Aplicación del procesamiento de lenguaje natural para verificación de identidad.

Felipe Orlando López Pabón
BSc. student in Electronic Engineering

Advisor: **Prof. Juan Rafael Orozco Arroyave Ph.D.**

Co-Advisor: **MSc. Juan Camilo Vazquez Correa**

GITA research group, University of Antioquia.

forlando.lopez@udea.edu.co

June 4, 2019

Outline

Introduction

Context: Virtual education

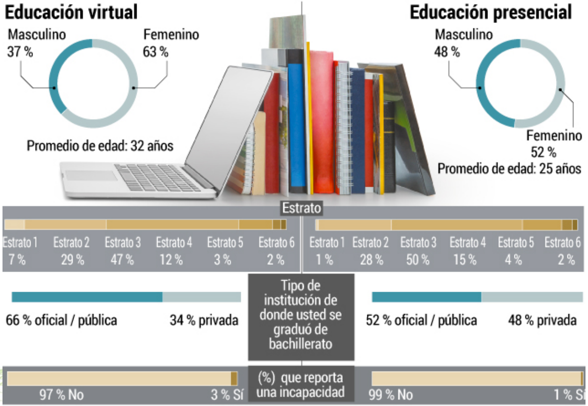


Figure: Virtual education in Colombia. Figure taken from M. Díaz 2018.

Hypothesis and objectives

Hypothesis

People improve their redaction skills as they advance in their university career, therefore, the linguistic style of a person in the first levels is different from the of a person in intermediate levels and also different from the of a person in last levels or the of a person with a university degree.

Hypothesis and objectives

Objectives

General Objective To develop algorithms that allow to differentiate the linguistic styles of people that belong to the university community and that are registered in a web platform through natural language processing (NLP) techniques.

Hypothesis and objectives

Objectives

Specific Objectives

1. To evaluate the usefulness of NLP measures to differentiate linguistic styles.
2. To extract relevant linguistic features associated to written texts made by the users of the web page.
3. To implement classification systems and build user models that allow differentiating people according to their linguistic style.
4. To measure the performance of the systems through percentages of: accuracy, precision, sensitivity, specificity and F1-score, also with confusion matrix.

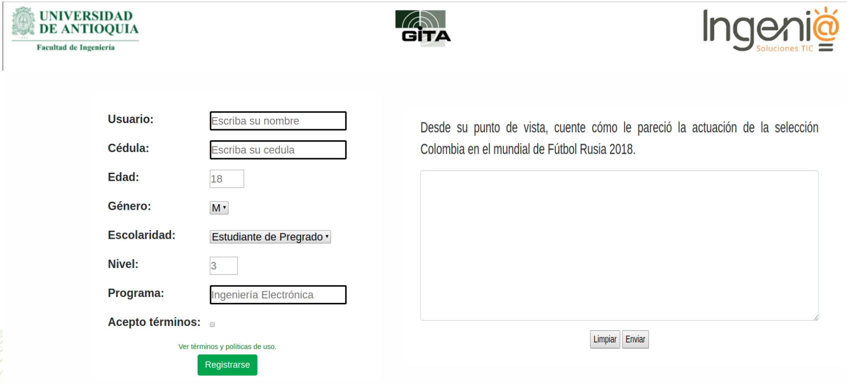
Hypothesis and objectives

Contribution of this work

The Vector Support Machine (SVM) and Random Forest (RF) classifiers are implemented, as well as the Gaussian Mixture Models (GMM) in order to distinguish the linguistic style of three groups of people, which is considered through different features such as Bag of Words (BoW), Term frequency - Inverse document frequency (TF-IDF), Word2vec, Global Vectors (GloVe) and Grammatical features.

Database

Web page



UNIVERSIDAD DE ANTIOQUIA
Facultad de Ingeniería
GITA
Ingeni@
Soluciones TIC
Usuario:
Escriba su nombre
Cédula:
Escriba su cedula
Edad:
18
Género:
M
Escolaridad:
Estudiante de Pregrado
Nivel:
3
Programa:
Ingeniería Electrónica
Acepto términos:
Ver términos y politicas de uso.
Registrarse
Desde su punto de vista, cuente cómo le pareció la actuación de la selección Colombia en el mundial de Fútbol Rusia 2018.
Limpiar
Enviar

Tasks and Database

Performed tasks

Table: Description of the performed tasks for the construction of the database.

Task	**Description**
1	Desde su área profesional, argumentar una posible solución frente a la contaminación de fuentes hídricas que está sufriendo el país actualmente.
2	Desde su punto de vista, cuente cómo le pareció la actuación de la selección Colombia en el mundial de Fútbol Rusia 2018.

Tasks and Database

General information about users.

Table: Information about all the participants in this study. μ: average, σ: standard deviation.

	Total		Group 1 (G1)		Group 2 (G2)		Group 3 (G3)	
	Men	Women	Men	Women	Men	Women	Men	Women
Number of subjects	96	45	38	14	29	17	29	14
Age ($\mu \pm \sigma$)	23.7 ± 5.5	24.5 ± 7.5	20.5 ± 3.2	20.2 ± 1.3	24.0 ± 4.9	23.5 ± 4.2	27.8 ± 5.9	29.9 ± 10.9
Bachelor students	85	37	38	14	29	17	18	6
Professionals	3	5	-	-	-	-	3	5
Magisters	4	-	-	-	-	-	4	-
Doctors	4	3	-	-	-	-	4	3

Methodology

Methodology

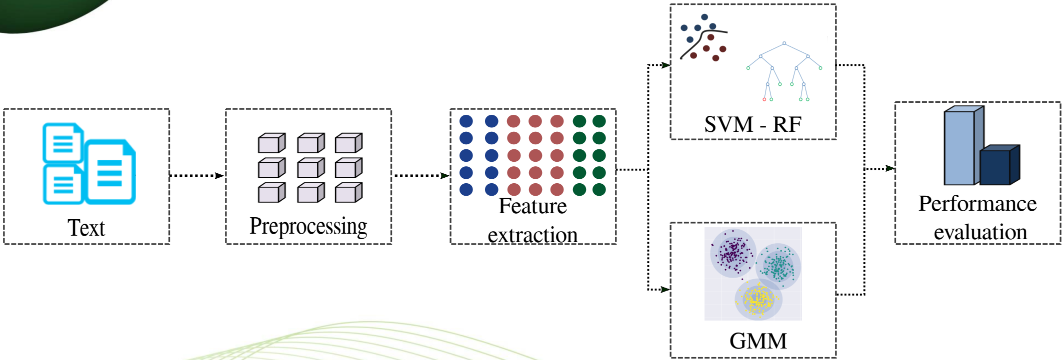


Figure: Block diagram of the methodology implemented in this study.

Feature Extraction

Bag of Words (BoW)

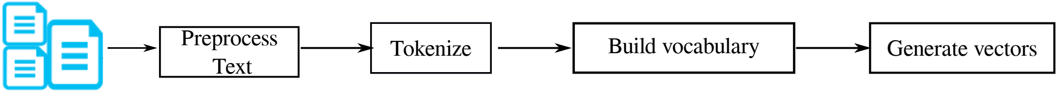


Figure: Scheme of the BoW method. Figure adapted from P. Dubey 2016.

Bag of Words (BoW)

campar	jugar	gustar	cada	jugador	esforzar	poder	pensar	partir	director	tecnico	malo
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure: Example of the BoW values obtained with the second task.

Term frequency – Inverse document frequency (TF-IDF)

In equations 1, 2 and 3 is shown the way to obtain the TF-IDF value (N. S. Sarwan 2017).

$$TF(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}} \tag{1}$$

$$IDF(t) = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}\right) \tag{2}$$

$$\mathrm{TF} - \mathrm{IDF} = TF \cdot IDF \tag{3}$$

Term frequency – Inverse document frequency (TF-IDF)

campar	jugar	gustar	cada	jugador	esforzar	poder	pensar
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0927744677391947
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.13577135174611427	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.14688053574178667	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0913853343510375	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.09141021465778124	0.0	0.0

Figure: Example of TF-IDF values obtained with the second task.

Word2vec

Word2Vec uses nearby words to represent target words with a shallow neural network whose hidden layer encodes the representation of the word. The aim is to represent the words as a vector in a multidimensional space, where similar or related words are represented by nearby points (C. Bellei 2018).

Word2vec

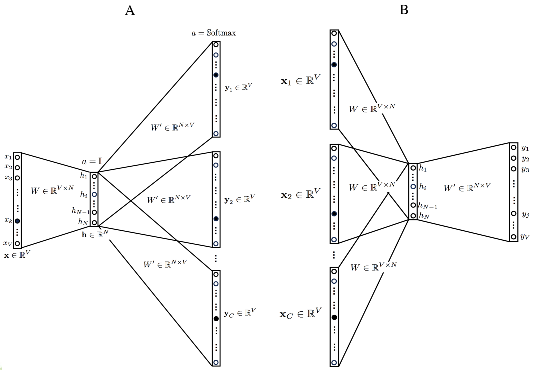


Figure: Topology of the models used in Word2Vec. A) *Skip Gram*, B) CBOW. Figure adapted from C. Bellei 2018.

Global Vectors for Word Representation (GloVe)

The GloVe model obtains word vectors when examining the co-occurrences of them within a corpus. Before training the model, it must be build a co-occurrence matrix X, where a cell X_{ij} tabulates the number of times that the word j appears in the context of the word i. Then, this co-occurrence data is used instead of the corpus (Pennington, Socher, and Manning 2014).

Global Vectors for Word Representation (GloVe)

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{ij})(\vec{w_i}^T \vec{w_j} + b_i + b_j - \log(X_{ij}))^2 \tag{4}$$

$$f(X_{ij}) = \begin{cases} \left(\frac{X_{ij}}{x_{\max}}\right)^{\alpha}, \text{si } X_{ij} < x_{\max} \\ 1, \text{en otro caso.} \end{cases} \tag{5}$$

Grammatical features

The following eight grammatical features are taken into account (Kincaid et al. 1975):

$$FR = 206.835 - 1.015 \frac{\#\ \text{words}}{\#\ \text{sentences}} - 84.6 \frac{\#\ \text{syllabes}}{\#\ \text{words}} \tag{6}$$

$$FG = 0.39 \frac{\#\ \text{words}}{\#\ \text{sentences}} + 11.8 \frac{\#\ \text{syllabes}}{\#\ \text{words}} + 15.59 \tag{7}$$

$$DP = \frac{\#\ (\text{verbs} + \text{adjectives} + \text{prepositions} + \text{conjunctions})}{\#\ \text{words}} \tag{8}$$

$$DC = \frac{\#\ (\text{verbs} + \text{nouns} + \text{adjectives} + \text{adverbs})}{\#\ \text{words}} \tag{9}$$

Grammatical features

$$NVR = \frac{\#\ \text{nouns}}{\#\ \text{verbs}} \tag{10}$$

$$NR = \frac{\#\ \text{nouns}}{\#\ (\text{nouns} + \text{verbs})} \tag{11}$$

$$PR = \frac{\#\ \text{pronouns}}{\#\ (\text{pronouns} + \text{nouns})} \tag{12}$$

$$SCR = \frac{\#\ (\text{subordinated conjunctions})}{\#\ (\text{coordinated conjunctions})} \tag{13}$$

Classification

The aim of a SVM is to discriminate data samples by finding a separating hyperplane that maximizes the margin between classes (Bishop 2006). The decision function of a soft-margin SVM is expressed according to Equation 14.

$$y_n \cdot (\mathbf{w}^T \phi(\mathbf{x}_n) + b) \geq 1 - \xi_n, \quad n = 1, 2, 3, \cdots, N \tag{14}$$

Classification: Support vector Machine (SVM)

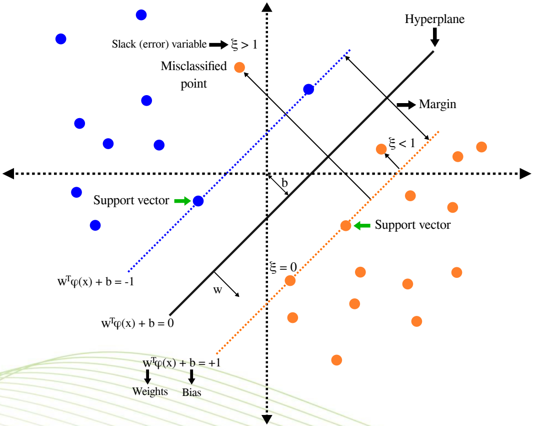
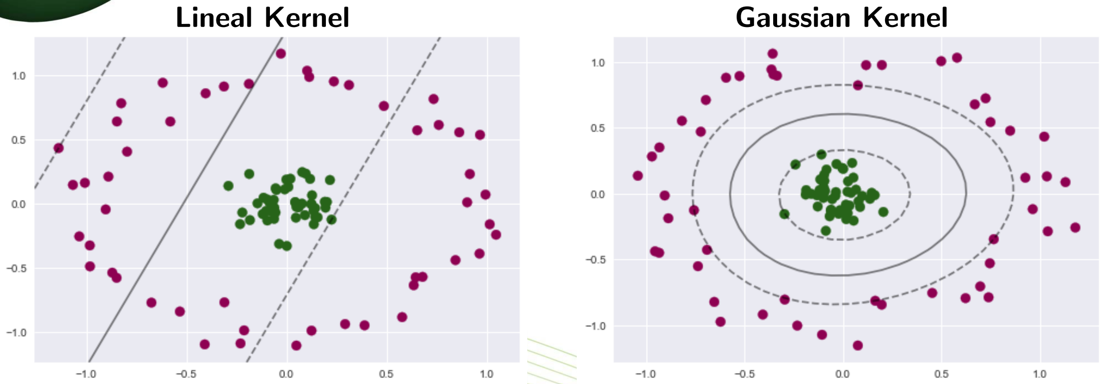


Figure: Soft-Margin SVM. Figure adapted from Sandipan Dey 2018.

Classification: Support vector Machine (SVM)

UNIVERSIDAD DE ANTIOQUIA 1803



Classification: Random Forest (RF)

This is one of the most common ensemble methods, which is based on the combination of multiple algorithms to make the final decision. Particularly, the RF combines several classifiers such as the decision trees (Gislason, Benediktsson, and Sveinsson 2006).

Classification: Random Forest (RF)

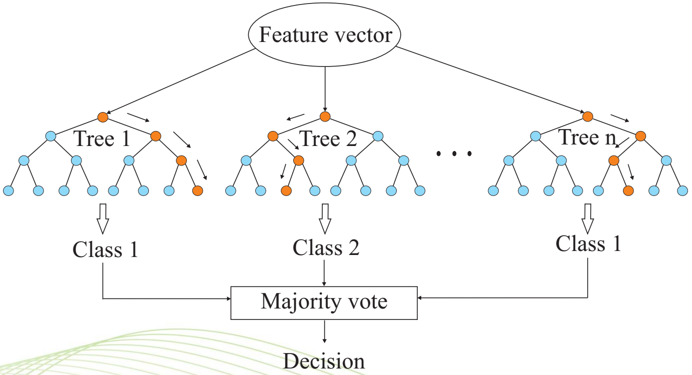


Figure: RF classifier scheme.

Generation of user models

Gaussian Mixture Model (GMM)

- GMM searchs a mixed of gaussian probability distributions that best model any dataset.
- Soft version of K-Means: EM algorithm for GMM.

$$p(\boldsymbol{x}) = \sum_{m=1}^{M} \frac{c_m}{(2\pi)^{\frac{D}{2}} \left|\boldsymbol{\Sigma_m}\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu_m}\right)^{T} \boldsymbol{\Sigma_m}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu_m}\right)\right] \tag{15}$$

Where $\boldsymbol{\mu_m}$ and $\boldsymbol{\Sigma_m}$ are the vector of means and the covariance matrix of the random vector x respectively. c_m is the weight associated with the m-th Gaussian component and meets $\sum_{m=1}^{M} c_m = 1$.

Gaussian Mixture Model (GMM)

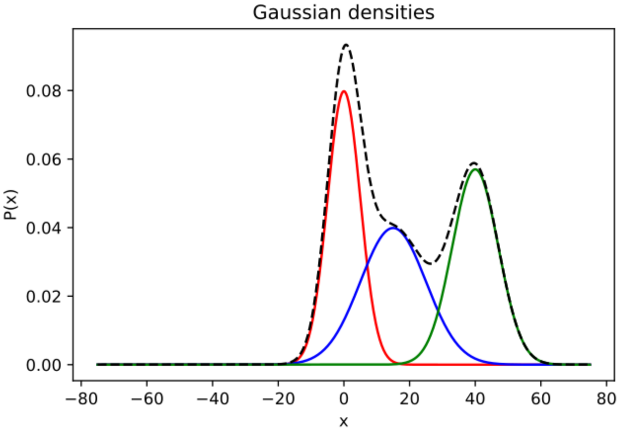


Figure: Gaussian mixture model.

Gaussian Mixture Model (GMM)

Bhattacharyya distance

The Bhattacharyya distance, D_b, measures the similarity of two probability distributions and is closely related to the Bhattacharyya coefficient. For example, if $f(x)$ and $g(x)$ are probability distributions, the D_b between them would be the way:

$$D_b(f(x), g(x)) = -ln(B_C(f(x), g(x))), \tag{16}$$

where B_C it is known as the Bhattacharyya coefficient and is defined for continuous probability distributions as

$$B_C(f(x), g(x)) = \int \sqrt{f(x)g(x)}. \tag{17}$$

Experiments and Results

Experiments

The following 3 experiments were carried out:

- Biclass classification (G1 vs G3) and triclass classification (G1 vs G2 vs G3) using the SVM classifier,
- Biclass classification (G1 vs G3) and triclass classification (G1 vs G2 vs G3) using the RF classifier, and
- Biclass classification (G1 vs G3) and triclass classification (G1 vs G2 vs G3) considering GMMs and Bhattacharyya distance.

Experiments

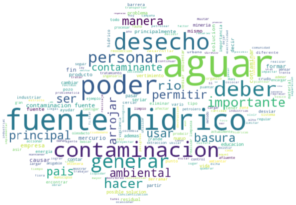
Also, it was taken into account the following:

- Cross validation (CV), of 10 partitions for the training process, that is, the data is divided into 10, chosen at random, 9 of them are used for training and 1 for test.
- It is used optimization of the parameters: C and γ (for SVM), *n-estimators* and *max-depth* (for RF), this, in order to obtain better results.
- Training with Task 1 using the parameters who obtained the best results in CV and test with Task 2.

Word Cloud

A

B



C

Figure: Word Cloud for users who made the Task 1. A) Group 1, B) Group 2, C) Group 3.

Word Cloud

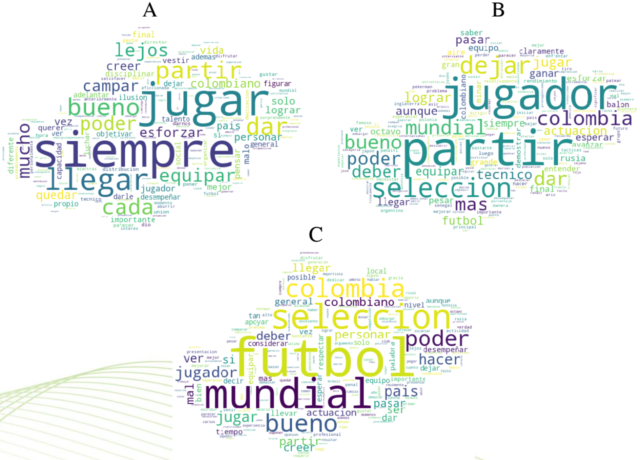


Figure: Word Cloud for users who made the Task 2. A) Group 1, B) Group 2, C) Group 3.

Dimensionality reduction using LDA.

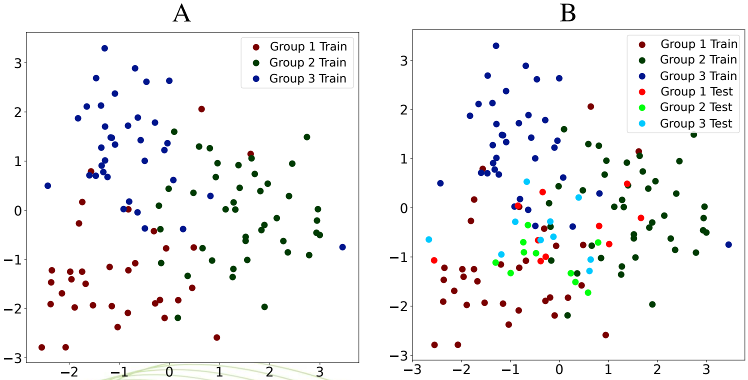


Figure: G1, G2 and G3 in a two-dimensional space of A) 111 users who performed Task 1 and B) 111 Users of Task 1 (“ Train ”) plus the 30 users who performed Task 2 (“ Test ”).

Results: SVM Biclass

Table: Classification with SVM of the texts of the G1 vs texts of the G3, training and testing with 75 users who carried out the Task 1.

Feature	K	C	γ	Acc (%)	F1 (%)	Sen (%)	Spe (%)	Mat
Fusión	Rbf	0.5	0.0001	64.1 ± 3.0	59.6 ± 3.9	75.9 ± 3.8	49.3 ± 8.5	[18 15] [12 30]
BoW	Rbf	0.5	1	63.9 ± 3.2	58.6 ± 3.6	77.7 ± 7.9	46.3 ± 5.8	[17 16] [9 33]
TF-IDF	Rbf	0.05	1	62.6 ± 4.2	57.6 ± 5.1	73.7 ± 6.4	48.8 ± 5.3	[18 15] [13 29]
Word2vec	Rbf	0.001	1	58.7 ± 2.1	48.5 ± 3.7	81.6 ± 7.3	28.4 ± 12.9	[7 26] [5 37]
GloVe	Rbf	10	1	**66.8 ± 2.5**	64.2 ± 3.4	78.2 ± 4.2	52.7 ± 6.3	[18 15] [9 33]
Grammatical	Linear	0.001	-	58.6 ± 1.3	50.8 ± 1.8	73.9 ± 5.8	39.8 ± 8.7	[14 19] [14 28]

Results: SVM Biclass

Table: Classification with SVM of the texts of the G1 vs texts of the G3, training and testing with 75 users who carried out the Task 1, applying LDA.

Feature	K	C	γ	Acc (%)	F1 (%)	Sen (%)	Spe (%)	Mat
Fusión	Rbf	0.5	0.0001	61.2 ± 1.9	52.9 ± 2.8	84.6 ± 3.7	31.0 ± 6.6	[18 15] [12 30]
BoW	Rbf	0.05	0.0001	61.4 ± 2.6	53.3 ± 4.1	83.5 ± 3.5	33.6 ± 7.1	[17 16] [9 33]
TF-IDF	Rbf	0.5	0.0001	59.2 ± 1.9	46.6 ± 3.6	93.7 ± 5.2	15.0 ± 9.4	[18 15] [13 29]
Word2vec	Linear	0.05	-	63.6 ± 2.6	61.4 ± 3.5	64.4 ± 6.5	62.6 ± 6.3	[7 26] [5 37]
GloVe	Linear	0.05	-	$\mathbf{65.5} \pm \mathbf{3.0}$	63.4 ± 3.6	73.2 ± 4.1	55.8 ± 8.1	[18 15] [9 33]
Grammatical	Rbf	0.5	0.0001	62.8 ± 2.5	59.4 ± 2.7	64.9 ± 6.6	60.1 ± 7.3	[14 19] [14 28]

Results: SVM Biclass

Table: Classification with SVM of the texts of the G1 vs texts of the G3, training with 75 users of Task 1 and testing with 20 users of Task 2.

Feature	K	C	γ	Acc (%)	F1 (%)	Sen (%)	Spe (%)	Mat
Fusión	Rbf	0.5	0.0001	50.0	33.3	100.0	0.0	[0 10] [0 10]
BoW	Rbf	0.5	1	50.0	33.3	0.0	100.0	[10 0] [10 0]
TF-IDF	Rbf	0.05	1	55.0	52.0	80.0	30.0	[3 7] [2 8]
Word2vec	Rbf	0.001	1	50.0	33.3	100.0	0.0	[0 10] [0 10]
GloVe	Rbf	10	1	**55.0**	54.9	50.0	60.0	[6 4] [5 5]
Grammatical	Linear	0.001	-	50.0	33.3	100.0	0.0	[0 10] [0 10]

Results: SVM Biclass

Table: Classification with SVM of the texts of the G1 vs texts of the G3, training with 75 users of Task 1 and testing with 20 users of Task 2, applying LDA.

Feature	K	C	γ	Acc (%)	F1 (%)	Sen (%)	Spe (%)	Mat
Fusión	Rbf	0.5	0.0001	50.0	33.3	100.0	0.0	[0 10] [0 10]
BoW	Rbf	0.05	0.0001	50.0	49.5	40.0	60.0	[6 4] [6 4]
TF-IDF	Rbf	0.5	0.0001	65.0	62.7	90.0	40.0	[4 6] [1 9]
Word2vec	Linear	0.05	-	60.0	59.6	70.0	50.0	[5 5] [3 7]
GloVe	Linear	0.05	-	**75.0**	74.4	90.0	60.0	[6 4] [1 9]
Grammatical	Rbf	0.5	0.0001	50.0	33.3	100.0	0.0	[0 10] [0 10]

Results: RF Biclass

Table: Classification with RF of the texts of the G1 vs texts of the G3, training and testing with 75 users who carried out the Task 1.

Feature	N_t	M_d	Acc (%)	F1 (%)	Sen (%)	Spe (%)	Mat
Fusión	5	2	62.9 ± 3.9	58.8 ± 4.6	68.9 ± 5.9	55.8 ± 7.8	[17 16] [9 33]
BoW	5	10	61.4 ± 1.9	53.5 ± 3.4	79.9 ± 9.3	37.7 ± 12.6	[9 24] [4 38]
TF-IDF	20	10	63.3 ± 3.0	56.6 ± 4.5	76.4 ± 5.8	46.9 ± 10.3	[14 19] [8 34]
Word2vec	15	1	64.4 ± 3.3	61.1 ± 3.4	70.7 ± 6.9	56.3 ± 8.2	[16 17] [8 34]
GloVe	10	1	$\mathbf{64.8 \pm 4.3}$	61.7 ± 4.9	66.8 ± 7.7	62.6 ± 11.1	[26 7] [17 25]
Grammatical	5	1	62.7 ± 2.5	58.8 ± 3.3	68.5 ± 3.5	55.3 ± 5.5	[20 13] [13 29]

Results: RF Biclass

Table: Classification with RF of the texts of the G1 vs texts of the G3, training and testing with 75 users who carried out the Task 1, applying LDA.

Feature	N_t	M_d	**Acc (%)**	**F1 (%)**	**Sen (%)**	**Spe (%)**	**Mat**
Fusión	10	1	62.4 ± 2.3	56.3 ± 3.2	76.3 ± 7.9	44.8 ± 10.1	[17 16] [10 32]
BoW	5	1	63.4 ± 2.5	57.6 ± 2.8	77.0 ± 7.9	45.9 ± 8.4	[21 12] [14 28]
TF-IDF	10	1	58.9 ± 2.1	46.9 ± 3.0	92.3 ± 4.8	16.7 ± 9.1	[8 25] [5 37]
Word2vec	5	1	64.2 ± 3.4	61.6 ± 3.9	66.1 ± 5.1	61.9 ± 5.9	[19 14] [12 30]
GloVe	5	1	**66.6 ± 2.1**	64.4 ± 2.3	71.2 ± 6.9	60.8 ± 8.7	[22 11] [14 28]
Grammatical	5	1	63.3 ± 1.5	60.3 ± 1.9	65.8 ± 3.2	60.0 ± 5.1	[21 12] [15 27]

Results: RF Biclass

Table: Classification with RF of the texts of the G1 vs texts of the G3, training with 75 users of Task 1 and testing with 20 users of Task 2.

Feature	N_t	M_d	Acc (%)	F1 (%)	Sen (%)	Spe (%)	Mat
Fusión	5	2	**65.0**	60.1	100.0	30.0	[3 7] [0 10]
BoW	5	10	55.0	48.7	20.0	90.0	[9 1] [8 2]
TF-IDF	20	10	60.0	56.0	30.0	90.0	[9 1] [7 3]
Word2vec	15	1	60.0	60.0	60.0	60.0	[6 4] [4 6]
GloVe	10	1	55.0	53.9	70.0	40.0	[4 6] [3 7]
Grammatical	5	1	60.0	52.4	100.0	20.0	[2 8] [0 10]

Results: RF Biclass

Table: Classification with RF of the texts of the G1 vs texts of the G3, training with 75 users of Task 1 and testing with 20 users of Task 2, applying LDA.

Feature	N_t	M_d	**Acc (%)**	**F1 (%)**	**Sen (%)**	**Spe (%)**	**Mat**
Fusión	10	1	50.0	33.3	0.0	100.0	[10 0] [10 0]
BoW	5	1	55.0	53.9	40.0	70.0	[7 3] [6 4]
TF-IDF	10	1	60.0	56.0	30.0	90.0	[9 1] [7 3]
Word2vec	5	1	**65.0**	64.9	70.0	60.0	[6 4] [3 7]
GloVe	5	1	65.0	60.1	100.0	30.0	[3 7] [0 10]
Grammatical	5	1	55.0	43.6	100.0	10.0	[1 9] [0 10]

Results: GMM Biclass

Classification with GMM and the group of features Word2vec of the texts of the G1 vs texts of the G3, training with 75 users of Task 1 and testing with 20 users of Task 2.

N_{Gauss}	Acc (%)	F1 (%)	Sen (%)	Spe (%)	Mat
2	55.0	53.9	40.0	70.0	[7 3] [6 4]
3	55.0	54.9	60.0	50.0	[5 5] [4 6]
4	**65.0**	60.1	30.0	100.0	[10 0] [7 3]
5	55.0	54.9	50.0	60.0	[6 4] [5 5]
6	60.0	59.6	50.0	70.0	[7 3] [5 5]
7	60.0	58.3	40.0	80.0	[8 2] [6 4]
8	50.0	33.3	0.0	100.0	[10 0] [0 10]
9	50.0	45.1	80.0	20.0	[2 8] [2 8]

N_{Gauss}	Acc (%)	F1 (%)	Sen (%)	Spe (%)	Mat
10	50.0	40.5	90.0	10.0	[1 9] [1 9]
11	35.0	33.5	20.0	50.0	[5 5] [8 2]
12	40.0	37.5	20.0	60.0	[6 4] [8 2]
13	55.0	48.7	20.0	90.0	[9 1] [8 2]
14	55.0	48.7	20.0	90.0	[9 1] [8 2]
15	60.0	52.4	20.0	100.0	[10 0] [8 2]
16	60.0	52.4	20.0	100.0	[10 0] [8 2]

Results: GMM Biclass

Classification with GMM and the group of features GloVe of the texts of the G1 vs texts of the G3, training with 75 users of Task 1 and testing with 20 users of Task 2.

N_{gauss}	Acc (%)	F1 (%)	Sen (%)	Spe (%)	Mat
2	50.0	33.3	0.0	100.0	[10 0] [10 0]
3	55.0	53.9	40.0	70.0	[7 3] [6 4]
4	50.0	45.1	80.0	20.0	[2 8] [2 8]
5	**75.0**	74.9	70.0	80.0	[8 2] [3 7]
6	65.0	62.7	90.0	40.0	[4 6] [1 9]
7	75.0	73.33	100.0	50.0	[5 5] [0 10]
8	55.0	53.9	70.0	40.0	[4 6] [3 7]
9	50.0	33.3	100.0	0.0	[0 10] [0 10]

N_{gauss}	Acc (%)	F1 (%)	Sen (%)	Spe (%)	Mat
10	50.0	33.3	100.0	0.0	[0 10] [0 10]
11	50.0	33.3	100.0	0.0	[0 10] [0 10]
12	50.0	33.3	100.0	0.0	[0 10] [0 10]
13	50.0	33.3	100.0	0.0	[0 10] [0 10]
14	50.0	33.3	100.0	0.0	[0 10] [0 10]
15	45.0	31.0	90.0	0.0	[0 10] [9 1]
16	45.0	31.0	90.0	0.0	[0 10] [1 9]

Results: SVM Triclass

Table: Classification with SVM of the texts of the G1 vs texts of the G2 vs texts of the G3 training and testing with the 111 users who carried out the Task 1.

Feature	K	C	γ	Acc (%)	F1 (%)	κ	Mat
Fusión	Rbf	0.001	0.0001	41.3 $\pm$ 2.8	38.4 $\pm$ 3.5	0.105 $\pm$ 0.044	[10 15 8] [7 24 11] [10 11 15]
BoW	Rbf	5	0.0001	39.9 $\pm$ 1.9	35.8 $\pm$ 1.9	0.077 $\pm$ 0.025	[7 18 8] [3 31 8] [8 18 10]
TF-IDF	Rbf	0.005	0.0001	39.9 $\pm$ 2.9	36.9 $\pm$ 2.9	0.081 $\pm$ 0.043	[7 16 10] [6 25 11] [8 15 13]
Word2vec	Rbf	10	0.0001	40.9 $\pm$ 2.8	36.9 $\pm$ 2.1	0.092 $\pm$ 0.041	[5 24 4] [5 31 6] [5 21 10]
GloVe	Rbf	10	0.0001	**43.7 $\pm$ 1.9**	41.1 $\pm$ 1.5	0.148 $\pm$ 0.025	[12 14 7] [6 22 14] [8 12 16]
Grammatical	Linear	0.05	-	34.2 $\pm$ 1.3	27.2 $\pm$ 1.7	-0.026 $\pm$ 0.020	[3 24 6] [1 31 10] [4 27 5]

Results: SVM Triclass

Table: Classification with SVM of the texts of the G1 vs texts of the G2 vs texts of the G3 training and testing with the 111 users who performed the Task 1, applying LDA.

Feature	K	C	γ	Acc (%)	F1 (%)	κ	Mat
Fusión	Rbf	0.5	0.1	38.8 $\pm$ 1.2	30.9 $\pm$ 1.7	0.037 $\pm$ 0.020	[7 22 4] [2 32 8] [3 28 5]
BoW	Rbf	0.5	0.1	35.9 $\pm$ 1.9	27.2 $\pm$ 1.8	-0.009 $\pm$ 0.033	[3 23 7] [3 34 5] [2 30 4]
TF-IDF	Rbf	1	0.1	35.9 $\pm$ 2.7	23.7 $\pm$ 2.1	-0.015 $\pm$ 0.039	[4 29 0] [3 35 4] [1 33 2]
Word2vec	Rbf	0.5	0.0001	35.4 $\pm$ 3.7	33.1 $\pm$ 3.6	0.024 $\pm$ 0.057	[10 12 11] [9 17 16] [8 12 16]
GloVe	Rbf	0.05	0.0001	**43.2 $\pm$ 1.9**	40.9 $\pm$ 2.0	0.135 $\pm$ 0.029	[12 15 6] [10 23 9] [7 15 14]
Grammatical	Rbf	5	0.0001	35.3 $\pm$ 4.3	31.9 $\pm$ 4.5	0.012 $\pm$ 0.066	[4 17 12] [5 24 13] [3 19 14]

Results: SVM Triclass

Table: Classification with SVM of the texts of the G1 vs texts of the G2 vs texts of the G3, training with the 111 users of Task 1 and testing with the 30 users of Task 2.

Feature	K	C	γ	Acc (%)	F1 (%)	κ	Mat
Fusión	Rbf	0.05	0.0001	36.7	27.9	0.050	[9 0 1] [7 1 2] [9 0 1]
BoW	Rbf	0.005	0.0001	**43.3**	39.6	0.149	[7 3 0] [5 5 0] [9 0 1]
TF-IDF	Rbf	0.005	0.0001	36.7	35.1	0.050	[6 3 1] [6 3 1] [7 1 2]
Word2vec	Rbf	10	0.0001	26.7	19.7	-0.100	[0 9 1] [3 7 0] [2 7 1]
GloVe	Rbf	10	0.0001	26.7	27.5	-0.100	[3 5 2] [7 2 1] [4 3 3]
Grammatical	Rbf	10	0.0001	33.3	16.7	0.000	[0 10 0] [0 10 0] [0 10 0]

Results: SVM Triclass

Table: Classification with SVM of the texts of the G1 vs texts of the G2 vs texts of the G3, training with the 111 users of Task 1 and testing with the 30 users of Task 2, applying LDA.

Feature	K	C	γ	Acc (%)	F1 (%)	κ	Mat
Fusión	Rbf	0.5	0.1	33.3	16.7	0.000	[0 10 0] [0 10 0] [0 10 0]
BoW	Rbf	50	0.1	40.0	38.3	0.099	[4 1 5] [3 2 5] [4 0 6]
TF-IDF	Rbf	1	0.1	43.3	38.3	0.150	[4 2 4] [2 1 7] [2 0 8]
Word2vec	Rbf	0.5	0.0001	**43.3**	41.2	0.150	[4 6 0] [2 7 1] [2 6 2]
GloVe	Rbf	0.05	0.0001	23.3	21.7	-0.149	[4 4 2] [8 2 0] [5 4 1]
Grammatical	Rbf	1	0.0001	30.0	24.0	-0.050	[1 8 1] [2 7 1] [1 8 1]

Results: RF Triclass

Table: Classification with RF of the texts of the G1 vs texts of the G3 training and testing with the 111 users who carried out the Task 1.

Feature	N_t	M_d	Acc (%)	F1 (%)	κ	Mat
Fusión	50	2	39.0 ± 3.7	35.1 ± 3.9	0.059 ± 0.056	[8 17 8] [6 28 8] [2 24 10]
BoW	15	10	38.9 ± 3.2	31.9 ± 3.9	0.045 ± 0.048	[5 22 6] [5 33 4] [4 26 6]
TF-IDF	50	10	38.6 ± 3.0	32.6 ± 3.3	0.046 ± 0.045	[6 15 12] [6 29 7] [2 24 10]
Word2vec	50	2	**39.9** ± **4.5**	36.9 ± 4.7	0.080 ± 0.068	[2 19 12] [3 30 9] [4 17 15]
GloVe	50	5	39.2 ± 2.9	36.2 ± 3.1	0.067 ± 0.047	[9 14 10] [5 24 13] [10 14 12]
Grammatical	5	1	33.8 ± 3.4	28.4 ± 3.5	-0.201 ± 0.052	[1 22 10] [3 25 14] [3 20 13]

Results: RF Triclass

Table: Classification with RF of the texts of the G1 vs texts of the G3 training and testing with the 111 users who performed the Task 1, applying LDA.

Feature	N_t	M_d	Acc (%)	F1 (%)	κ	Mat
Fusión	50	2	37.9 ±1.8	31.4 ± 2.1	0.036 ± 0.028	[5 23 5] [7 28 7] [7 19 10]
BoW	20	5	33.8 ± 2.6	27.3 ± 2.7	-0.029 ± 0.041	[3 18 12] [3 28 11] [6 23 7]
TF-IDF	15	5	33.9 ± 2.6	25.0 ± 2.5	-0.033 ± 0.036	[0 25 8] [1 29 12] [2 24 10]
Word2vec	15	2	33.7 ± 2.8	31.2 ± 3.5	0.003 ± 0.042	[11 12 10] [14 13 15] [9 13 14]
GloVe	50	2	**43.5 ± 3.6**	40.8 ± 3.5	0.139 ± 0.054	[10 16 7] [9 26 7] [6 16 14]
Grammatical	15	1	34.8 ± 3.0	30.8 ± 3.2	0.006 ± 0.046	[4 13 16] [8 21 13] [4 17 15]

Results: RF Triclass

Table: Classification with RF of the texts of the G1 vs texts of the G3, training with the 111 users of Task 1 and testing with the 30 users of Task 2.

Feature	N_t	M_d	**Acc (%)**	**F1 (%)**	κ	**Mat**
Fusión	50	2	**40.0**	28.2	0.099	[0 9 1] [0 10 0] [0 8 2]
BoW	15	10	16.7	14.8	-0.250	[0 9 1] [6 4 0] [3 6 1]
TF-IDF	50	10	40.0	37.8	0.099	[7 3 0] [7 3 0] [6 2 2]
Word2vec	50	2	26.7	25.2	-0.100	[2 4 4] [1 5 4] [2 7 1]
GloVe	50	5	20.0	19.1	-0.200	[1 8 1] [6 4 0] [6 3 1]
Grammatical	5	1	30.0	23.5	-0.050	[3 0 7] [6 0 4] [4 0 6]

Results: RF Triclass

Table: Classification with RF of the texts of the G1 vs texts of the G3, training with the 111 users of Task 1 and testing with the 30 users of Task 2, applying LDA.

Feature	N_t	M_d	**Acc (%)**	**F1 (%)**	κ	**Mat**
Fusión	50	2	36.7	26.0	0.050	[2 0 8] [4 0 6] [1 0 9]
BoW	20	5	33.3	25.9	0.00	[1 0 9] [0 1 9] [1 1 8]
TF-IDF	15	5	30.0	27.6	-0.050	[5 3 2] [5 1 4] [7 0 3]
Word2vec	15	2	**46.7**	46.0	0.199	[6 4 0] [0 5 5] [5 2 3]
GloVe	50	2	23.3	22.7	-0.149	[4 4 2] [8 1 1] [4 4 2]
Grammatical	15	1	40.00	40.9	0.099	[3 1 6] [5 4 1] [4 1 5]

Results: GMM Triclass

Classification with GMM and the group of features Word2vec of the texts of the G1 vs texts of the G2 vs texts of the G3, training with the 111 users of Task 1 and testing with the 30 users of Task 2.

N_{Gauss}	Acc (%)	F1 (%)	κ	Mat
2	30.0	27.6	-0.050	[1 6 3] [2 5 3] [1 6 3]
3	33.3	31.1	0.000	[2 4 4] [2 2 6] [3 1 6]
4	**46.7**	42.3	0.200	[9 1 0] [6 2 2] [7 0 3]
5	30.0	29.4	-0.050	[3 6 1] [4 1 5] [3 2 5]
6	30.0	30.6	-0.050	[2 7 1] [5 2 3] [3 2 5]

N_{Gauss}	Acc (%)	F1 (%)	κ	Mat
7	30.0	25.9	-0.050	[5 3 2] [7 0 3] [6 0 4]
8	30.0	15.8	-0.050	[9 1 0] [10 0 0] [9 1 0]
9	33.3	31.0	0.000	[1 6 3] [2 4 4] [1 4 5]
10	33.3	21.9	0.000	[1 1 8] [2 0 8] [1 0 9]
11	20.0	20.4	-0.200	[3 5 2] [9 1 0] [5 3 2]

N_{Gauss}	Acc (%)	F1 (%)	κ	Mat
12	20.0	21.00	-0.200	[3 6 1] [9 1 0] [5 3 2]
13	33.3	30.1	0.000	[1 8 1] [3 7 0] [4 4 2]]
14	33.3	30.1	0.000	[1 8 1] [3 7 0] [4 4 2]]
15	23.3	21.1	-0.150	[5 5 0] [9 1 0] [4 5 1]
16	36.7	27.9	0.050	[9 1 0] [9 1 0] [7 2 1]

Results: GMM Triclass

Classification with GMM and the group of features GloVe of the texts of the G1 vs texts of the G2 vs texts of the G3, training with the 111 users of Task 1 and testing with the 30 users of Task 2.

N_{Gauss}	Acc (%)	F1 (%)	κ	Mat
2	33.3	21.3	0.000	[9 1 0] [9 1 0] [10 0 0]
3	36.7	35.5	0.050	[5 2 3] [4 2 4] [4 2 4]
4	33.3	24.0	0.000	[2 0 8] [2 0 8] [2 0 8]
5	**53.3**	46.8	0.300	[8 0 2] [4 1 5] [3 0 7]
6	36.7	31.8	0.050	[2 4 4] [4 1 5] [1 1 8]
7	43.3	39.0	0.150	[4 2 4] [4 1 5] [0 2 8]
8	40	40.4	0.100	[3 2 5] [0 4 6] [1 4 5]
9	30.0	23.9	-0.050	[0 5 5] [0 4 6] [0 5 5]
10	30.0	23.9	-0.050	[0 5 5] [0 4 6] [0 5 5]
11	36.7	28.8	0.050	[0 5 5] [0 4 6] [0 3 7]
12	26.7	18.1	-0.100	[0 3 7] [0 1 9] [0 3 7]
13	26.7	18.1	-0.100	[0 3 7] [0 1 9] [0 3 7]
14	30.0	21.9	-0.050	[0 2 8] [0 2 8] [0 3 7]
15	23.3	16.7	-0.150	[0 2 8] [1 1 8] [1 3 6]
16	30.0	16.2	-0.050	[0 1 9] [1 0 9] [1 0 9]

UNIVERSIDAD DE ANTIOQUIA
1803

Conclusions

Conclusions

- The main objective is achieved, which is to find differences between the writing styles of the users belonging to the university community according to their school level, because a maximum efficiency in Biclass classification (G1 vs G3) of 75.0% is achieved and of 53.3% for Triclass classification (G1 vs G2 vs G3).

Conclusions

- The main objective is achieved, which is to find differences between the writing styles of the users belonging to the university community according to their school level, because a maximum efficiency in Biclass classification (G1 vs G3) of 75.0% is achieved and of 53.3% for Triclass classification (G1 vs G2 vs G3).
- In general, the best results are obtained with GloVe, which indicates that this type of feature is useful when you want to differentiate between texts by the way they are written and by their content.

Conclusions

- The main objective is achieved, which is to find differences between the writing styles of the users belonging to the university community according to their school level, because a maximum efficiency in Biclass classification (G1 vs G3) of 75.0% is achieved and of 53.3% for Triclass classification (G1 vs G2 vs G3).
- In general, the best results are obtained with GloVe, which indicates that this type of feature is useful when you want to differentiate between texts by the way they are written and by their content.
- If you want to distinguish between users with a low level of education and users with a high level of education, the indicated method to classify is considering an SVM or GMM. On the other hand, for triclass classification, GMM is superior (53.3% efficiency when classifying) to the SVM and RF approaches (43.3% and 46.7% respectively).

Future Work

As future work, it is proposed to extract features that take into account deeply the linguistic style of the users, such as lexical, syntactic, structural and content specific features from the original text, without carrying out any kind of pre-processing, and to measure again the performance with these features using the classification algorithms worked here.

References

Bishop, Christopher M. *Pattern recognition and machine learning*. springer, 2006.

C. Bellei. *The backpropagation algorithm for Word2Vec*. [Online; accessed 03-April-2019]. 2018. URL: http://www.claudiobellei.com/2018/01/06/backprop-word2vec/.

Gislason, Pall Oskar, Jon Atli Benediktsson, and Johannes R Sveinsson. "Random forests for land cover classification". In: *Pattern Recognition Letters* 27.4 (2006), pp. 294–300.

Kincaid, J Peter et al. "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel". In: (1975).

M. Díaz. *Una educación cada vez menos física*. [Online; accessed 23-Mayo-2019]. 2018. URL: https://www.elespectador.com/noticias/educacion/una-educacion-cada-vez-menos-fisica-articulo-735695.

N. S. Sarwan. *An Intuitive Understanding of Word Embeddings: From Count Vectors to Word2Vec*. [Online; accessed 01-April-2019]. 2017. URL: https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/.

P. Dubey. *An introduction to Bag of Words and how to code it in Python for NLP*. [Online; accessed 01-April-2019]. 2016. URL: https://medium.freecodecamp.org/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04.

References II

Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

Sandipan Dey. *Implementing a Soft-Margin Kernelized Support Vector Machine Binary Classifier with Quadratic Programming in R and Python*. [Online; accessed 22-Mayo-2019]. 2018. URL: https://www.datasciencecentral.com/profiles/blogs/implementing-a-soft-margin-kernelized-support-vector-machine.