Gustafson, "The effects of birth order on personality," Master's thesis, Alfred Adler Graduate School, 2010.
5. Y. Oshima-Takane, E. Goodz, and J. L. Deverensky, "Birth order effects on early language development: Do secondborn children learn from overheard speech?," *Child Development*, vol. 67, pp. 621–634, 1996.
6. I. M. Zambrana, E. Ystrom, and F. Pons, "Impact of gender, maternal education, and birth order on the dev. of language comprehension: a longitudinal study from 18-36 months of age," *J. Dev. Behav. Pediatr*, vol. 33, pp. 146–155, 2012.
7. M. H. Bornstein, D. B. Leach, and O. M. Haynes, "Vocabulary competence in first and secondborn siblings of the same chronological age," *Journal of Child Language*, vol. 31, no. 4, pp. 855–873, 2004.
8. S. Stolt, L. Haataja, H. Lapinleimu, and L. Lehtonen, "Associations between lexicon and grammar at the end of the second year in finnish children," *Journal of Child Language*, vol. 36, no. 4, pp. 779–806, 2009.

9. K. Keller, L. Troesch, and G. A., "First-born siblings show better second language skills than later born siblings," *Frontiers in Psychology*, vol. 6, no. 705, 2015.
10. M. Koppel, S. Argamon, and A. R. Shimoni, "Automatically categorizing written texts by author gender," *Literary and Linguistic Computing*, vol. 17, no. 4, pp. 401–412, 2002.
11. A. P. López-Monroy, M. Montes-y Gómez, H. J. Escalante, and L. Villaseñor-Pineda, "Using intra-profile information for author profiling," in *Working Notes for CLEF 2014 Conference*, vol. 1180, pp. 1116–1120, 2014.
12. W. Youyou, M. Kosinski, and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans," *Proceedings of the National Academy of Sciences*, vol. 112, no. 4, pp. 1036–1040, 2015.
13. J.-V. Cossu, N. Dugué, and V. Labatut, "Detecting real-world influence through twitter." https://hal.archives-ouvertes.fr/hal-01164453, 2015. (hal-01164453).
14. J. W. Pennebaker, *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury USA, 2011.
15. K. Cruz-Sanchez, V. Reyes-Meza, M. Martínez-Gómez, R. Hudson, and A. Bautista-Ortega, "Effects of birth order and number of siblings on personality and stress response," *Developmental Psychobiology*, vol. 57, no. S:10, 2015.
16. S. R. Garner, "Weka: The waikato environment for knowledge analysis," in *In Procceding of the New Zealand Computer Science Research Students Conference*, pp. 57–64, 1995.
17. H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of the International Conference on New Methods in Language Processing*, (Manchester, UK), 1994.