Universidad Autónoma de Madrid

Escuela Politécnica Superior

22
23

Trabajo fin de grado

Author Profiling

Luna Mancebo Rodríguez

Escuela Politécnica Superior
Universidad Autónoma de Madrid
C\Francisco Tomás y Valiente nº 11

excelencia UAM CSIC

www.uam.es

Universidad Autónoma de Madrid

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Grado en Computer Engineering

TRABAJO FIN DE GRADO

Author Profiling

An analysis of the differences in writing styles in formal and informal texts

Autora: Luna Mancebo Rodríguez
Tutora: Lara Quijano-Sánchez

Marzo 2022

Luna Mancebo Rodríguez
Author Profiling

Luna Mancebo Rodríguez
C\ Francisco Tomás y Valiente nº 11

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

Agradecimientos

Resumen

Palabras clave

Otras

Abstract

Keywords

Algunas

ÍNDICE

1 Introduction **1**

2 State of the Art **3**

3 Methods **9**

3.1 Datasets 9

3.1.1 Twitter dataset 9

3.1.2 Police dataset 10

3.2 Data pre-processing 10

3.2.1 Twitter dataset 10

3.2.2 Police dataset 11

3.3 Feature selection 11

3.3.1 Stylistic features 11

3.3.2 Twitter features 12

3.3.3 N-grams features 13

Bibliografía **14**

Listas

Lista de figuras

Lista de tablas

2.1 Bibliography research: stylometry query 4

3.1 Stylistic features 12

1

Introduction

With the increased use of new technologies and algorithms in sensible tasks such as hate speech detection or predictive policing, it is crucial that not only the algorithm has a high performance, but also that the data which it is trained with is fair and represents every population group. As stated in FRA 2022 [?] "algorithms are only as good as the data that are used to develop them".

The term 'bias' can have different meanings depending on the field of study. Bias can refer to any of the following:

- Differential treatment based on protected characteristics: this refers to an inclination toward a certain group of people based on characteristics such as gender, ethnicity, religion...
- Statistical bias: it exists when the data isn't adequately measuring what they are intended to measure. For example, if a sample of the general population contains more men than women, it is said to be biased towards men.

Biases in algorithmic systems may lead to discrimination, but it is important to differentiate between the two terms. Not all forms of bias relate to protected characteristics. For instance, an algorithmic model that differentiates between people based on whether or not they have pets, doesn't target a protected characteristics. Moreover, even if the target is a protected characteristic, if the result doesn't lead to a disadvantageous situation for a group of people, it isn't considered discriminatory. Therefore the problem arises when bias in algorithms results in direct discrimination when the target is a protected characteristic and it leads to a less favourable treatment for a certain group of people.

Bias most frequently occurs when the data used to train an algorithm only reflect certain demographic groups or reflect social biases. Determining where bias comes from is a challenging exercise, and as it is a fairly new subject, many studies are being made on this matter. How bias can be detected and eliminated in AI applications are at the center of discussions around regulating AI. Unified regulations and policies regarding this subject are crucial in order to eradicate bias from algorithms, so that we can all benefit, without discrimination, from such a powerful tool.

EU institutions have become more engaged in the area of AI, bias and fundamental rights, since they have acknowledged it is a problem that needs mitigation through policy and legislative proposals. Since 2017, institutions such as the European Parliament, the European Council and the European Commission, have been working towards a proposal for regulating AI applications. On 21 April of 2021, the Commission published its proposal for an AIA, which forms part of its digital strategy, and is a key aspect for the EU to make its law fit for the digital age.

Not only researchers or policymakers are interested on the subject of bias in AI, the press are very sceptical about these algorithms, and if it is possible they could be making more harm than good. For instance, several articles [? ?] are being written stating that the AI VeriPol, whose goal is to predict if a police report is fake or not, is biased because of the differences in writing styles depending on the complainant's demography. These statements are subjective, as no studies have been made that demonstrate that differences in writing styles exist depending on where the authors of the text are from. Even if these claims where true, and differences in writing styles existed based on the demography of the author, it hasn't been proven that the prediction of the algorithm is less favourable towards a group of people. Therefore it is extremely important to conduct studies that uncover bias, in order to eliminate it, or on the other hand that demonstrate that bias doesn't exist and silence false claims.

NLP refers to the branch of computer science, more specifically the branch of AI, concerned with giving computers the ability to understand text and spoken words. It is a field that is rapidly growing and many progress has been made recently. With it, a new area susceptible to bias has opened. Algorithms that predict hate speech or sentiment analysis, for instance, have been proven to be biased [? ?] and many studies have been conducted in order to tackle this problem, building up methods to mitigate bias [? ? ?]. Other studies have worked on the task of creating a framework to detect bias in text [? ?].

The main goal of this study is to extensively analyze the differences in writing styles, lexical and syntactical, depending on the author's demographics. In this paper we don't just focus on demographic traits like gender and age, but introduce the novelty of studying if differences in writing styles exist depending on the geographic region of the author of a text.

As social media is the most common used tool to express emotions and thoughts, most of the studies conducted on the field of extracting differences in writing styles based on demographic traits have used datasets containing texts of social media platforms like Twitter. The second point of study of this paper is, using the same demographic characteristics of gender, age and region for extracting writing style differences, compare the obtained results when using a dataset of informal texts (i.e tweets), or formal texts (i.e police reports).

If the results obtained for the police dataset are statistically significant, which would mean differences in writing style exist, the next point to tackle in this paper would be to study whether or not the algorithm VeriPol is less favourable towards a specific group. In case it was discriminatory, the final posed question for this research would be: what techniques can be used to eliminate the existing bias? Are they useful? Do they mitigate bias, or is the algorithm still discriminatory?

The paper is organized as follows: in chapter 2 we present the state of the art and summarize the most relevant findings of the bibliography research.

2

State of the Art

One of our research questions is whether or not significant differences in writing style exist, depending on the author's gender, age or region. That is why we want to analyze other studies that address one or several of our tasks, and see which stylometry techniques they have employed, and the drawn conclusions. Stylometry is the study of writing style, and it dates back several decades [?]. Computational stylometry distinguishes several subtasks such as determining and verifying author identity, and author profiling.

Most of the studies analyzed in our bibliography research, address the task of author profiling, trying to predict the author's gender. Other papers, like [? ?] also try to predict the political ideology of the author. Another common demographic trait to predict in the author profiling task is age [? ? ? ? ? ? ? ? ?]. Research papers that participated in the PAN 2015 [? ?] also include five different personality traits in their prediction targets. Studies that were submitted to PAN 2019 [? ? ?] address the task of predicting whether the author of a text is human or a bot, and in case it is human, predict it's gender. In the paper [?] not only do they address the gender prediction task, but they also predict the Spanish language variety of the author (between eight different varieties of Spanish). This could be treated as a similar problem as the geographic region characteristic, as the language variety depends on the region of the author. We can therefore conclude that most of the research made thus far focuses on studying differences in writing depending on the gender and age of the author. In this paper we introduce a new demographic trait, the geographic region of the writer.

Besides author profiling, authorship attribution also uses stylometry to help predict the author of a text. For instance, paper [?] uses three different group of features and a SVM classifier to predict the author of a tweet. Another example is the study conducted in [?], were they use three groups of features (lexical, structural and idiosyncratic features); and three distances (Euclidean, Manhattan and Cosine) to measure the distance between the feature vectors of each author.

As we mentioned in the introduction, social media is a common platform were people tend to express their ideas and thoughts, therefore most of the studies are conducted using datasets based on tweets. Some papers [? ? ? ? ?] use hotel reviews or blog entries as their input text. Others, like [?] use more formal texts, like TEDTalks. The research conducted in [?] compares the gender prediction task when using a formal and informal text dataset. This is one of the research questions posed in our study, as not many studies focus on this approach and we believe it is interesting the comparison between these two types of datasets.

There are different tasks were stylometry is useful besides the common author profiling. For instance, the studies's [? ?] main goal is adversarial stylometry, which consists in rewriting the input text such that its style changes and the stylometric differences are blurred, in order to standarize the texts so that it is difficult to predict demographic traits of the author. This could be a possible mitigation technique in case VeriPol was discriminatory.

Moreover, stylometry is nowadays being used in literary tasks, such as identifying the similarities between novels, or studying the writing style of the authors. For instance, the paper [?] uses stylometry to prove or refute the hypothesis that *Madam Bovary* had a significant influence on *La Regenta*. The principal feature used is *Most Frequent Words* using n-grams (with n being 1, 2 or 3) and choosing between 100 and 5000 words. The main difference of this paper with the rest of the bibligraphy, is that their study is purely statistical, no classifier is used, and therefore no predictions are made. To perform the contrast of hypothesis, they use two different distances, the Euclidean and the Delta of Burrows. Another example is the problem addressed in [?] where they use stylometry to study the writing style of a poet through handwritten manuscripts. The experiment is based on the extraction of the most frequent words (100, 300 and 500 most frequent) and applying three different methods: characteristic curve, Chi-squared and Delta of Burrows distance.

For our first research question, a specific search query was made in order to obtain papers whose main focus was stylometry and feature selection (statistical, stylometric, lexical...). Many classifiers were used, and all of them employed accuracy as the evaluation metric to test the performance of the model. Some of the models that obtained great results were RF used in [? ?], SVM used in [? ? ?] and CNN used in [? ?]. We summarize this findings in table 2.1

Tabla 2.1: Bibliography research: stylometry query

ID	Demographic traits	Dataset	Data preprocessing	Feature Extraction	Classifier	Results
[?]	Gender Age Political ideology	Twitter	Remove hyperlinks Lowercase text Remove non-alphabetical tokens	N-gram based features Linguistic features Embeddings-based features	CNN BERT BiGRU	Gender: 72.02 (BiGRU) Age: 46.68 (BiGRU)
[?]	Gender Profession Political ideology	Twitter		Word frequency Statistical characteristics RoBERTa embeddings	LR RF DT MLP	Gender: 67 (LR)

ID	Demographic traits	Dataset	Data preprocessing	Feature Extraction	Classifier	Results
[?]	Bot Gender	Twitter		Character based features Word based features Syntax based features Twitter features	LR RF SVM CNN KNN	Gender: 88.88 (RF)
[?]	Gender	Hotel reviews	Stop-word removal Stemming Remove punctuation marks	Content based features Syntactic features	NB RF	Gender: 93.25 (RF with 8000 frequent terms and 3000 POS n-grams)
[?]	Gender	Twitter YouTube News posts		Lexical features Morphological features Syntactical features Character-based features	LR	Gender: 62.8 (lexical features)
[?]	Bot Gender	Twitter	Unify tweets in a single document Remove stop words Stemming Lemmatization Spell correction Splitting hashtags	Psycholinguistic features	GBDT	Gender: 88
[?]	Bot Gender	Twitter	Replace emojis, URL, mentions, special characters Lowercase text Trim repeated characters Remove N-grams repeated in every document	N-grams features	LR SVM MLP	Gender: 76
[?]	Gender Language variety	News posts	Eliminate source code Lowercase text Replace emojis, URL, hashtags and numbers with special characters	N-grams features Bag of Words Word embeddings Sentence embeddings	CNN SVM	Gender: 75.61 (SVM with 8-gram characters) Language variety: 94.16 (SVM with combination of features)

ID	Demographic traits	Dataset	Data preprocessing	Feature Extraction	Classifier	Results
[?]	Gender	Novels		Character-based features Word-based features Sentence-based features Dictionary-based features Syntactic features Discourse features	SVM	Gender: 88.94 (syntactic features)
[?]	Gender Age	Twitter		Character-based features Word-based features Semantic features Syntactic features Vocabulary richness Readability-based features	CNN NB DT RF SVM	Gender: 97.7 (CNN) Age: 90.1 (CNN)
[?]	Gender	News posts		Character-based features Word-based features Sentence-based features Discourse features Syntactic features Dictionary based features	RF	Gender: 71.23 (Character-based features)
[?]	Gender Age	Twitter	Unify tweets in a single document Remove emojis, hashtags, links... Remove all non-letter characters	Structural Stylometry Second Order Attributes (SOA)	NB DT RF SVM	Gender: 75 (N-grams combined with weighed SOA) Age: 54 (SOA)
[?]	Gender Age	Twitter	Lowercase Remove mentions, hashtags, retweets... Remove all non-alphabetic characters Remove stopwords	Word unigrams that occur at least two times Word bigrams Character 4-grams Average spelling error Punctuation feature	LR	Gender: 75.64 Age: 51.79

ID	Demographic traits	Dataset	Data preprocessing	Feature Extraction	Classifier	Results
[?]	Gender Age	Twitter	Replace numbers, URL, mentions, picture links, emojis and slang words for special characters Split punctuation marks from adjacent words	Character ngrams (affixes, words and punctuation)	LR SVM	Gender: 66 (LR with SOA) Age: 44 (LR with SOA)
[?]	Gender Age Personality traits	Twitter		Stylometry-based features (29 features)	NB SVM RF LR	Gender: 73 Age: 53
[?]	Gender Age Personality traits	Twitter	Unify tweets Remove HTML tags, hashtags, URL, mentions, replies... Remove duplicate tweets	Stylometry-based features Structural features	SVM	Gender: 95 Age: 78
[?]	Gender Age Personality traits	Twitter		Stylometry-based features Content-based features	SVM RF	Gender: 76 Age: 41 (RF with 2000 trees)
[?]	Author identification	Twitter	Grouping tweets of the same author	Lexical features Syntactical features Twitter-based features	SVM	91.11

3

Methods

3.1. Datasets

As mentioned in chapter 1, the second research question is to compare the results obtained for the main task of this study (whether or not differences in writing style exist depending on the gender, age and region of the author of the text), when the employed dataset contains formal or informal texts. Therefore this study uses two different datasets: the Twitter dataset (with informal texts) and the police dataset (with formal texts).

3.1.1. Twitter dataset

The first step in order to build the Twitter dataset is to select which users will conform it. This selection is done manually, as the Twitter API doesn't provide information about the age, gender or nacionality of its users. Therefore the selected users must have this information publicly available in thier profile description. Each demographic trait is grouped as follows:

- Gender: female and male
- Age: 18-24, 25-34, 35-44, 45-54 and +55
- Region: the selected users nacionality is spanish, as the goal of the study is to analyze different writing styles in the spanish language. Thus the selected users are grouped based on their autonomous community. In Spain there are 17 autonomous communities, all of which are present in the dataset except for "Islas Baleares".

The second step is to obtain the 100 most recent tweets of each of the selected users. The python library *tweepy* is a useful resource to easily access the Twitter API. It provides a function *search all tweets* that permits to indicate a specific search query. For this study, the parameters specified in the query are the language and the number of returned tweets, set to "Spanish" and "100" respectively.

The dataset is composed of **1146** users and 114.600 tweets, of which:

	Gender		Age				
	Female	Male	18-24	25-34	35-44	45-54	+55
Count	542	604	187	221	232	267	239
Frequency (%)	47.29	52.71	16.32	19.28	20.24	23.3	20.86

	Region															
	Catalunya	Andalucia	Murcia	Canarias	La Rioja	Madrid	Asturias	Castilla-La Mancha	Navarra	Valencia	Galicia	Pais Vasco	Castilla y Leon	Aragon	Cantabria	Extremadura
Count	91	87	81	81	79	76	76	73	73	73	73	63	61	55	53	51
Frequency (%)	7.94	7.59	7.07	7.07	6.89	6.63	6.63	6.37	6.37	6.37	6.37	5.50	5.32	4.80	4.62	4.45

Table 2: Proportions of Twitter dataset

3.1.2. Police dataset

The dataset is build of 3899 police reports each of which provides the following information:

- Police officer: gender, age and region of birth.
- Complainant: gender age and region of birth.
- Complaint text: the text has been previously anonimezed replacing sensitive data such as locations, telephone numbers, people's names and dates with specific tags.

Hence, after analyzing the dataset, the proportions of each class are:

	Gender		Age				
	Female	Male	18-24	25-34	35-44	45-54	+55
Count	878	3020	18	414	2348	918	200
Frequency (%)	22.5	77.5	0.46	10.61	60.28	23.53	5.12

	Region																	
	Catalunya	Andalucia	Murcia	Canarias	Madrid	Asturias	Castilla-La Mancha	Navarra	Valencia	Galicia	Pais Vasco	Castilla y Leon	Aragon	Cantabria	Extremadura	Islas Baleares	Melilla	Alemania
Count	78	661	104	42	405	121	485	11	1189	55	10	171	20	27	454	15	47	3
Frequency (%)	2	16.97	2.65	1.08	10.38	3.1	12.48	0.28	30.47	1.41	0.26	4.38	0.51	0.69	0.38	1.2	0.08	

Table 3: Proportions for police officers of Police dataset

	Gender		Age				
	Female	Male	18-24	25-34	35-44	45-54	+55
Count	1688	2185	597	638	792	753	1093
Frequency (%)	43.58	56.42	15.41	16.47	20.45	19.44	28.22

	Region															
	Catalunya	Andalucia	Murcia	Madrid	Asturias	Castilla-La Mancha	Navarra	Valencia	Galicia	Castilla y Leon	Aragon	Cantabria	Extremadura	Melilla	Latino	No hispanohablante
Count	66	461	96	115	18	387	4	1031	13	70	15	3	338	10	386	824
Frequency (%)	1.7	11.9	2.48	2.97	0.46	9.99	0.1	26.62	0.34	1.81	0.39	0.08	8.73	0.26	9.97	21.28

Table 4: Proportions for complainant of Police dataset

3.2. Data pre-processing

Text pre-processing is a crucial step before moving on to feature extraction.

3.2.1. Twitter dataset

Tweets are short texts of a maximum length of 280 characters. By themselves they don't represent with enough extent the writing style of their author. Therefore, in order to create each user's text document, the retrieved tweets are concatenated in a single document per user.

The previous step before feature extrtaction must be cleaning the data, in this case, processing the text and what characters are going to be usefull for the classifier. For the *Twitter-based features*, no cleaning is made, but the rest feature groups must undergo a common cleaning step consisting on:

- Remove URLs
- Replace emojis with its text counterpart
- Remove mentions: words starting with the character “@”
- Remove hashtags: words starting with the character “#”

For the *Character-based and Structural-based features* it is enough to just perform this step. However, before extracting *Syntactical-based features* punctuantion symbols are removed, and on top of that, for *Word-based features* stop words are eliminated. In order to do so, a file containing a list of Spanish stop words is used, instead of using the common NLTK library. This decision was made as the file used contained more words than the Spanish NLTK stopword function.

3.2.2. Police dataset

For the police dataset, the reports have been previously anonimized, and sensitive data such as names, places, telephone numbers or dates have been replaced with tags. With this dataset the first cleaning step of removing twitter characteristics isn’t necessary as it was for the Twitter dataset, instead, the tags related with the anonimization of the data are removed. The rest of the steps are the same, before extracting syntactical-based features punctuation symbols are removed, and for the word-based features stopwords are eliminated.

3.3. Feature selection

Feature extraction is a crucial step in studying differences in writing style. Through our literature, presented in chapter 2, we have been able to select the best performing features for similar tasks. The features used in our study are purely statistical and are divided in three categories: stylistic features, N-grams features and Twitter features (the later only for the Twitter dataset).

3.3.1. Stylistic features

It focuses on the differences on the arrangements of aspects of the text (word, sentence, paragraph). We can divide these kind of features in four subgroups: word based features, character based features, structural features and syntactical features. The following table 3.1 presents the features used and their category.

Tabla 3.1: Stylistic features

Category	**Feature description**
Word based features	Count of words
	Count of positive words
	Count of negative words
	Unique words count
	Count of words that occur twice
	Average word length
	Maximum length of a word
	Count of words with numbers
	Count of words with length greater than 6
	Count of words with length smaller than 3
	Count of stop words
Character based features	Character count
	Count of capital letters
	Count of punctuation marks
Structural features	Sentence count
	Average count of sentences per paragraph
	Average count of words per paragraph
	Average count of characters per paragraph
	Variation in tweets length
Syntactic features	Determiners count
	Prepositions count
	Singular noun count
	Plural nouns count
	Adverbs count
	Adjectives count
	Proper nouns count
	Pronouns count
	Conjunctions count
	Count of past tense verbs
	Count of future tense verbs

3.3.2. Twitter features

We have applied this method only to the Twitter dataset, as we believe it can contribute with interesting insights to help build an answer to the main question of this study. The features in question are:

- Number of retweets
- Number of mentions
- Number of URLs
- Number of emojis
- Number of hashtags

3.3.3. N-grams features

Word and character N-grams are popular features in Information Retrieval tasks. We have based our approach in the method used by [?], as they provided good results and we want to establish a robust method as a baseline. The counterpart of this technique is that it is computationally expensive and it doesn't take into account the surrounding n-grams (it is context-less). On the other hand, it is an approach that captures the content of words. This can be both a positive and negative aspect, as it might draw conclusions that are specific for the topics of the texts of the dataset (loss of generality).

We extracted word and character n-grams using TF-IDF, using unigrams, bigrams and trigrams for words. For characers we combined sequences between 2 and 7 character lenght without word boundaries. We combined both feature sets into one, and applied LSA to reduce the dimension of the vector, obtaining one of 100 components. We calculated a vector per tweet, and then averaged them, resulting in the author's vector.

Universidad Autónoma de Madrid