DEEP LEARNING: APPLICATION OF THE LSTM MODEL IN THE CATEGORIZATION OF TWEETS ON VIOLENCE IN THE CITY OF BELÉM

Jonatã Paulino da Costa, Aurea Milene Teixeira Barbosa dos Santos, Eulália Carvalho da Mata,
Marcelino Silva da Silva, Priscila Siqueira Aranha and Carlos Renato Lisboa Francês

UFPA – Universidade Federal do Pará

Rua Augusto Corrêa, 0 – Guamá – CEP 66075-110 – Belém-Pará-Brazil

ABSTRACT

Twitter has more than 313 million users, is considered a subjective social network of content generation on the Internet, not only used to disclose personal information, but shares opinions and information about events and events in general, being a great source of research for the discovery of knowledge among relevant data. In this context, the present work presents a study on the implementation (application) of the Recurrent Neural Network LSTM for the analysis and classification of tweets that are related to crime and not crime. The set of data that is the object of the survey comes from an account of a newspaper in the city of Belém, containing 20000 records in a period of two months. The results obtained with the use of RNR LSTM proved to be quite satisfactory, reaching accuracy in some cases of 90% correctness.

Based on the results of this research, we observed the effectiveness of this method in the analysis of feelings in comparison with other algorithms such as Nayve Bayes, Convolutional Neural Networks used in the literature. Because it is a network that uses a Long-Term Memory technique, it can adapt to large time intervals, such as Twitter texts.

KEYWORDS

Deep Learning, Twitter (Tweets), Python, Embeddings, Categorization, Analysis of Feelings

1. INTRODUCTION

The internet has evidently revolutionized the world and people as a whole. With the advancement of technology, people are using their computers or mobile devices more often to access the web and establish their social networks. According to a survey carried out by the Regional Center for the Development of the Information Society (CETIC, 2014), in 2014, 94.2 million Brazilians had Internet access, 76% of whom used it to participate in social networks, indicating that virtual communication is increasingly common and present in the lives of Brazilians. Comments are openly shared as suggestions and feedback on particular things, product, merchandise, political affairs and crimes, such as murder, femicide and other types of crimes (Kaur, 2014). Most of these comments, being publicly shared, can be easily accessed from the web (Schuller, 2014).

Social Networks are networks formed from interactions between people, groups or institutions motivated by common interests or goals that relate through digital media. In Social Networks there are opinions, personal information or events in general. Among the tools that allow the creation of social networks is Twitter that has more than 313 million monthly active users (TWITTER, 2017).

Twitter emerged in 2006, a social network with an unusual premise - each post, called tweet, should not exceed the limit of 140 characters. Due to its growing popularity, the network has gradually become more efficient and accessible, becoming of paramount importance, giving its users the power to create and share ideas and information instantly, without barriers. Twitter is a subjective network of content generation on the Internet, according to the region and the culture of the population inserted in the context. Another factor that influences is related to the use given to this network by its users. They not only use it to disclose personal information as originally intended, but also to share opinions and information on facts and events in general (Naaman, 2010).

This usage feature makes Twitter a people-based news feed (Zhao, 2009), making it a very important tool for two types of user profiles: those who seek information and those who want to spread their information widely (Java, 2007). The combination of these factors makes Twitter an important source of opinions and feelings about events and events that can be analyzed and subsequently used in several areas such as politics, social and even business marketing, as verified by (Li, 2011).

With the emergence of Computational Intelligence techniques, in particular Deep Learning, it is now possible to analyze several texts, including social networks. These techniques give us quite satisfactory results according to the effectiveness of the implemented model.

In this research, the Twitter social network was used where data from @doldiarionline (referring to a local newspaper in the city of Belém) were collected to train a classifier and to be able to categorize if a tweet contains a crime report or not, in a binary way. To classify the tweets, a Recurrent Neural Network (RNN) with LSTM extension (Long-Short Term Memory) was applied. The selection of this network was based on the observation of the works in the literature as (Iucca, 2013), which employs this same technique of analysis and demonstrates satisfactory results for classification. This network is also used in applications involving handwriting recognition (Doetsch, 2014), handwriting generation (Graves, 2013), language translation (Luong, 2015), acoustic modeling of speech (Sak, 2014), synthesis of speech (Fan, 2014), audio analysis (Marchi, 2014), video analysis (Zhu, 2017).

The importance of this classification is due to the discovery of knowledge among the relevant data of the social network (Twitter), providing an understanding of the cause and effect relationships in relation to the various issues that refer to crimes such as homicide, femicides, homophobia, and many other types. Such phenomenon is influenced by parameters of the real world society. You can also search for social networking information by distributing projects.

The article is organized in the following sections: Section 2 discusses Recurrent Neural Networks LSTM; Section 3 presents the methodology; section 4 discusses the experiment and the results and in section 5 the final considerations.

2. RECURRENT NEURAL NETWORKS LSTM

These networks use a specific mechanism in their hidden layers, called memory cells. This allows the RNN LSTM to be able to 'remember' the information it has stored even after several iterations occur. It can understand not only the words in the text, but also the order in which they are found (Ghazi, 2019).

There is a dependence of the sequence with which the data occur in time dimension, and the variation occurs in time, where the sense of time in this context means position, sequence, that is, $t + 1$, $t + 2$. An example of its operation is that when the phrase "The house" is read, the letter A follows to the matrix the front where is the word house, that is, the letter A comes from the previous matrix and combines with the matrix the front where is the word house and so on. This way when working with longer sentences your memory is not lost, and when the network reaches the end of the reading there will be an understanding of the complete sentence, just like our brain, where we take into account what happened before (DEEP LEARNING BOOK, 2019).

There are several LSTM drive architectures. A common architecture is composed of a cell (the memory part of the LSTM unit) and three "regulators", commonly called ports, of the information flow inside the LSTM unit: an input port, an output port, and a forgetfulness port. Some variations of the LSTM drive do not have one or more of these gates or perhaps have other gates. For example, blocking recurring units (GRUs) do not have an exit port (Sercu, 2017).

There are connections inside and outside the LSTM ports, some of which are recurrent. The weights of these connections, which need to be learned during training, determine how the doors operate. In Figure 1, we can see an LSTM unit with its internal processes.

Figure 1. Unit LSTM

The processes are detailed step by step in the items below for a more detailed understanding.

- Cell State:** Partially solves the vanishing problem. In the cell state there is no function, the gradient can flow through the cell state besides passing through the sums and multiplications that ponder what will enter the cell. There is permission to fix all times, partially solving the problem. If we read very long times we can have problems, but there are some adaptations that can read a thousand, five thousand ten thousand times.
- Forget Gate LSTM:** Gates are the multiplicative units. The Forget Gate is responsible for forgetting internal state information that comes from the previous state that will be discarded. This gate is a sigmoid that varies between 0 and 1. At value 1 the intact internal state is maintained. At value 0 the internal state is discarded, it is totally cleaned based on the previous state and also on the input. All operations work with the same idea, that is, based on what is known based on what is being received now, if it determines what to do, the internal state discarded, 1 internal state intact.
- Input Gate:** It is who decides what is going to be inserted as information in the cellular state, which will enter again. Besides forgetting something, there needs to be an update. In this case there is another sigmoid operation, 1 enters all information, 0 does not enter any information. At the moment a $\hat{C}$ is created that is a candidate to a new cellular state and based on that there is an update soon after.
- Output Gate:** It is who determines what new information will be passed to the next time and to exit the Network in its internal state, and if there is a new time, is what will be sent to this new time, also based on the input and the previous state.

Figure 2 illustrates the processes that were respectively cited above.

Figure 2. LSTM processes

3. METHODOLOGY

The approach proposed in the preparation of this work follows the following steps: (i) data collection in the social network Twitter, (ii) pre-processing of the data, in order to remove noises and irrelevant terms; (iii) extraction of characteristics of texts for entry into the Network; (iv) Experiments and results through the training of the Neural Network to classify existing categories on the basis of crime or non-crime.

Figure 3. Methodology Flowchart

3.1 Data Collect

The data collection stage for this work was done on twitter, in a period of 2 months on interspersed days, with 3,200 tweets being extracted each day. 20,000 tweets were collected with publications that related to crimes that happened in the city of Belém and others that had no relation with crimes.

3.2 Preprocessing

Preprocessing is the step performed immediately after data collection. This step consisted of separating the tweets into tokens, soon afterwards they removed all the mentions like URLs and emoticons of each of them. Subsequently, the tokens are normalized, that is, they undergo transformations. After normalization, stopwords were removed and soon after the elimination of stemming.

Of a total of 20,000 tweets collected, 1,500 tweets were separated to be labeled in crime and 1,500 tweets to be labeled in non-crime, serving as a basis for learning the Network. This separation and labeling of data in crime, not crime, is given as a supervised model, and in this type of model the Neural Network makes decisions based on the experiences contained in predefined examples (labels) that serve as a sample for its learning. This information has been used explicitly so that the Neural Network can be able to show expected results at the exit of the Network. This labeled base served for the training and learning of the Neural Network. Table 1 shows an example of the database labeled.

Table 1. Table Example of Tweets that Have Been Labeled in crime and naocrime for Network Training

Tweets	Labels
A woman is found dead inside the cemetery	crime
Thieves attack ALEPA to steal copper cables	crime
Remains of four people are found in the interior	crime
Man arrested on charges of raping two children	crime
Fernanda Gentil is tight with language in Russia: It's easy	non-crime
20 years after the defeat, 98's selection finally has revenge	non-crime
Dol wants to know: do you think spanking educates you?	non-crime
Ex-BBB Marcos Harter will be a candidate for federal deputy	non-crime

3.3 Extraction of Characteristics

After the preprocessing step and extraction of the raw data characteristics, an LSTM Recurrent Neural Network was used, where these networks according to the state of the art are widely used for text classification (Poria, 2016), given its effectiveness when it comes to solving problems related to time series. Figure 7 shows the network architecture.

The model for conducting the experiments consists of an architecture that contains four layers: the first is the input layer that receives the documents in time intervals, the second layer is the embedding layer that brings an array with the pre-trained vector representation of the word vocabulary. The third layer is an LSTM layer, where these vector representations enter, and the cell output of the LSTM Network is then fed into a densely connected layer and finally a softmax activation function with two categorical outputs. The number of outputs (units, nodes or neurons) is the dimension of the representational space of the layer, that is, the amount of freedom that the network can learn by an internal representation.

Figure 4. Network Architecture

4. EXPERIMENTS AND RESULTS

The experiments will be presented and the results obtained after the evaluation of the proposed method.

In this step the model was trained and evaluated using a set of hyper parameters with the objective of obtaining better results in the classification of the categories of tweets, crime or non-crime. Table 2 describes the main hyper parameters used in the Neural Network.

Table 2. Configuration of Hyper Parameters

Hyper parameters	description	value
word embeddings	Layer size word embeddings	50
batch_size	Number of samples to be used for each gradient update	30
max_fatures	Reflects the maximum number of words in the dataset maintained in the vocabulary	5000
max_sequence_length	It limits the maximum of the sequences that will be used in the training, the size of the sequences in the dataset are varied, so it is necessary to limit this size and will be used from the beginning to the end of the training	300

4.1 Neural Network Performance

To evaluate the performance of the model, the tweeter tests are not included in the training data set, where the model is presented as a performance indicator after the insertion of the new tweeters as shown in table 3, where we see the Input, which are the entries of new tweets, the description, where the model indicates the due class referred to the Input and finally the percentage of hit that is reached by the model.

A binary_crossentropy loss function, an adam optimizer, and the f1-score metric were used in this network. The metric f1-score is one of the most suitable metrics for unbalanced databases, which is the case of the base used in this work. The data were divided into 20% for the tests, 80% for training. Sixty iterations were performed with the model.

Observing the graphs of figure 5 and figure 6, where a comparison between the training data and the validation data is shown, we can see that the model presented efficient results based on 90% peak accuracy at the end of 60 seasons. When the model is submitted to new tweeters.

Table 3. Table Demonstration Percentage of Correctness of new Description Tweet

Input	Description	Probability
Woman is killed with two head shots	crime	88%
Roberto Carlos show in Belem will be held on Saturday	non-crime	88%
Seven people were murdered at dawn in Belem	crime	89%
Some church members held a fraternization party	non-crime	89%

Figure 5. Graphic f1-score of the Model

Figure 6. Graphic Crossentropy of the Model

5. CONCLUSION

In the present research was developed and analyzed the application of a Recurrent Neural Network LSTM for the categorization of tweeters in crime or non-crime in the city of Belém. For the development of the same it was necessary to make collections of twitters that had relations with crimes and those that did not had no relation to crimes to have a substantial sampling of the problem domain, thus generating the data set for training and testing, this stage was one of the most laborious because of the need for manual labeling of all used tweets, considering that for the context of the research there is no reference indicating a pre-existing basis.

The results obtained were quite satisfactory, presenting a great advantage in detecting these feelings by means of the classification method, with 90% accuracy, demonstrating that the Recurrent Neural Network LSTM is an efficient method in the proposal of this work. Since the purpose of the research was to demonstrate the potential of the use of the recurrent neural networks LSTM in the context of the classification of tweets related to crime.

In future works, comparisons and evaluations will be carried out with other learning methods used in the literature of feelings analysis in social networks, such as Naïve Bayes and Convolutional Neural Networks to show which would be the best method to categorize tweets related to crimes, as well as change of metrics by analyzing their performance.

Another contribution of this research will be to use the best method of categorizing social networks to create a database of crimes in cities that do not have an organized public database with data on crime, a situation that is recurrent in several Brazilian cities. The creation of this base will enable the obtaining of important information in the understanding of the society, besides assisting in studies and demands to combat the decrease of crime.

ACKNOWLEDGEMENT

A UFPA (Universidade Federal do Pará), PPGEE, CAPES, CNPq.

REFERENCES

- CETIC (2014). Apresentac ao dos principais resultados tic domicílios: Acesso em 10 fev 2019.
- Data Science Academy (2019). Long short-term memory (lstm). Acesso em: 11 fev. 2019.
- Doetsch, P., Kozielski, M., and Ney, H. (2014). Fast and robust training of recurrent neural networks for offline handwriting recognition. In 2014 14th International Conference on Frontiers in Handwriting Recognition, pages 279–284. IEEE.
- Fan, Y., Qian, Y., Xie, F.-L., and Soong, F. K. (2014). Tts synthesis with bidirectional lstm based recurrent neural networks. In Fifteenth Annual Conference of the International Speech Communication Association.
- Ghazi, M. M., Nielsen, M., Pai, A., Cardoso, M. J., Modat, M., Ourselin, S., Sørensen, L., Initiative, A. D. N., et al. (2019). Training recurrent neural networks robust to incomplete data: Application to alzheimer’s disease progression modeling. *Medical image analysis*, 53:39–46.
- Graves, A. (2013). Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850.
- Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pages 56–65. ACM.
- Kaur, A. and Gupta, V. (2014). N-gram based approach for opinion mining of punjabi text. In International Workshop on Multi-disciplinary Trends in Artificial Intelligence, pages 81–88. Springer.
- Li, Y.M. and Li, T.-Y. (2011). Deriving marketing intelligence over microblogs. In Proceedings of the 2011 44th Hawaii International Conference on System Sciences, HICSS ’11, pages 1–10, Washington, DC, USA. IEEE Computer Society.
- Lucca, G., Pereira, I. A., Prisco, A., and Borges, E. N. (2013). Uma implementac ao do algoritmo naive bayes para classificac ao de texto. Centro de Ciencias Computacionais-Universidade Federal do Rio Grande (FURG).
- Luong, M.-T. and Manning, C. D. (2015). Stanford neural machine translation systems for spoken language domains. In Proceedings of the International Workshop on Spoken Language Translation, pages 76–79.
- Marchi, E., Ferroni, G., Eyben, F., Gabrielli, L., Squartini, S., and Schuller, B. (2014). Multi-resolution linear prediction based features for audio onset detection with bidirectional lstm neural networks. In 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 2164–2168. IEEE.
- Naaman, M., Boase, J., and Lai, C.-H. (2010). Is it really about me?: Message content in social awareness streams. In Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW ’10, pages 189–192, New York, NY, USA. ACM.
- Poria, S., Cambria, E., and Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108:42–49.
- Sak, H., Senior, A., and Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth annual conference of the international speech communication association.
- Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., and Zhang, Y. (2014). I (special session)***** the interspeech 2014 computational paralinguistics challenge: Cognitive & physical load. In Fifteenth Annual Conference of the International Speech Communication Association.
- Sercu, T., Saon, G., Cui, J., Cui, X., Ramabhadran, B., Kingsbury, B., and Sethy, A. (2017). Network architectures for multilingual speech representation learning. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5295–5299. IEEE.
- Twitter (2017). Twitter milestones: a selection of memorable moments.
- Zhao, D. and Rosson, M. B. (2009). How and why people twitter: The role that micro-blogging plays in informal communication at work. In Proceedings of the ACM 2009 International Conference on Supporting Group Work, GROUP’09, pages 243–252, New York, NY, USA. ACM.
- Zhu, X., Wang, Y., Dai, J., Yuan, L., and Wei, Y. (2017). Flow-guided feature aggregation for video object detection. In The IEEE International Conference on Computer Vision (ICCV).