Characterizing Influential Leaders of Ecuador on Twitter Using Computational Intelligence

JohnnyTorres*, GabrielaBaquerizo†, CarmenVaca* and Enrique Peláez*
*Escuela Superior Politécnica del Litoral (ESPOL)
Emails: {jomatorr,cvaca,epelaez}@fiec.espol.edu.ec
†Universidad Casa Grande
Email: gbaquerizo@casagrande.edu.ec

***Abstract*—The popularity of social networks, such as Twitter, have provided users around the world the ability to share information, express opinions or sentiments about any topic. Twitter has become the preferred social network platform used by researchers for measuring popularity or influence of users in social networks. This study seeks to extend analysis of influential users in the spatial context of Ecuador, applying computational intelligence techniques in order to identify influential users and for those users calculate its ranking. The results show that a careful selection and normalization of features found in the twitter user's profile, allows us to detect influential users with high degree of accuracy, and then calculate the ranking only over those users. This approach provide a quicker method compared to previous techniques for determining the ranking by filtering non-influential users.**

I. Introduction

The widespread dissemination of information on Internet and the development of new interaction platforms like blogs, microbloggings and social networks generated that users have greater access to information from a variety of themes [1]. At the same time, it allows users to share information, express their opinions and feelings about any subject.

The impact of social networks has led the researchers interested in the phenomenon to propose methodologies to study among other things, users with high degree of influence over other users known as "followers" and, development some strategies to how measure it. These "influentials" generate information that become viral [2] and propagate in cascade through the social network. Researchers have observed that although the number of followers is a good indicator to identifying these users, but isn't an absolute predictor of influence [1], [3], [4].

Influential leaders in social networking has become an interesting topic of research in recent years as its users have raised dramatically all over the world [5]. Some researchers say that users who post large amounts of information in the network and have a large number of followers are called 'influencers'. Other authors call 'influencers' to people that 'produce content that is relevant to a significant portion of the community' [6].

These 'influential' people some times are called opinion leaders. For Rogers [7] an opinion leader is an individual who influence in others' opinions, their character and conduct constantly. He proposes the differentiation between opinion leader and their followers and says that 'opinion leaders have a greater exposure to mass media than their followers; they are more cosmopolite than their followers and; they have greater change agent contact than their followers'.

This goes together with the proposal of traditional approach that 'assumes that a minority of members in a society possess qualities that make them exceptionally persuasive in spreading ideas to others. These exceptional individuals drive trends on behalf of the majority of ordinary people' [1].

If we transfer the mass media traditional measuring system to social Networks, we will measure the number of followers by user and we could say if we have more followers we are more influential. However, [8] the mechanism of influence doesn't only relate with followers of a user, the really important thing is considering the interactions that the follower has with the tweets that opinion leader published, because the information that they share is interesting to their followers [9].

Another view of opinion leader definition is that those are people that aren't a 'head formal organizations nor are they public figures such as newspaper columnists, critics, or media personalities, whose influence is exerted indirectly via organized media or authority structures' [10]. They are 'individuals who are highly informed, respected, or simply connected' [10].

However, for this study we have considered an initial list of 20 leaders of public opinion in Ecuador which in most cases are politicians, celebrities, journalists, athletes, etc. People who have a large number of followers and are very active on twitter. The goal of this study is to identify influential users through computational intelligence techniques on Twitter in the spatial context of Ecuador. An overview of related work in this area in the section II. Then the dataset used in the experiments is explained in section III. Next the methodology for identifying influential users is discussed in section IV. Finally, in section V the conclusions are presented and future work is proposed.

II. Related Work

There are some studies from various fields that are seeking to identify who are the influencers on social networks. Silva et.al [6] suggest detecting leader users from relevant information posted by the users. They proposed ProfileRank, a model inspired in PageRank 'that exploits the principle that relevant content is created and propagated by influential users

and influential users create relevant content' [6]. These authors investigate influence and relevance of information.

On marketing cases its very important the study of consumer behavior. Which brands consumers buy? and Why they prefer some ones over others? In that situation the decision of influential people is invaluable for customers. In this sense influence is 'contextual and temporal, depending on the subject, the speaker's credibility, and a variety of other factors' [11]. Identifying influencers is an important tool for viral marketing strategies. The consumption behavior of influential people is invaluable for purchase decision of customers.

At the same time, there are several studies that seek to characterize influential, that is to define and measure the influence of individuals and to know the dynamics of spread of information. A methodology to measure the user's influence on Twitter has been proposed by Weng et al. [12], TwitterRank, as an extension of the famous PageRank algorithm used initially by Google to measure the influence of users taking into account both the topical similarity between them and the link structure. The authors mention important connotations in the Twitter platform, which is based on the model of "following", in which each user selects another user to follow, it was found that 72.4% of the users follow more than 80% of their followers signaling the presence of "reciprocity".

Topical authorities in microblogging services such as Twitter has been studied by Pal et al. [2], emphasizing the diversity of hundreds of millions of users in Twitter, consider not a weakness but a strength, with its own challenges. One of the main concerns for a user in Twitter is to find interesting and respected authors for any given topic. Addressing this challenge, the authors proposed the establishment of a set of features for characterizing social media authors for nodal and topical metrics, basically employing a probabilistic clustering over the content generated by the users, plus a ranking procedure inside each cluster.

III. Dataset

The dataset used in this study has been generated using the application programming interface (API) provided by Twitter [13], specifically the RESTful API which is a lightweight framework compared to SOAP Web Services for exposing massive data into the Web. With the use of this API, 10K (thousands) user's profiles and 2.1 Million tweets have been downloaded from June/2015 until Aug/2015. JSON is the format provided by Twitter for its RESTful services, and this format has been collected in a local database with a size just over 10GB. The initial seed of influential users were obtained and consolidated from specialized websites in Twitter Statistics like Social Bakers [14] and local providers like a consulting company Llorente & Cuenca [15]. After the data was collected the user's profiles were manually labeled by type of user as considered influential or not according to the rankings in the websites analyzed and the metrics observed for each user.

IV. Identifying Influentials

In order to address the objective of this study, the identification of influential users of Ecuador in Twitter, a basic analysis was performed on the dataset to understand the composition and its distribution. The feature selection on the raw data obtained of Twitter was defined, using a specific number of variables as shown in table I.

Table I: Dataset Variables

User Profile	
Name	*Full Name.*
Screen Name	*Twitter username.*
Created At	*Creation Date.*
Statuses Count	*Number of tweets created or retweeted by the user.*
Favorites Count	*Number of tweets favorite by the user.*
Followers Count	*Number of followers of the user.*
Friends Count	*Number of followings of the user.*
Listed Count	*Number of list created by the user.*
Protected	*If true, user's tweets are only visible to followers.*
Geo-Enabled	*If true, each tweet is geo-localized.*
Verified	*Used to establish authenticity of identities of key individuals and brands.*
Tweets	
Created By	*User who created or retweeted a tweet*
Created At	*Date and time of the tweet creation*
Favorite Count	*Number of favorites received*
Retweet Count	*Number of retweets received*

For this and the subsequent steps, the analysis was based on a machine learning platform for Python known as scikit-learn [16]. Figure 1 shows features scored based on ANOVA analysis, based on that score, 20% of the lowest scoring percentage were removed. The following features were discarded: Geo-Enabled, Contributors Enabled, Favorites Received.

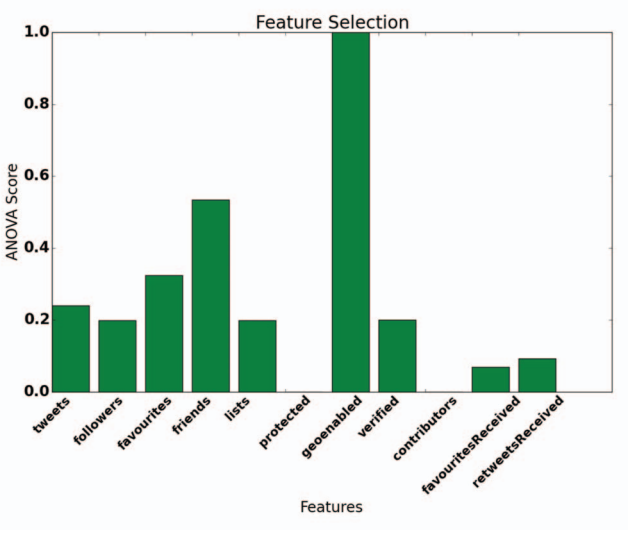

Figure 1: Feature selection based on ANOVA.

The next step in this analysis consists on preprocessing the data to generate the matrix of values. The preprocessing includes scaling and normalizing the features. Figure 2 shows

a histogram of the number of followers, a feature that measure the level of popularity of a user. The values have been normalized using $\log_{10} a$ to avoid bias towards popular users in the later analysis, and improve the prediction accuracy.

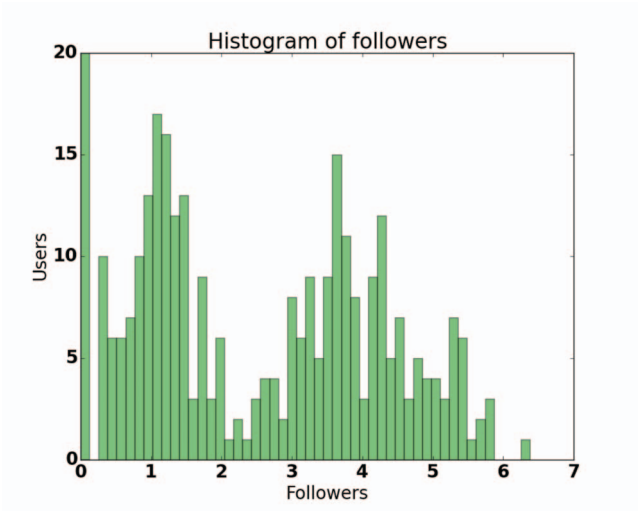

Figure 2: Histogram of normalized number of followers.

A. Dimensionality

To be able to visualize the data, a principal component analysis (PCA) [17] was performed ensuring variability of the data is maintained in a minimum threshold of 80%. At the beginning a high degree of variability in the dataset was observed, due to the presence of outliers corresponding to popular users. Figure 3 shows a PCA visualization with the normalization and standardization of the features, blue dots indicate influential users. After applying PCA the variability of the dataset is maintained up to 86%.

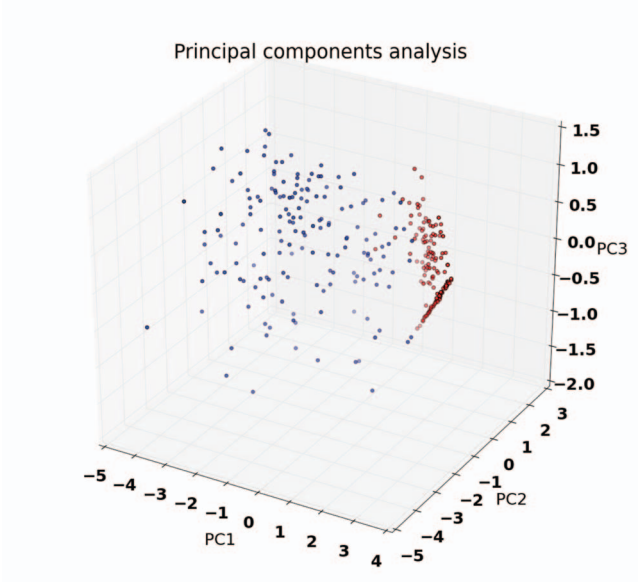

Figure 3: PCA dimensionality reduction.

B. Location

In order to filter the users from Ecuador, the field "Location" was used. The content of this field was compared using regular expression again a list of cities and countries. The analysis shows that 70.71% of the users have defined the location in their profile information. Of those users with location information, 82.75% belongs to users of Ecuador or its cities.

C. Classification of Influential Users

For classifying the influential users a comparative analysis was conducted on the dataset, the following machine learning algorithms for binary classification were used: Support Vector Machines (SVM), Nearest Neighbors, Naive Bayes.

Support Vector Machines [18] consist a set of supervised learning algorithms used for classification, regression, and outliers detection. SVM decision function on classification depends on some subset of the training data, called the support vectors. The kernel used can be used according to the problem at hand, the options are linear: polynomial, radial basis function (RBF), sigmoid, precomputed, among others. In our experiments the RBF kernel was used because presented better results compared with other kernels.

Nearest Neighbors [19] Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point.

Naive Bayes [20] consist of a set of supervised learning algorithms based on applying Bayes theorem with the "naive" assumption of independence between every pair of features.

The dataset was split 60% for training the classifiers and 40% for testing purposes. The experiments showed the classification without normalization of the features scores prediction accuracy between 84% and 86%, while using normalization the score improve up to 99%. Prediction scores in the classification task were calculated on test data. The first plot in figure 4 corresponds to the dataset without any classification, while next plots to the right shows the decision boundary for each of the algorithms used in the experiments.

D. Ranking

To calculate the ranking, we consider three coefficients for those users classified as influentials. First, the volume of activity refers to how much activity the user is having in Twitter over a period, and it is given by the tweets created by the user or retweets of content from another user, plus the number of favorites given to other tweets. Due the activity of favorites is less important for an influential user, it's given multiplied by a factor to lower its weight on the total activity volume for a user.

$$V_{\text{ac}} = \log\left(tweets + \sqrt{favs}\right) \tag{1}$$

The attention volume refers to the attention the followers are given to the user based on the tweets created or retweets.

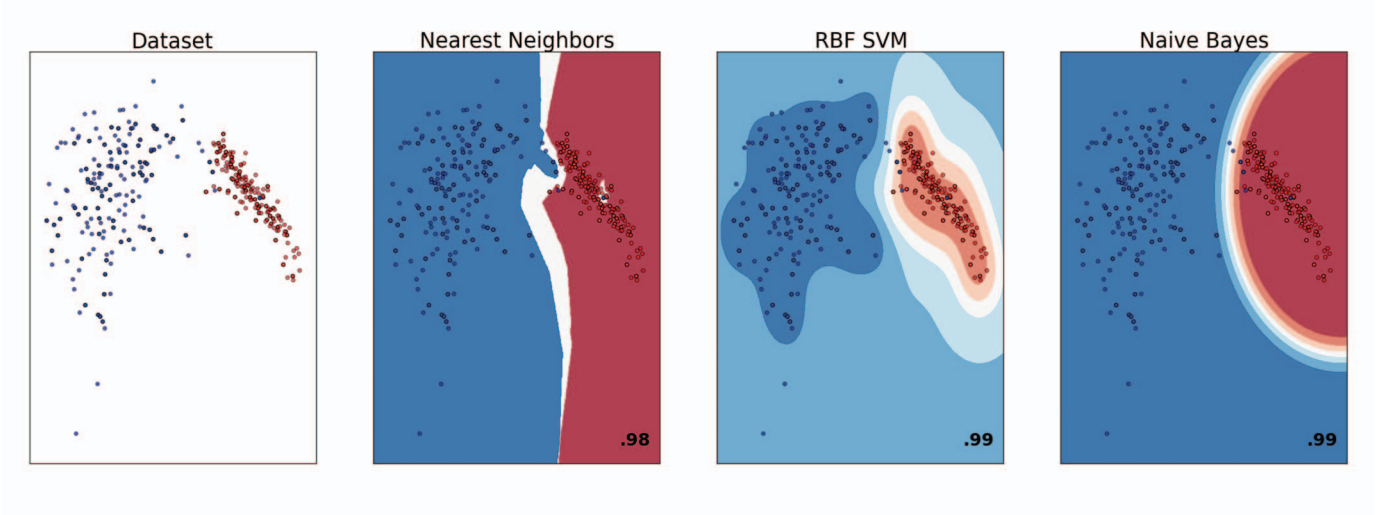

Figure 4: Comparative analysis of algorithms for Twitter's user classification.

Table II: Ranking comparison

Position	Name	Ranking LC
1	MashiRafael	77.9
2	ppsesa	50.4
3	jaimenebotsaadi	47.6
4	MauricioRodasEC	45.7
5	CarlosVerareal	44.4
Position	**User**	**Ranking**
1	MashiRafael	165.9
2	ppsesa	133.2
3	jaimenebotsaadi	126.5
4	CarlosVerareal	125.4
5	jimmyjairala	121.6

$_{o}$ denotes the content created by the user, and $_{n}$ non-original content.

$$V_{at} = \log{(rt + fv)_o} + \log{\sqrt{(rt + fv)_n}} \quad (2)$$

N_i refers to the network influence, which indicates the relationships the users have, meaning followers and friends, and its weight over the influence of a user in Twitter.

$$N_i = log(fo) \quad (3)$$

The ranking in equation 4 is calculated using activity index (1), attention index (2), and network influence (3).

$$R_i = (N_i + V_{ac}) * V_{at} \quad (4)$$

Figure 5 shows the ranking for some of the users, it was found that 80% of the top 5 most influential users are similar to those found in referential site Llorente-Cuenca (LC) for year 2015 as shown in the table II. The ranking positions showed a similarity when compared to previous techniques like TwitterRank, although in that technique the ranking is calculated in relation to specific topics, and not a global ranking.

Figure 5: Ranking for influential users in Ecuador.

V. Conclusions and Future Work

The results of the classification tasks for all three algorithms show high accuracy score when using normalized data. Table III shows Support Vector Machine performs slightly better compared to others algorithms.

Table III: Classification accuracy comparison

Algorithm	Accuracy score
Nearest Neighbors	98%
Support Vector Machine	99%
Naive Bayes	98%

It was observed that some user's profile features have little effect on the accuracy of the algorithms for classification as shown in figure 1. The ranking based on users classified as influentials shows similar results compared with referential sources and other techniques as TwitterRank, but the proposed methodology is more scalable because the ranking is calcu-

lated only over the influential users predicted by the classifier instead of using the entire dataset.

In a future work, we will consider an extension of the characterization of users on social networks, including topology and cognitive maps analysis to understand the factors and relations between users, as well ranking scores based on topics and trends that arise in Twitter.

REFERENCES

[1] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in twitter: The million follower fallacy." *ICWSM*, vol. 10, no. 10-17, p. 30, 2010.

[2] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 45–54.

[3] S. Asur, B. A. Huberman, G. Szabo, and C. Wang, "Trends in social media: Persistence and decay," *Available at SSRN 1755748*, 2011.

[4] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *Machine learning and knowledge discovery in databases*. Springer, 2011, pp. 18–33.

[5] INTEL. (2013) What happens in an internet minute? INTEL. [Online]. Available: http://goo.gl/ML21F2

[6] A. Silva, S. Guimarães, W. Meira Jr, and M. Zaki, "Profilerank: finding relevant content and influential users based on information diffusion," in *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM, 2013, p. 2.

[7] E. M. Rogers, *Diffusion of innovations*. Simon and Schuster, 2010.

[8] L. Bruni, "A methodological framework to understand and leverage the impact of content on social media influence," Ph.D. dissertation, Italy, 2014.

[9] C. Francalanci and I. Metra, "Content-based discovery of twitter influencers."

[10] D. J. Watts and P. S. Dodds, "Influentials, networks, and public opinion formation," *Journal of consumer research*, vol. 34, no. 4, pp. 441–458, 2007.

[11] H. Jenkins, S. Ford, and J. Green, *Spreadable media: Creating value and meaning in a networked culture*. NYU Press, 2013.

[12] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 261–270.

[13] Twitter rest api para desarrolladores. [Online]. Available: https://dev.twitter.com/rest/public

[14] Social baker estadísiticas y marketing en redes sociales. [Online]. Available: http://www.socialbakers.com/

[15] C. L. . Cuenca. (2015) Mapa del poder. [Online]. Available: http://www.mapadepoderecuador.com/

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[18] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *Intelligent Systems and their Applications, IEEE*, vol. 13, no. 4, pp. 18–28, 1998.

[19] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, no. 1, pp. 21–27, 1967.

[20] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Citeseer, 1998, pp. 41–48.