Age Classification from Spanish Tweets
The Variable Age Analyzed by using Linear Classifiers

Luis G. Moreno-Sandoval[1], Joan Felipe Mendoza-Molina[1], Edwin Alexander Puertas[1],
Arturo Duque-Marín[1], Alexandra Pomares-Quimbaya[2] and Jorge A. Alvarado-Valencia[3]
[1]Colombian Center of Excellence and Appropriation on Big Data and Data Analytics (CAOBA), Bogotá, Colombia
[2]Department of Systems Engineering, Pontificia Universidad Javeriana, Bogotá, Colombia
[3]Department of Industrial Engineering, Pontificia Universidad Javeriana, Bogotá, Colombia

Keywords: SVM, SGD, Classification Problem, Age Classification, Twitter, Spanish.

Abstract: Text classification or text categorization in social networks such as Twitter has taken great importance with the growth of applications of this process in diverse domains of society. Literature about text classifiers is significantly wide especially in languages such as English; however, this is not the case for age classification whose studies have been mainly focused on image recognition and analysis. This paper presents the results of testing linear classifiers performance in the task of identifying Twitter users age from their profile descriptions and tweets. For this purpose, a Spanish Lexicon of 45 words around the concept "*cumpleaños*" was created and the Gold Standard of 1541 users with age correctly identified was obtained. The experiments are presented with the description of the algorithms used to finally obtain the best seven models that permit to identify the user's age with accuracy results between 66% and 69 %. Considering the information-retrieval layer, the new results showed that accuracy was increased from 69,09% to 72,96%.

1 INTRODUCTION

Text classification (TC) or text categorization consists of the process of automatically assigning one or more predefined categories to text documents. (Sun & Lim, 2001). Text classification problem has been widely studied by the database, data mining, machine learning and information retrieval practitioners and scientists because of its wide range of applications in diverse domains like news filtering and organizations, document organization and retrieval, opinion mining, email classification and spam filtering, among others (Aggarwal & Zhai, 2012).

The emergence and rapid growth in volume and extension of information on the web have led to extending statistical and machine learning classification techniques to text data categorization with the corresponding challenges in terms of the complexity of unstructured data. In this sense, special attention has given to linear classifiers such as neural networks and SVM classifiers because of their effectiveness in attending data characteristics and their possibilities of incorporation of linkage information into the classification process.

Age classification studies reported in the literature are mainly based on images recognition and analysis. According to Peersman et al., (2011), advances in natural language processing technology have enabled computational linguists to perform automatic linguistic analyses of the variation of linguistic characteristics in texts according to the authors' age or gender. However, age prediction studies have been focused on distinguishing, for example, adults from adolescents in long texts like blogs (Tam and Martel, 2009, Nguyen et al., 2011).

Twitter texts characteristics, however, present higher challenges for computational linguistics because short texts often contain non-standard language. Besides, most of the studies have been in the English language which establishes a research need in the field.

The results presented in this paper were developed within the Digital Segmentation project of Nutresa Business Group, in association with the Center of Excellence and Appropriation in Big Data and Data Analytics (CAOBA) (Vargas-Cruz, Pomares-Quimbaya, Alvarado-Valencia, Quintero-Cadavid,

Moreno-Sandoval, L., Mendoza-Molina, J., Puertas, E., Duque-Marín, A., Pomares-Quimbaya, A. and Alvarado-Valencia, J.
Age Classification from Spanish Tweets.
DOI: 10.5220/0006811102750281
In *Proceedings of the 20th International Conference on Enterprise Information Systems (ICEIS 2018)*, pages 275-281
ISBN: 978-989-758-298-1

& Palacio-Correa, 2017). This project uses analytical processes to characterize users of social networks as consumers through their online behavior. Considering the previous facts, this paper is focused on text classification in Spanish from twitter available information.

This paper presents an analysis of the effectiveness of 7 text classifiers in determining the age of users on Twitter. The principle is to enhance the recognition of users age from the application of an algorithm that uses a Spanish lexicon constructed around the concept "*cumpleaños*" in the category age.

We describe the steps we follow to test the performance of linear classification algorithms in the identification of users age in twitter. The paper presents the training database construction process in two parts (a) Identification and extraction of people accounts in twitter and (b) Identification of age expressions in accounts descriptions based on a Spanish Lexicon created around the concept of "*cumpleaños*". The process of analysis for the derived variable "age range" form the Gold Standard of 1541 users, yielded in the identification of seven classification models with accuracies between 66% and 69%. Considering the information-retrieval layer, the new results showed that accuracy was increased from 69,0972% to 72,963%.

The paper is organized as follows: the second part briefly describes previous related work on text classifiers. In the third section, we explain the methodology used to construct the database and the Spanish lexicon. Section four presents the experiments carried out and discusses the main results obtained. Finally, we outline conclusions and further work.

2 RELATED WORK

Aggarwal y Zhai defines the classification problem as follows: "We have a set of training records D = {X1..., XN}, such that each record is labeled with a class value drawn from a set of k different discrete values indexed by {1 ...k}. The training data is used in order to construct a classification model, which relates the features in the underlying record to one of the class labels. For a given test instance for which the class is unknown, the training model is used to predict a class label for this instance" (Aggarwal & Zhai, 2012, p.163).
Classification problem has been widely studied by data mining, database, and information retrieval communities. Duda et al., (2000, 2012) made an exhaustive review of pattern classification as Dumais y Chen (2000) did it for hierarchical classification. More recently Kotsiantis (2007) made a supervised machine learning classification techniques review.

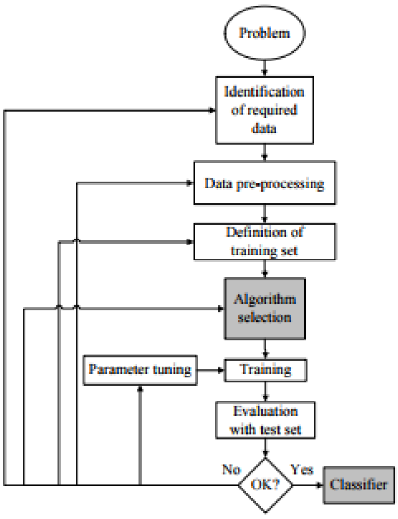


Figure 1: The process of supervised ML by Kotsiantis (2007).

According to the author, "Inductive machine learning is the process of learning a set of rules from instances (examples in a training set), or more, generally speaking, creating a classifier that can be used to generalize from new instances" (Kotsiantis,2007, p. 250). Based on the process of applying supervised ML to a real-world problem described in Figure 1, we analyze the different classifiers registered in the literature.

Kotsiantis (2007) proposes three categories for classifying techniques: Artificial Intelligence (Logical/Symbolic techniques), Perceptron-based techniques and Statistics (Bayesian Networks, Instance-based techniques), shown in Table 1.

Table 1: Supervised ML Techniques by Kotsiantis (2007).

Logical/Symbolic techniques	**Perceptron-based techniques**	**Statistics techniques**
Decision Trees Rule-based Classifiers	Single layered perceptrons	Bayesian networks
	Multilayered perceptrons	Instance-based learning
	Radial Basis Function (RBF) networks	Support Vector Machines

Aggarwal y Zhai (2012) present a classification in five categories: Decision Tree Classifiers, Rule-based Classifiers, Probabilistic and Naïve Bayes Classifiers, Proximity-based Classifiers and Linear Classifiers which include SVM Classifiers, Regression - based Classifiers, and Neural Network Classifiers.

For the purposes of the study, attention was centered on Linear Classifiers since the experiments were made with SVM, even first tests proved other classifiers of this group. Linear Classifiers have been developed independently, however, they are similar at a basic conceptual level and the main differences are in terms of the details of the objective function which is optimized, and the iterative approach used in order to determine the optimum direction of separation (Aggarwal y Zhai, 2012).

Support Vector Machines are based on the principle of determining separators in the search space which can provide the best separation of the different classes. According to Aggarwal y Zhai (2012), this is essentially a quite scalable semi-supervised approach because of its use of unlabeled data in the classification process and its use of a number of modified quasi-newton techniques, which tend to be efficient in practice.

On the other hand, age classification problem has registered in the literature studies specially dedicated to images analysis (Ramesha et al., 2010; Gao & Ai, 2009; Ylioinas et al.,2012; Levi & Hassner 2015, Rybintsev et al., 2015). Although some of them use the algorithms mentioned before, literature about text-based age classification is not broad.

3 METHODOLOGY

This section presents the steps we follow to test the performance of linear classification algorithms in the identification of users age in twitter. In the first section, we present the training database construction process in two parts (a) Identification and extraction of people accounts in twitter and (b) Identification of age expressions in accounts descriptions based on a Spanish Lexicon created around the concept of “*cumpleaños*”. The second section shows the process of analysis for the derived variable “age range”.

3.1 Training Database Construction

Before creating the classifier, it was necessary to have a database that allows not only to train a text classification model but also to extract age-related information from the account profile description. The information was collected from Twitter’s rest API during august, 2017, only in Spanish twitters. All the Datasets obtained were manually validated by the CAOBA experts group.

3.1.1 Algorithm for Identifying People Accounts from Twitter Rest API

The process carried out to develop the experiment consisted initially in the search of Twitter user accounts of people who were in Colombia, using as input the Twitter accounts of the 113 universities registered in the country. They were found 114.953 users linked, from which only 82.147 were active. On the other hand, 19,378 Colombian celebrity accounts were taken. In these accounts, 1,241,248 linked users were identified and it was possible to obtain data from 967,660 of them.

A validation process was carried out to identify that it was a profile related to a person, for which the construction of a manual list of 8701 people names was required. This analysis was performed on the *name* and *screen_name* fields of the account description through a process of BoW (Bag of Words) (Moreno et al., 2017). It was possible to determine if it was a person's account, this means it had the name of a person associated or it was a corporate account. In total, 50,819 accounts of people linked to universities and 734,037 accounts of people linked to celebrities were obtained.

3.1.2 Algorithm for Identifying Ages Expressions in Accounts Descriptions from Spanish Lexicon.

Once the accounts of people to analyze were identified, a lexicon of Spanish expressions related to the theme "*cumpleaños*" in age category was created obtaining 50 words. Having this lexicon constructed, we proceeded to analyze the descriptions of each user accounts using BoW techniques. 1159 accounts linked to the universities that had information associated with age texts and 41183 accounts linked to celebrities related to the same lexicon were identified.

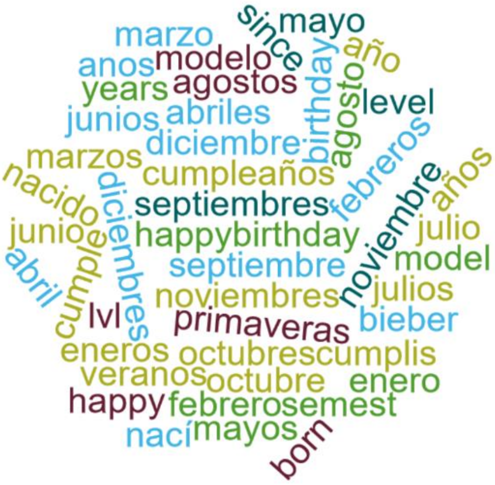
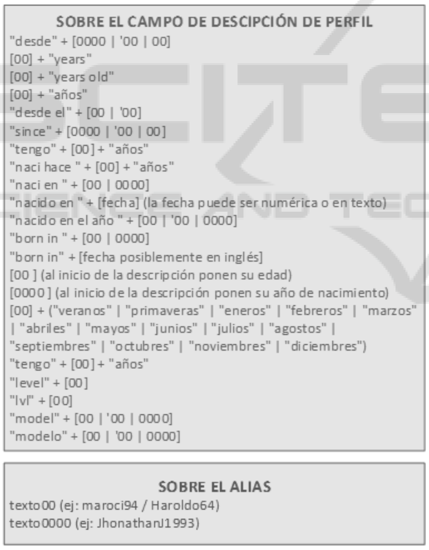


Figure 2: Spanish Lexicon “Cumpleaños”.

The algorithm takes regular expressions from the text, exploring linguistic rules that link the lexicon expressions with a numeric data related to age or birth year. Figure 2 shows the age patterns used to the analysis resulting from the rules established.



SOBRE EL CAMPO DE DESCIPCIÓN DE PERFIL

"desde" + [0000 | '00 | 00]
[00] + "years"
[00] + "years old"
[00] + "años"
"desde el" + [00 | '00]
"since" + [0000 | '00 | 00]
"tengo" + [00] + "años"
"naci hace " + [00] + "años"
"naci en " + [00 | 0000]
"nacido en " + [fecha] (la fecha puede ser numérica o en texto)
"nacido en el año " + [00 | '00 | 0000]
"born in " + [00 | 0000]
"born in" + [fecha posiblemente en inglés]
[00] (al inicio de la descripción ponen su edad)
[0000] (al inicio de la descripción ponen su año de nacimiento)
[00] + ("veranos" | "primaveras" | "eneros" | "febreros" | "marzos" | "abriles" | "mayos" | "junios" | "julios" | "agostos" | "septiembres" | "octubres" | "noviembres" | "diciembres")
"tengo" + [00] + "años"
"level" + [00]
"lvl" + [00]
"model" + [00 | '00 | 0000]
"modelo" + [00 | '00 | 0000]

SOBRE EL ALIAS

texto00 (ej: maroci94 / Haroldo64)
texto0000 (ej: JhonathanJ1993)

Figure 3: Age patterns extraction.

With the list of users fully automated whom the algorithm assigned an age, it was necessary to manually review to validate the accuracy of the result obtained. The automated list delivered 7359 users with a numerical value of age. A random selection was made of 2500 of these. The exact age for 1541 users was obtained. These results were combined with other sources of information from which the age of users was known and validated. Table 2 shows the summary of sources used and table 3 the distributions of users according to age ranges.

Table 2: Distribution of sources with age identified.

Source	Users
CAOBA Survey – Q1 2017	81
J.F. Mendoza’s survey – Q2 2017	119
Manual Inspection	524
Information Retrieval	1541
Total	2265

Table 3: Distribution of ages identified.

Age	Users
13-17	211
18-24	1067
25-34	566
35-49	267
50-64	109
65-xx	45
Total	2265

This means that our Gold Standard consists of 2265 users with age correctly identified, where 1541 of them were automatically identified using our information-retrieval algorithm.

3.1.3 Users Database Creation

The next step consisted on download tweets from the 2265 users with age identified and validated using two scripts in Python: one downloading the first 3240 tweets of each user (as maximum), and the second with the profile description of the users. An ETL using Pentaho Kettle 7 was created with the following operations:

- Load raw tweets previously downloaded (Load RAW tweets).
- Identify retweets (identify RT).
- Load raw tweets previously downloaded (Load RAW tweets)
- Concatenate tweets by user, excluding those that are RT (Create Test DB)
- Create a test database including username, age range, profile descriptions (previously downloaded), and concatenated tweets in a single document per user (Create Test DB).

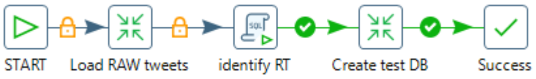


Figure 4: ETL steps to obtain the test database.

3.2 Derived Variable "Age Range"

The derived variable is composed of two main parts. The first is an algorithm that feeds from the training database, with the aim of training a tweet classification model; the second, is a 2-layer algorithm to classify users based on the model obtained in the previous step (*text-classifier layer*), as well as using extraction of information related to age from the description of the users and their alias (*information-retrieval layer*, based on the algorithm for identifying ages expressions in accounts descriptions from Spanish Lexicon).

3.2.1 Procedure to Train the Text Classification Model

The procedure to define the text classification model works as follows:

1. The number of categories to be classified is reduced in the training database. The database comes with six age categories ('13 -17 ', '18 -24', '25 -34 ', '35 -49', '50 -64 ', '65 onwards'), but to balance the number of actual occurrences by category, it was decided to group three categories ('13 -24 ', '25 -34', '35 onwards').
2. Training and test sets are defined, and the algorithm of training is defined and fed with the following fields:
   - *nickname:* name or alias that the user has on Twitter.
   - *description:* description of the user on Twitter.
   - *user_text: a* document that contains all the tweets of a user, previously concatenated and cleaned.
   - *file_trained_model:* text classification model obtained with the *train_text_classifier function*.

4 RESULTS

One of the main challenges of the text classifier is to find an algorithm that manages to assign users at an age range in the best possible way (that is, with the best accuracy) from some type of pattern analysis on the tweets they write.

Considering the text classification packages included in the scikit-learn package in Python, the families of Naive Bayes (NB) algorithms, Stochastic Descendant Gradient (SGD) and Support Vector Machines (SVM) seemed to be the best candidates to be evaluated. Thus, a code was designed to evaluate various models generated from each of these three families of algorithms by changing the values of their parameters.

The code of the function that allows to train and select the best text classification model was fed with the following fields:

- *TrainSet_data:* list in which each of its elements is a document that contains all the tweets of a user, previously concatenated and cleaned. Thus, each element represents a different user.
- *TrainSet_target*: list in which each of its elements contains the age range, corresponding to the same user in the exact order in which the elements of the TrainSet_data list appear.

4.1 Obtaining the Best Classification Models

When doing observation runs only for the *text-classifier layer*, it was noticeable that almost always NB models offered the worst results of accuracy, so they were discarded in the first instance. Hence, only SVM and SGD models where considered, running tests for 120 models (12 for SVM and 108 for SGD). For each run (from a total of 30 replicates), the test database was divided randomly into a training set and a test set, and then all the 120 models were evaluated. Therefore, considering their accuracy performance, the number of options was reduced to the best models obtained from the combination of the following parameters, obtaining 9 possible models:

Table 4: Parameters for classification models.

	SGD	**SVM**
Par. 1	loss = ['hinge', 'log', ‘huber’] *(discarded loss; ['modified_huber', 'squared_hinge', 'perceptron', 'squared_loss', 'epsilon_insensitive', 'squared_epsilon_insensitive'])*	Cost = [1,5,10] *(discarded Cost = [100])*
Par. 2	penalty = ['none','l2'] *(discarded penalty = ['l1','elasticnet'])*	kernel = ['linear'] *(discarded kernel = ['poly', 'sigmoid'])*
Par. 3	learning_rate = ['optimal'] *(discarded learning_rate = ['constant', 'invscaling'])*	
Total models	**6 (from 108)**	**3 (from 15)**

When running a test with 100 replicates, each one fed with a random training set close to 750 cases and a random test set close to 250, the following results were obtained in terms of accuracy as a measure of performance, only considering the text-classifier layer. The first tests showed that *SGD ‘huber’´mode*l reported the worst results with an accuracy of 60% for *SGDC_huber_none_optimal* and 55% for *SGDC_huber_l2_optimal*. So, the experiment finally provided 7 models with a significant performance:

Table 5: The 7 best classification models.

Model	Accuracy
SGDC_hinge_none_optimal	69,0972%
SVM_linear_5	68,9531%
SGDC_hinge_l2_optimal	67,3540%
SVM_linear_10	66,9039%
SGDC_log_l2_optimal	66,6667%
SGDC_hinge_none_optimal	66,3366%
SVM_linear_5_1	66,3194%

Afterwards, and the additional test was run considering the best model (*SGDC_huber_none_optimal*), this time considering the information-retrieval layer. The new results showed that accuracy was increased from 69,09% to 72,96%.

5 CONCLUSIONS AND FUTURE WORK

The experiments to identify the Twitter users age presented in this paper are based on the need for contributing in the research on the classification problem in the Spanish language, specifically in texts coming from social network interactions.

To achieve these purposes, it was necessary to pre-process 50,819 accounts of people linked to universities and 734,037 accounts of people linked to celebrities. A set of algorithms was created to identify people accounts and eliminate corporate accounts. Likewise, BOW (Bag of Words) process was used to analyze people accounts descriptions and tweets in order to apply an algorithm based on the Spanish lexicon around the concept “*cumpleaños*” also created for the experiment.

In this part of the experiments, it was possible to obtain 1541 users with age automatically assigned by the algorithm and manually validated which, together with other sources, confirmed a Gold Standard made up of 2265 users. Once obtained it, experiments permit us to validate the effectiveness of 120 models in the first instance and finally identify 9 models from 3 parameters in SGD and SVM classifiers. Two of these models were discarded to have a result of seven models reporting accuracy between 66% and 69%. After applying an additional layer to extract information from users’ alias and description, the maximum value of accuracy went up to 73%.

In conclusion, this work contributes not only in the evaluation of a big quantity of classifier models to identify the ones with the best performance but in creating a Spanish Lexicon of 45 words around Age topic and in constructing the first version of a Gold Standard to test other algorithms.

For future works, the Spanish Lexicon so as the Gold Standard could be tested into another kind of models and other sophistications, which would help to understand the performance of linear classifiers with different algorithms. Challenges in automatically classifying texts shared in social networks are still complex and studies in the Spanish Language need to continue advancing.

ACKNOWLEDGEMENTS

This research was carried out by the Center of Excellence and Appropriation in Big Data and Data Analytics (CAOBA). It is being led by the Pontificia Universidad Javeriana Colombia and it was funded by the Ministry of Information Technologies and Telecommunications of the Republic of Colombia (MinTIC) through the Colombian Administrative Department of Science, Technology and Innovation (COLCIENCIAS) within contract No. FP44842-anex46-2015.

REFERENCES

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. *Mining text data*, 163-222.

Dumais, S., & Chen, H. (2000, July). Hierarchical classification of Web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 256-263). ACM.

Gao, F., & Ai, H. (2009, June). Face age classification on consumer images with gabor feature and fuzzy LDA method. In *International Conference on Biometrics* (pp. 132-141). Springer, Berlin, Heidelberg.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

Levi, G., & Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 34-42).

Moreno-Sandoval L., Beltrán-Herrera P., Vargas-Cruz J., Sánchez-Barriga C., Pomares-Quimbaya A., Alvarado-Valencia J. and García-Díaz J. (2017). CSL: A Combined Spanish Lexicon - Resource for Polarity Classification and Sentiment Analysis. In Proceedings of the 19th International Conference on Enterprise Information Systems - Volume 1: ICEIS, ISBN 978-989-758-247-9, pages 288-295.

Nguyen, D., Smith, N. A., & Rosé, C. P. (2011, June). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 115-123). Association for Computational Linguistics.

Peersman, C., Daelemans, W., & Van Vaerenbergh, L. (2011, October). Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents* (pp. 37-44). ACM.

R. Duda, P. Hart, W. Stork. Pattern Classification, Wiley Interscience, 2000.

Ramesha, K., Raja, K. B., Venugopal, K. R., & Patnaik, L. M. (2010). Feature extraction based face recognition, gender and age classification.

Rybintsev, A. V., Lukina, T. Y. M., Konouchine, V. S., & Konouchine, A. S. (2013). Age estimation upon face image based on local binary patterns and a ranking approach. *Sistemy i Sredstva Informatiki [Systems and Means of Informatics]*, *23*(2), 62-73.

Sun, A., & Lim, E. P. (2001). Hierarchical text classification and evaluation. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE* International Conference on (pp. 521-528). IEEE.

Tam, J., and Martell, C. 2009. Age Detection in Chat. In Proceedings of the 3rd IEEE International Conference on Semantic Computing. (Berkeley, USA, September 14-16, 2009). DOI=10.1109/ICSC.2009.37.

Vargas-Cruz, J., Pomares-Quimbaya, A., Alvarado-Valencia, J., Quintero-Cadavid, J., & Palacio-Correa, J. (2017). Desarrollo de un Sistema de Segmentación y Perfilamiento Digital. *Procesamiento del Lenguaje Natural*, *59*, 163-166.

Ylioinas, J., Hadid, A., & Pietikäinen, M. (2012, November). Age classification in unconstrained conditions using LBP variants. In *Pattern recognition (icpr), 2012 21st international conference on*(pp. 1257-1260). IEEE.