Bag of Embedding Words for Sentiment Analysis of Tweets

Galvez Arias Pierina*, Guzman Ramos Pedro Jesús, Chipana Vila Luis Antonio, Trigoso Valeriano Carlos Alberto, Fabian Arteaga Junior
Universidad ESAN, Perú.

* Corresponding author. Email: 16100545@ue.edu.pe
Manuscript submitted January 9, 2019; accepted March 13, 2019.
doi: 10.17706/jcp.14.3.223-231

Abstract: This paper presents an alternative of solution based in artificial intelligence to simplify the human effort that implies the analysis of the impact for businesses of their publications in social networks services. This analysis is very important because the audience manifest its opinion mostly in texts that must be processed one by one to know their content and use it in benefit of the business, this implies the use of resources to read each comment and extract characteristics that make possible to determine whether the comments are, positive reactions or negative. Our solution can obtain most effective reports than the ones generated by manual procedures, it means that demands less resources and leads to the save of time and money during the extraction of the answers to a Twitter's publication. We use BOEW and Word2vec to generate the characteristic vector for each of the answers. Finally, to make the sentiment analysis we use statistic classification models to polarize comments.

Key words: Artificial intelligence, sentiment analysis, tweets, web scraping, Word2Vec.

1. Introduction

In the era of information in which we live, Internet becomes a repository of billions of data that we generate any time we use a product that is in the web or any time we do something on the web like in web pages, applications, tools and so on. Companies consider Internet as an enormous source of data that when is analyzed correctly, it helps them to know more about their clients and it also helps them to know more about their potential clients and what they are interested in, they can also use this data to improve the customer experience of the products or services they offer, in addition it helps shedding light on which new products could be offered, reducing the uncertainty associated to whether they will be accepted or not [1].

One of the most common trends in recent years, is to be a user of a Social Network Service (SNS) being the most popular Facebook, Twitter, Instagram and LinkedIn. The SNS that we have just mentioned are often erroneously called "social networks", since as Campuzano says, H. a social network refers to a form of interaction between a few people who share common interests or activities, and the current concept of social network refers to the way in which this group develops its relationship activity, that is to say, those that are part of this social network do not have a face-to-face relationship but rather develop it virtually through the Internet, so Online social networks have their origin and are developed through electronic channels and in view of this need, the SRS are born and the service they offer is basically to publish content of interest in many different formats, from multimedia to text only and to share it with other users. [2] In

this way, Social Network Services become a very attractive source to analyze and obtain from them what people think or think about current events, companies, services, products, etc. A type of content that is frequently generated in the environments we have just mentioned are the comments that are expressions of opinions that express the reactions aroused in users by the content they consume in social networks services. Interpreting the trends of the comments is very important for an organization, since it describes the perception that the audience has of a specific company or product and of the image that the organization projects to the public too [3].

Likewise, exploiting and analyzing the content of social networks has caused interest in researchers as well as in the companies themselves and, therefore, analytical techniques have been developed which allow to know in broad strokes the satisfaction and the opinion of the customers respect to a certain topic, this type of study is known as "Opinion Mining" [4].

Opinion Mining is a set of multidisciplinary processing techniques that allow us to extract useful aspects of one or more texts, but how are those aspect useful? To enrich the information we extracted from the previous procedure, we use sentiment analysis to classify different opinions into groups or clusters according to their characteristics, commonly in the simplest techniques, opinions or groups are usually polarized in two, either positive or negative [5], but if the interests are more specific, it can be grouped into more groups depending on how you wish to work on them. Statistical classifiers that apply supervised or unsupervised classification techniques are used to achieve this classification. For our project we work with a database of comments with an assigned classification which makes our classifier supervised. [6]

So, what our solution suggests is that after entering the URL or link to a publication on Twitter (Tweet), the algorithm processes all the comments (response Tweets) that have been made in the original Tweet and indicates us the proportion of positive and negative comments so that the user can interpret them according to the business and speed up the decision making or assure a more objective conclusion about the effectiveness of a specific publication. Despite, our research topic is recent, its applications generate a high added value to companies. In order to publish this paper, all members afforded for it, we didn't receive any support from the institution.

2. State of Art

To make a system of sentiment analysis we must keep in mind that there exist phases and methods that help to complete this task. For this reason, each method must be validated individually and choose the one that gives us a better result in each phase. The phases in a sentiment analysis method are: Collection, Preprocessing, Mining and Analysis [7].

The comments we collect from the social network services are transformed into vectors and stored in a list. The process of converting words to vectors is done because analyzing words for an algorithm is simpler if these words are transformed to numbers, in this case vectors, help to get a better learning performance for the system grouping each word of the comment by its similarity or by the nearest.

Previous work presented by Mikolov Tomas at the conference on neural information processing systems 2013 (NIPS13) shows the skip-gram model, which consists of a larger learning method containing word vectors based on a large amount of data unstructured After using the groups of models we obtain a stock exchange the same ones mentioned by mikolov [8] in their work The Word2vec model and its applications have attracted the attention of a large automated learning community.

These vector representations of the words learned by the model have been shown to have semantic meanings, which are useful for performing different tasks such as the natural processing language. The Word2vec model is needed to use the architecture models to proceed to the distributed representation of words. These models are continuous bag of words (CBOW) or skip-gram. [9] In the CBOW architecture, the

model predicts the current word in a set of surrounding words from the same context. Skip-gram, contrarely, predicts the set of surrounding words using the current word.

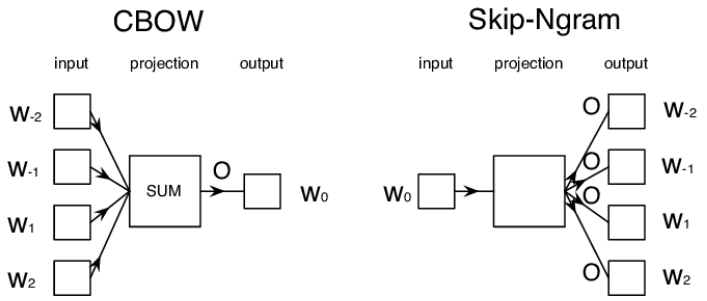

Fig. 1. Illustration of the skip-gram and continuous bag-of-word (CBOW) models.

These 2 architectures [10] are very used however, some authors CBOW is faster than skip-gram but this, works better with infrequent words. In the Word2vec model, it has input values which represent a vector with the values of the word we are dealing with, on the other hand, the output ones that are words related to the input vector depending on the architecture used.

The first step of the model is to build a vocabulary of training text and then through this, learn the vector representation of the words. The resulting vector can be used in many applications of natural language processing and automated learning. A simple way to find out the representations learned is to find the closest word according to the word specified by the user and by applying distance, find the closest value.

This model aims to create a numerical representation of a document, regardless of its length. The concept that Mikolov & Le, have used, is not complicated but is weighing in ways that using the Word2vec model to add an additional vector as an identifier. [8]

A small extension of the CBOW model is used but instead of using only words to predict a next word, a feature vector with unique identifier is added for the document. Word2vec is trained to complete the words surrounding the corpus, but it is also used to estimate the similarity or relationships between the words. It has been proven that with better adjustments, you can obtain a higher performance of the Word2vec model.

In order to determine if the comments are positive or negative, we use statistical classifiers that we train with each of the Word2Vec and Doc2Vec models that we generate. We chose to use SVM, K-means and Neural Networks, the first two classifiers are used because they are the most used for machine learning, as shown by Godoy, A. [11] and we use neural networks because as it is made up of very effective genetic algorithms for the selection of characteristics, they drastically reduce the number of characteristics and thus improve even its performance as a classifier [12].

3. Methodology

As we want to apply sentiment analysis to the comments of customers, we will use text mining techniques to classify these comments into two groups, positive reviews and negative reviews. For this, we have proposed to follow the following stages:

3.1. Stage 1: Database Preparation

For this project, we obtained a very significant number of records of real customer service comments from an important bank. The data was extracted from an Oracle PL / SQL database to a plain text format to facilitate the treatment of it. The data has approximately of 40,000 records.

For this project, we used the fields "Comment", which gives us the opinion about the attention received,

and the field "Evaluation" which gives the respective score of the "Comment" field in a scale from 1 to 5. It is important to highlight that the most recommended data extraction formats for this type of work are txt, csv and xlsx. To polarize the comments, we decided to transform the values of evaluation 1, 2 and 3 into 0 and 4,5 to 1, it is understood that 1 is equivalent to positive and 0 to negative so, we have two classification groups.

We needed to clean the database that's why we standardized it, this is because there are many null records or information that does not help us with the proposed objective. We have used the Python programming language to make use of its libraries in the cleaning of our data. First all the comments were passed in lowercase, then the punctuation marks are eliminated, likewise, eliminate null records and numbers.

3.2. Stage 2: Preprocess

After having extracted the comments, we proceeded to perform a preprocess by cleaning the comments you want to use for the sentiment analysis.

Therefore, the data must be standardized. This part of the treatment was also done in Python. The following techniques were performed:

- *Stops Words*: This technique eliminates the non-significant word of the data.
- *Upper To Lower*: Convert all the letters of the comment in lowercase.
- *Remove signs*: Remove the sign of comments and even convert some signs that have some meaning in a word (Example the sign ":)" represents "happy")

After doing the respective cleaning process, we proceeded to make a transformation to each comment which consists of tokenize (divide it into the minimum unit to study) all the words of the comment. This causes that the comment becomes an arrangement of words.

3.3. Stage 3: Generating Vectors

The multivariate predictive models or classification employ variables of the qualitative type, which was considered the main problem of the comments since it is not of quantitative type, that's the reason why the process of vectorization of the comments was carried out, this means finding a relative equivalent on a numerical scale.

In our research we found various vectorization techniques and we have considered Word2Vec for its efficiency and precision. Once the data was cleaned, we proceeded to perform the vectorization we mentioned before, for that we used Python programming language. We cleaned all the comments using a function developed in Python.

	Size	Min
1	50	2
2	50	5
3	50	10
4	100	2
5	100	5
6	100	10
7	200	2
8	200	5
9	200	10
	Max	

Fig. 2. Size vector and Min_count.

Likewise, we generated another function using Python's gensim library and the module Word2Vec to

generate the characteristic vectors of each word by passing as main parameters of the function the size of the vectors (number of dimensions of the characteristic vectors), the Min_Count (the amount of times a word should be repeated to be considered in the model), etc. Then we must generate the characteristic vector of the whole comment averaging the characteristics vectors of the words in each comment.

We did an exhaustive analysis to vectorize the comments we extracted from social network services and we used different Size_Vector and Min_Count to get the best model. The models we generated are listed in Fig. 2 [13].

Having the comments vectorized, any multivariate classification techniques can be used. Having numerical information helps greatly in explaining the situation. Specifically, we can build conglomerate or cluster analysis models. For this sentiment analysis we rely on three models: K-means, KNN, SVM and Neural Networks. The scikit-learn library was used, which contains the models mentioned above.

The summary of the applied methodology can be seen in the following figure [14].

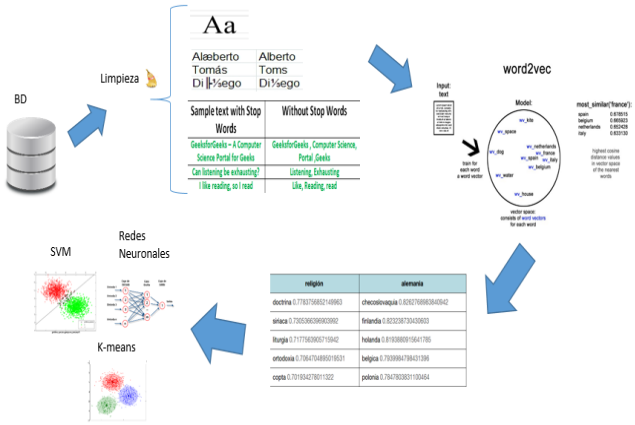

Fig. 3. Methodology.

3.4. Stage 4: Improvement of the Model Using BoEW

Once we have a base of clean comments and we have also generated several Word2Vec models with different values for size, min_count and windows, having already tested their accuracy with statistical classifiers, we choose the one that gives us the best results for our database of comments. If a comment of n words is processed by a word2vec model, it will result in n vectors of size equal to the size parameter that the word2vec model has chosen.

Now, if we pass all our base of comments through a given model word2vec, we get as many vectors as words have the bag of words of the model, all with the same size that is equal size parameter of the model, graphically the word bag of the model is represented as a cloud of points in space.

To group this point cloud, we apply K-means, that is, we define a value k equal to the number of centroids or groups that we want to have to build our BoEW model. The ideal number of k, we will not know until we perform the accuracy tests of each model we generate, but according to the experiments we did, it is advisable to test k values that correspond to the minimum number of words in the comments to be processed, the average number of words They have the comments and the maximum number of words that the comments have.

Finally, for each comment of our database described as n vectors per word, we create a vector A of size k (number of centroids or groups) in which each component of vector A represents a centroid and we initialize it to 0. We proceed to fill it according to the number of words per centroid the comment has as can be seen in the following image[15].

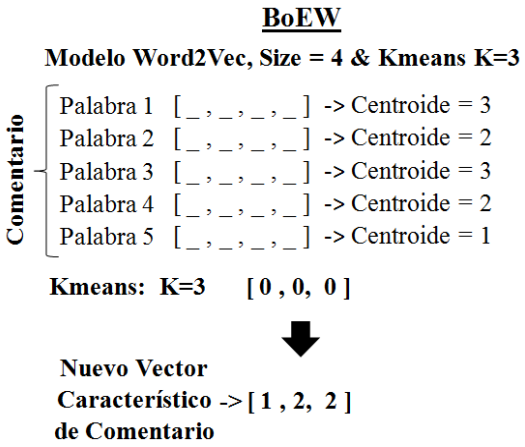

Fig. 4. BoEW model.

The numbers that we assign to the vector initialized in zero is according to how many words in the comment belong to each centroid. Assuming that the first word is closer to the centroid 3, the first component of the vector is added 1, if the next word of the comment is closer to the centroid 2, 1 is added to the component 2 of the characteristic vector and so on. In this way, a vector is constructed that represents the number of words per centroid of a given comment. The sum of all the components of the characteristic vector is the total number of words in the comment.

Finally, the characteristic vectors obtained are standardized and statistical classifiers are used to determine the polarity of the comments, to which category they belong or of what type they are. All the codes that we use to implement the solution that we present in this paper can be found at the following link: *https://github.com/tutorin/analisis_post/*

3.5. Results

Each point mentioned in the methodology of the present project has been registered, giving proof of the use of the Word2Vec model and the different steps mentioned.

For this, the following figure [16] with the respective results is presented:

Experimentos	Size	Min.	Redes Neuronales	SVM	K-means
1	50	2	0.823	**0.820**	0.285
2	50	5	0.821	0.817	0.291
3	50	10	0.820	0.813	0.282
4	100	2	0.813	0.815	**0.712**
5	100	5	0.817	0.813	0.291
6	100	10	0.823	0.811	0.286
7	200	2	0.822	0.813	0.286
8	200	5	**0.825**	0.811	0.288
9	200	10	0.816	0.803	0.286
	Máx.		**0.825**	**0.820**	**0.712**

Fig. 5. Accuracy by Word2Vec experiment.

Here the use of Word2Vec is shown using different Size and Min_count that were proposed for the work, in addition the accuracy is shown with different methods to know the conformity of the measured value with its true value. It should be noted that the best accuracy found is that of 0.825 using a size of 200 and a min_count of 5 and using the statistical classifier of neural networks. Therefore we proceed to test different values of K to obtain the best classification accuracy using BoEW. The results can be seen in the following tables [17]-[20].

K=5					
Size	**Min**	**Redes Neuronales**	**SVM**	**K-means**	**KNN**
200	5	**0.767**	**0.767**	0.378	0.751

Fig. 6. Accuracy for BoEW experiment with K =5.

K=15					
Size	**Min**	**Redes Neuronales**	**SVM**	**K-means**	**KNN**
200	5	0.788	**0.793**	0.362	0.775

Fig. 7. Accuracy for BoEW experiment with K =15.

K=20					
Size	**Min**	**Redes Neuronales**	**SVM**	**K-means**	**KNN**
200	5	0.799	**0.804**	0.363	0.779

Fig. 8. Accuracy for BoEW experiment with K =20.

K=25					
Size	**Min**	**Redes Neuronales**	**SVM**	**K-means**	**KNN**
200	5	0.797	**0.810**	0.641	0.787

Fig. 9. Accuracy for BoEW experiment with K =25.

We can conclude that greater accuracy is obtained when K takes the value of 25 and using the SVM classifier and is 0.81.

4. Conclusions

A system of sentiment analysis is very useful and of valuable application in companies. Since it adds value in the organization by providing concise and accurate information about the opinions of people made on social networks. The proposed model shows us a great precision when distinguishing a negative tone comment from a positive tone comment. It is important to emphasize that we can improve the proposed

model, by adding more comments in the training, and trying with different parameters for each method.

On the other hand, as a group we would like to launch a web platform so that it can be accessed by companies through the creation of an account and a subscription so that they can obtain Insights and analyze the performance of the posts in their social networks services and measure the impact on customers. To reduce the margin of error, Cross Validation techniques and apply deep learning techniques can be used.

References

[1] Larcón, G. (2018). 76% de empresas privadas formales usan internet. *Gestión.* Retrieved from https://gestion.pe/economia/empresas/76-empresas-privadas-formales-internet-225011

[2] Campuzano, H. Las redes sociales digitales: Concepto, clases y problemática jurídica que plantean en los albores del siglo XXI. *Actualidad Civil, 1,* 1-20.

[3] Escortell, A. (2017). *El Impacto de las Emociones en el análisis de la Polaridad en Textos con Lenguaje Figurado en Twitter.* Universitat Politècnica de València. España: Valencia.

[4] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Abril del 2018. *de Foundations and Trends in Information*. Retrieved from http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf

[5] Gutierrez, G., & Troyano, J. (2015). Procesamiento del lenguaje natural aplicado al análisis del sentimiento de opiniones. *Escuela Técnica Superior de Ingeniería Informática*, 7. España: Sevilla, Tesis. Retrieved from http://www.10.13140/RG.2.1.2827.3366

[6] Chaovalit, P., & Zhou, L. (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches. *Proceedings of the Hawaii International Conference on System Sciences (HICSS).*

[7] Chang, J., O'Reilly, C., & Pontika, N. (2018). *What is Text Mining, How Does It Work and Why is It Useful?* Retrieved from https://www.fosteropenscience.eu/content/text-mining-101uence=1

[8] Quoc, L., & Mikolov, T. (2009). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on International Conference on Machine Learning: Vol.* 32.

[9] Meyer, D. (2016). *How Exactly Does Word2vec Work?* Retrieved from http://www.1-4-5.net/~dmm/ml/how_does_word2vec_work.pdf

[10] Ling, W. (2015). *Illustration of the Skip-gram and Continuous Bag-of-Word (CBOW) Models*. Retrieved from https://www.researchgate.net/figure/Illustration-of-the-Skip-gram-and-Continuous-Bag-of-Word-CBOW-models_fig1_281812760

[11] Godoy, A. Técnicas de aprendizaje de máquina utilizadas para la minería de texto. *Universidad Federal de Santa Catarina,* 103-126. México. Investigación.

[12] Estévez, P. (1999). Selección de características para clasificadores neuronales. *Anales del Instituto de Ingenieros de Chile,* 65-74.

[13] Own authorship. *Size Vector and Min_Count. [Fig. 2].*

[14] Own authorship. *Methodology. [Fig. 3].*

[15] Own authorship. *BoEW Model. [Fig. 4].*

[16] Own authorship. *Accuracy by Word2Vec Experiment. [Fig. 5].*

[17] Own authorship. *Accuracy for BoEW Experiment with K =5. [Fig. 6].*

[18] Own authorship. *Accuracy for BoEW Experiment with K =15. [Fig. 7].*

[19] Own authorship. *Accuracy for BoEW Experiment with K =20. [Fig. 8].*

[20] Own authorship. *Accuracy for BoEW Experiment with K =25. [Fig. 9].*

Galvez A. Pierina was born in Lima, Peru in March, 19th 1994. She is currently studying at Esan University IT and Systems Engineering career and she is about to graduate.

She works as Intern in the area of marketing and customer experience in Banco de Credito del Perú (BCP). BCP is located in Lima, Peru. She has worked as an intern in the commercial division in the same place before. She has strong interests in voice of customer's analysis.

Guzmán R. Pedro Jesus is a student of the ninth cycle of information technology and systems engineering at the ESAN University - Peru. Interested in artificial intelligence issues and its applications in mobile devices.

Chipana, Luis was born in Lima, Peru on December 20, 1994. He is a systems engineering student at the university ESAN in Peru. He studied the primary school at the Maria Montessori School and the secondary school at the Shuji-kitamura School. He did internship at ESAN University in the area of marketing as a database analyst. Participant in "PERU WITH SCIENCE" organized by concytec 2018 presenting the project "HORUS".

He is a speaker of a robotic project "HUMIDITY DETECTOR" in the "IV SCIENTIFIC MEETING" of the naval school of Peru and the air force of Peru representing the ESAN University.

Trigoso V. Carlos was born in Lima, Perú, in 1994. He is an undergraduate student in ESAN's University and will be graduating in 2019 with a BS in IT & Systems engineering.

He is currently working at Telefónica del Perú as a intern in Lima, Perú. He has a strong interest in the field of database administration.