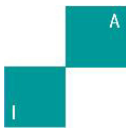
Inteligencia Artificial, 23(66) (2020), 66-84
doi: 10.4114/intartif.vol23iss66pp66-84



INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

Sentiment Analysis Applied to Analyze Society's Emotion in Two Different Context of Social Media Data

Marilyn Minicucci Ibañez[1,2,A], Reinaldo Roberto Rosa[3,1], Lamartine N. F. Guimarães[4,1,5]

[1]Applied Computing Graduate Program (CAP), National Institute for Space Research, São José dos Campos, São Paulo 12227-010, Brazil
[2] Federal Institute of São Paulo - IFSP-SJC, São José dos Campos, São Paulo 12223-201, Brazil
[A] marilynminicucciibanez@gmail.com
[3]Lab for Computing and Applied Mathematics (LABAC), National Institute for Space Research, São José dos Campos, São Paulo 12227-010, Brazil
[4] Nuclear Energy Division (ENU), Institute for Advanced Studies (IEAv), São José dos Campos, São Paulo 12228-001, Brazil
[5]Space and Technology Science Graduate Program (PG-CTE), Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, São Paulo 12228-900, Brazil

Abstract
In the last few decades, the growth in the use of the Internet has generated a substantial increase in the circulation of information on social media. Due to the high interest of several areas of society in the analysis of these data, a study of better techniques for the manipulation and understanding of this type of data is of great importance so that this enormous volume of information can be interpreted quickly and accurately. Based on this context, this study shows two approaches of sentiment analysis to verify the emotion of the population in different context. The first approach analyses the positive and negative sentiment about 2018 presidential elections in Brazil considering data from the Twitter social network. The second approach performs analysis of data from social media to identify threats sentiment level of armed conflicts considering data off the conflict between Syria and the USA in 2017. To achieve this goal, machine learning techniques such as auto-encoder and deep learning will be considered in conjunction with NLP text analysis techniques. The results obtained show the effectiveness of the approaches used in the classification of sentiment within the domains used according to the methodology developed for this work.

Keywords: Machine Learning, Deep Learning, Auto-encoder, Natural Language Processing, Sentiment Analysis, Social Media

1 Introduction

The evolution of the Internet has enabled the advent of social media as one of the main means of circulation of personal, political and dissemination information. As the consequence, the amount of information generated daily in these means of communication gradually increases annually. This excessive amount of

ISSN: 1988-3064 (on-line)

information has drawn the attention from different areas of knowledge who have realized the importance of using this information in a suitable way to verify the opinions and sentiment of the population in a certain subject. In this way, the need to use modern techniques, as machine learning and sentiment analysis, is identified to assist in the precise verification of specific information among these huge volumes of data [11], [22], [30].

In accordance with this need, this paper presents two approach of applying sentiment analysis to verify the opinion of the population on two main themes: the presidential elections in Brazil in 2018 considering data from social networks and an analysis of social media data for the identification of threats of armed conflicts considering data on social media about the conflict between Syria and the USA in 2017. Emphasizing that sentiment analysis is a machine learning area that allows the identification of human emotions in texts, images, sounds, etc. considering a given domain [6].

For the analysis of the election for president of Brazil, data were collected in Portuguese from the social network Twitter candidates Bolsonaro and Haddad. For the analysis of tweets, a vocabulary was built with words related to the choice of the topic. Each vocabulary word was polarized at 0 (negative word) and 1 (positive word) considering the theme of the election as a domain of analysis. In the analysis of the armed conflict, from the news collected, the one that best represented the idea of threats among government officials in the countries of the analyzed conflict was selected and which served as a basis for calculating the similarity with the others analyzed news. Syrian conflict data was collected on social networks around the world and in English, such as Reuters [28], CNN [9], The Guardian [17], etc.

The analysis performed on the data collected on the social network Twitter generated as results the identification of the percentage of positive and negative messages (tweets) for each candidate analyzed. This analysis allowed to identify which candidate was more favorable to win the elections. The analysis of information from social media, on the other hand, allowed to identify the percentage or level of threat among those involved in the conflict. This result can be used to identify how these threats favored the start of the armed conflict.

To achieve the results in the work, applied sentiment analysis together with Natural Language Processing (NLP) and machine learning techniques. For the application of NLP, the word embedding APIs NLTK [5], Tensorflow [16] and SpaCy [31] were used. For the application of machine learning techniques such as Auto-encoder, Deep Auto-encoder and Deep Learning, was used the Deep Learning Keras API [8].

The objective of this work is to show two sentiment analysis approaches using two different public information sources, as social networks and social media. These approaches identify the positive/negative sentiment and threat degree emotions expressed in that information sources by the considered population.

2 Related Works

This section presents the state of the art of some of the main work related to the emotion analysis in social network and social media. There is no intention in this work to make a detailed bibliographic review of the published articles.

In 2015, the article A Multilingual Approach for Sentiment Analysis, [27], presented a comparison between different sentiment analysis tools using nine different languages: Portuguese, French, Spanish, Italian, Turkish, Russian, Arabic, Dutch, German and English. The database was initially in English and than was translated into the other languages using the Python Goslate API for translation. The sentiment analysis tools used in the comparison were: Linguistic Inquiry and Word Count (LIWC), SentiStrength, SentiWordNet, SenticNet, SASA - SailAil Sentiment Analyze, Happiness Index, PANAS-t - Positive Affect Negative Affect Scale, NRC Emotion Lexicon, NRC Hashtag Sentiment Lexicon, Sentiment140 Lexicon, OpinionLexicon, VADER - Valence Aware Dictionary for sEntiment Reasoning and Emoticons. In the analysis it was found that the best accuracy of the tools was for the English language.

In 2016, [14] published a work entitled A novel deep learning architecture for sentiment classification that proposed a hybrid deep learning architecture with a two-layer Boltzmann Restricted Machine (RBM) and a Probabilistic Neural Network (PNN) for the classification of sentiments.
In the first stage, RBM performs the dimensionality reduction. In the next step, the PNN performs the classification of sentiments. The work tested five different sets of data and compared the results with

current works. The proposed method showed the best precision of 93.3 %, 92.7 %, 93.1 %, 94.9 % and 93.2 % for the Movies, Books, DVD, Electronics and Utensils approaches kitchen, respectively.

In 2017, [29] published the article Stacked Denoising Auto-encoders for Sentiment Analysis: A review in which different types, topologies and learning methods that use Auto-encoders in the analysis of sentiment in texts multi-domain and multilingual. The main subjects approach in the article are presented as follows:

- overview of the state of the art of Sentiment Analysis, highlighting the application Auto-encoders (AEs), Denoising Auto-encoder (DAs) and Stacked Denoising Auto-encoders (SDAs).
- use of marginalized SDAs (mSDAs) and their variants, such as: Heterogeneous Hybrid Transfer Learning (HHTL) and Stacked Instance Denoising Auto-encoders.
- comparison between the models presented.

The author concludes that SDAs can be further deepened and that cross-cultural analysis is an area to be further explored in sentiment analysis.

In 2018, [1] show a study that analyzes the different approaches that Eastern and Arab media apply to crisis events. In the work, the messages published on Twitter about the November 2015 terrorist attack in Beirut and Paris were used as a case study. In the analysis, 2390 tweets were used with the sentiment of sympathy classified for the training of a regression model of a deep convolutional neural network. This model was also used to predict the sentiment of sympathy for the upcoming crisis events. Three forms of analysis were used: bias coverage, which verifies if there was any difference in the coverage or volume of news in both countries; the News media sympathy bias that analyzes how the degree of sympathy of the messages could affect the population and information propagation that verified if the sentiment of positive sympathy spread in the messages. As a result, it was found that both countries had similar coverage, 79% accuracy in predicting the sentiment of sympathy was also achieved, and it was observed that the retweets were impartial in terms of whether a tweet was sympathetic or not.

In 2019, [18] show a study which present on how the avoidance of Hong Kong residents' actions in relation to tourism and how these actions are affected by socioeconomic factors. This analysis was carried out using 72,755 newspaper news from 2003 to 2015. In order to carry out the analysis, a sentiment analysis framework was developed in the work in news published in the media in Chinese language. This framework uses SVM and NB techniques to classify news. The prediction results of this analysis are: for the SVM (Support Vector Machine) classifier (accuracy = 0.913; F-measure = 0.914); for the classifier NB (Naive Bayes) (accuracy = 0.839, F-measure = 0.840).

In 2020, [13] presents a study on how the negative influence of the news published in the news media influences the perception of society in relation to the understanding of the use of Artificial Intelligence in society, specifically in the area of health. The work uses quantitative and qualitative approaches to analyze news published on AI in the period from 1956 to 2018. For this purpose, the Google Sentiment analysis tool, the Cloud Natural Language API, was used. As a result, the work concluded that the data analyzed failed to support the theory that the negative sentiment of the media about AI had any influence on its use.

3 Theoretical Foundation

This section shows briefly the theory used in developing this work. The section 3.1 explains how Natural Language Processing (NLP) applies to this work. The section 3.2 presents the sentiment analysis technique related to this work. The section 3.3 shows some important definition of machine learning, related to this work, with the techniques: MultiLayer Perceptron, Auto-encoder and Deep Auto-encoder.

3.1 Natural Language Processing - NLP

Natural Language Processing or NLP is a sub-area of artificial intelligence and enables the development of systems that allow computer-human iteration using human natural language whether by text or by speech. ems that allow computer-human iteration using human natural language whether. The NLP can be divided into the following stages [21] which are detailed below.

- Tokenization is very well characterized in artificial languages as programming languages, since they do not have much ambiguity. However, in natural language the same character may have several meanings depending on its context. In this way, the tokenization was used considering the approach **Languages Delimited by Spaces** - like the European languages, word boundaries are only indicated by the insertion of blank spaces. However, the symbols between the spaces are not necessarily the tokens needed for the next processing. This is due to the ambiguous nature of writing systems as to the range of tokenization conventions that exist.

- Lexical Analysis performs text analysis at the level of the word. A basic task of logical analysis is to relate morphological variants to their respective keywords. These words are found in a keyword dictionary grouped together with their semantic and syntactic information. In the case of the analysis of natural language, logical analysis dismantles the words of a sentence in its grammatical components (noun, adjective, pronoun, etc.).

- Syntactic Analysis is the task of recognizing a sentence and assigning it a syntactic structure. These syntactic structures are attributed by Context Free Grammar (CFG). The application of CFG by means of specific algorithms generates a representation in a tree structure. These trees analyze an important intermediate state of representation for semantic analysis.

- Semantic Analysis is the representation of intermediate meanings. They are only composed of linguistic expressions. These representations of meanings are only attributed to phrases based on the knowledge acquired with the logical and grammatical phases. Thus, this type of analysis is used in the understanding of the meaning of a sentence. It is also widely used for elimination of ambiguities.

- Pragmatic Analysis is the last stage of the analysis of natural language. In this phase, the meaning is elaborated on the basis of contextual knowledge and logical form and is mapped to the final language of representation of the subject matter. This type of analysis is also used to validate the semantic analysis. In this analysis it is considered that words can be associated by meanings (water, swimming) or subject proximity (water, well).

In this work, the tokenization stage was used to eliminate unwanted characters from social media news and Twitter messages (tweets), along with the machine learning techniques present in the NLTK [5] and Tensorflow [16] APIs. NLTK is an open source platform for creating Python pro-grams to work with human language data. The platform has an easy to use interface with the word processing library defined for classification, tokenization, derivation, markup, semantic reasoning and analysis, and libraries related to Natural Language Processing (NLP) [5].

3.2 Sentiment Analysis

The Sentiment Analysis area refers to the tasks of analysis, identification and classification of all information that is characterized in an emotional way, a subjective way, or an opinion generating information, be it information in text, image or sound format [10]. For the accomplishment of these characterization tasks, it is usually used the Natural Language Processes statistics and/or machine learning methods. According to [10], these tasks are usually divided as follow.

- Subjectivity Classification: deals with the identification of parts of the texts that have a sense of subjectivity.

- Polarity Classification: determines that fragments of texts are classified as positive or negative sentiments.

- Intensity Classification: works with the emotional intensity expressed in the text. This type of approach is usually divided into classes: strongly positive, positive, neutral, negative or strongly negative sentiments.

- Sentimental Analysis Based on Topics or Features: is the verification of existing features related to sentiments about the subject matter.

- Opinion Mining: is related to the retrieval of information from a query. Thus, it allows to consult a specific topic and to classify it in a certain category.

In recent years, Sentiment Analysis has been applied in different ways of expressing sentiments in alternative textual forms, mainly through information from social networks. In the current literature of the area, there are works that identify sentiments through sounds, using emoticons from social networks [4], and images [33]. In addition, Sentiment Analysis is also being used in the most diverse sectors of society. In the last few years, the network and the social media sector has turned its attention to the study of the sentiments of the population and the government. In many countries, the sentiments of the population are being studied in relation to the policies and programs launched by the different areas of their governments (municipal (county), state and federal) [3] [32].

In the work of [7] it is emphasized that a way of controlling intelligence activities is the knowledge of the public opinion. In Brazil, research is already being carried out in the area involving the view of society in relation to the public programs and government policies. One of these surveys is the work [2] that analyzes the hashtags of social networks to assess the population's adherence, by state, to the campaign of the Aedes Aegypti mosquitoes that transmits the dengue virus among other diseases.

This work uses the sentiment analysis Polarity Classification to analyze election and sentiment analysis based on topics or features to analyze armed conflict data.

3.3 Machine Learning

The Machine Learning deals with the computational algorithms that allow learning and, consequently, its improvement through the repetition of experiments. Within the Machine Learning there are several applications ranging from Data Mining that allows the discovery of general rules in a large volume of data, to systems that automatically learn the needs of one user [19].

3.3.1 MLP - MultiLayer Perceptron

In the Multilayer Perceptron network each unit performs a weighted sum of their inputs and transmit that level of activation through a transfer function to produce an output. The network therefore has a simple interpretation as an input-output model, with weights and biases as free model parameters. Such networks can model functions of arbitrary complexity with the number of layers and the number of units in each layer determining the complexity of the function [24] [19]. The learning of the MLP network is performed using the backpropagation algorithm. This algorithm is based on two basic steps.

- Propagation, in this step an input pattern is presented and its result is propagated, layer by layer. The synaptic weights are fixed and at the end a set of network output is released.

- Backpropagation, in this step the network output is compared to the output that is desired to calculate the error correction parameter. The weights are adjusted according to the result of the error correction parameter calculation. This adjustment is applied, layer by layer - from the output layer to the input layer [19].

The figure 1 shows an example of a computational model for the MLP network.

The MLP neural network was applied in this research to classify the positives and negatives sentiments in tweets about running candidates to the 2018 Brazilian presidential election.

3.3.2 Deep Learning

Deep Learning consists of a feed forward neural network that has a large depth related to the number of layers between the input of the network and its output. Feed forward neural networks aim to map a function $y = f(x, \theta)$ and learn the value of the parameter that represents the best fitting of the function [15]. In the Deep Learning model the first layers are used in an unsupervised way, and later values in the next layers are used as initial values for supervised learning. Several knowledge areas that are using Deep Learning, Natural Language Processing (NLP) are making a significant increase in its use. The use of Deep Learning in the NLP area made a great optimization of information processing possible, as can

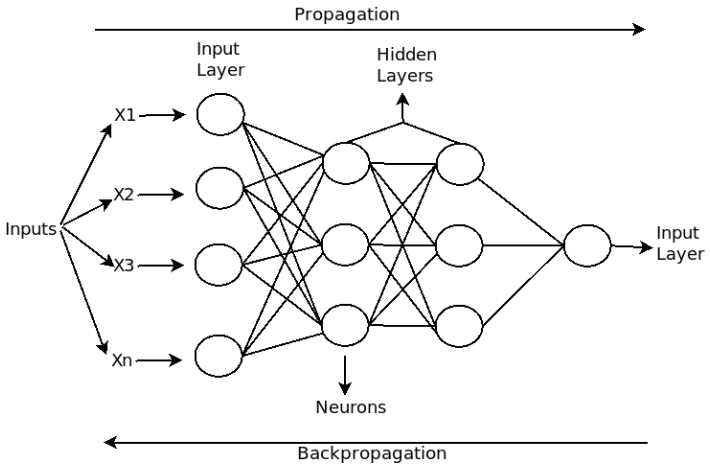


Figure 1: MultiLayer Perceptron Representation. Adapted of [19].

be seen in [6].

In this work, the Deep Learning is used by SpaCY API to calculate similarity levels in news about armed conflict.

3.3.3 Auto-encoder

Auto-encoder is a learning algorithm that uses a neural network to represent the dimensionality reduction of its input in its output. According to [15], this algorithm internally has a hidden layer z that describes the code used to represent data entry. Self-encryption consists of two parts.

- The Encoder function is defined by $z = f(x)$ and compresses the input into a latent space representation.
- The Decoder function is defined by $r = g(h)$ and which reconstructs the output of this representation.

The Figure 2 shows an example of the neural network structure Auto-encoder.

The auto-encoder is typically used for dimensional reduction or feature learning [20]. For training one can use both the Feed-forward neural networks and the Backpropagation [19]. According to [15] there are various types of Auto-encoder.

- Sparse Auto-encoder, it has a training criterion involving a scattered penalty (h) on the code z layer. In which we have, $L(x, g(f(x))) + h$, where $g(h)$ is the decoder output and $h = f(x)$ is the encoder output. Sparse Auto-encoder is typically used for sorting tasks [23].
- Denoising Auto-encoder, it receives corrupted data as input and is trained to reproduce the original data and not corrupted as a result. This Auto-encoder minimizes the equation $L(x, g(f(x)))$, where x represents a copy of data that has been corrupted by some form of noise. Denoising Auto-encoder is commonly used for the retrieval of information with noise [23].
- Convolutional Auto-encoder, it differs from conventional Auto-encoders, since their weights are shared between places at the entrance, preserving the spatial location. The reconstruction is therefore due to a linear combination of small parts of the data based on the initial code. Convolutional Auto-encoder is typically used for image pattern recognition [23].
- Deep Auto-encoder, it is an algorithm consisting of an input layer, multiple layers of encoding and multiple layers of decoding. It can be pre-trained as a stack of single-layer Auto-encoders. It is

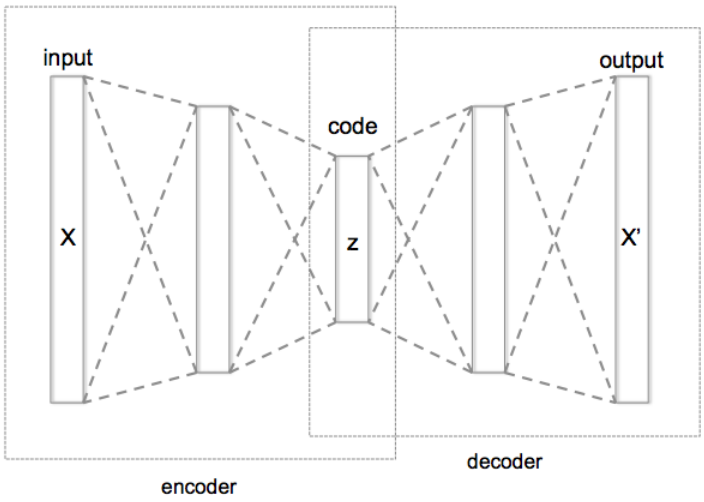


Figure 2: Structure of Auto-encoder neural network. Adapted of [23]

commonly used for the dimensionality reduction. The Figure 3 presents a example of the Deep Auto-encoder structure.

The Figure 3 presents a example of the Deep Auto-encoder structure.

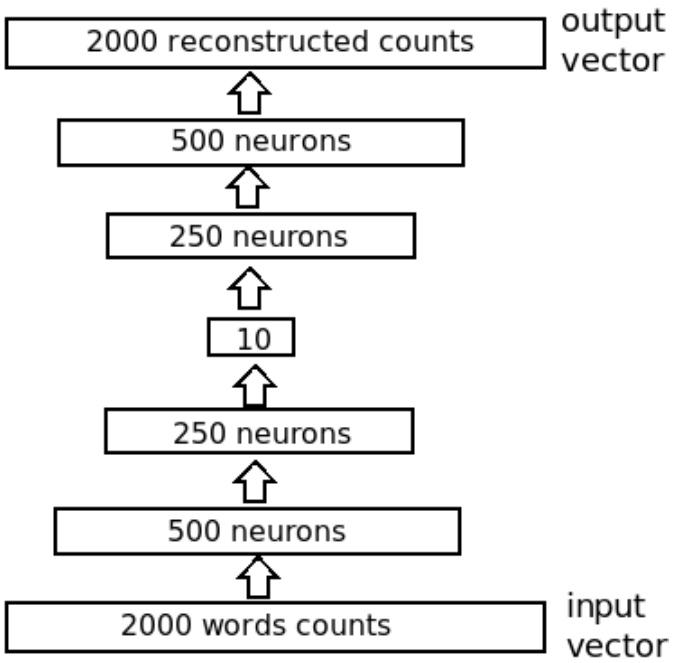


Figure 3: Structure of Deep Auto-encoder neural network. Adapted of [23]

The Auto-encoder and the its variation Deep Auto-encoder were used in this work to detect features in tweets. These features are applied in the MLP neural network to classification of positive and negative

sentiments about candidates to the Brazilian presidential elections.

4 Methodology

The methodology applied in this work was divided in two phases. The first phase, showed in the Figure 4, was developed to analyze the Brazilian president 2018 election to identify positives and negatives sentiments of the population. The second phase, showed in Figure 5, was to analyze data reference to Syrian and the USA armed conflict to identify threats degree about this conflict in the public media news. These methodologies are detailed in the next sections.

4.1 Phase 1: Methodology applied to analyze the Brazilian president 2018 election

The methodology developed for analyzing the Brazilian president 2018 election data, was collected from Twitter. It consists of the following steps: collect from tweets (Twitter messages) about Bolsonaro and Haddad candidates, during the period from June to September, 2018. Pre-process these tweets to eliminate characters and symbols without meanings. Then, processing of the resulting data applying the MLP network, Auto-encoder - MLP and Deep Auto-encoder - MLP. These processing result in the percentage of positive and negative sentiment found in tweets published about each candidates. Figure 4 presents the scheme of this methodology.

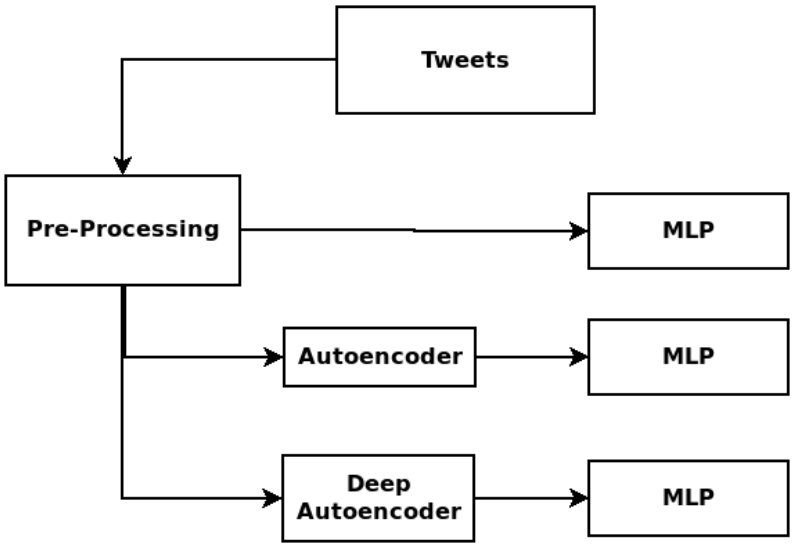


Figure 4: Methodology for analyzing the data collected from Twitter about Brazil's presidential election in 2018.

4.1.1 Tweets

The data used in this work were collected from the Twitter social network. Twitter has an API that allows you to easily and freely collect the tweets [12]. To validate the developed methodology in this work, 35,000 tweets were collected for each presidential candidate who participated in the second round of the Brazilian election in the period of June to September, 2018.

An example of the tweets referring to candidate Bolsonaro used in this work is shown as follows.

```
#Brasil #BolsonaroPresidente
Agora e a hs pra varrer essa quadrilha do poder,
nem fome
#goBolsonarogo #Bolsonaro @jairbolsonaro #BrasilDecide
Doria mostrando seu apoio a #BolsonaroPresidente
falou por todos nos!
#Bolsonaro17  #ELESIM
#eleicoes2018
#BrasilAcimaDeTudo #DeusAcimaDeTodo
Quem estiver ao lado de Lula sera derrotado.
#PTNuncaMais
#PTNAO
#ForaPT
CORRUPTO - SAFADO - PRESO
```

An example of the tweets referring to candidate Haddad used in this work is shown as follows.

```
#HaddadPresidente
#Haddad13
#eleicoes2018 #elenao
TUDO, MENOS O BOLSONARO!!!!!!
Vamos divulgar as PROPOSTAS do Haddad.
#AgoraEHaddad
So #Haddad13 pode nos livrar de
tao mitologica tragedia
Professores contam por que querem Haddad
presidente do Brasil
RejeiÃ§Ã£o das mulheres:
Haddad 49\% x 41\% Bolsonaro.
```

4.1.2 Pre-Processing

The tweets were collected in the Portuguese language. To pre-process this text this phase was divided in two steps, as follows.

- Tokenization, in which one eliminates characters that have no meaning in the text, such as α, #, http,:, ?, !.

- Word Embedding, in which one generates the vector embedded words that contains each word of the set collected tweets. In this vector each word is presented in the form of an integer, which represents the number of times a word is repeated in the set of analyzed tweets.

  To generate embedding words vector, one uses the Embedding Tensorflow API function, which comes from an open source library for machine learning for a wide variety of tasks [16], rather than using ready-made techniques such Word2vec, Sense2vec, etc. Such choice was made with the intention to obtain the representation of the word set in the form that were collected without modifications in its meaning. After Embedding words, the values in the vector were normalized considering the maximum value existing in this vector.

4.1.3 Processing

In this phase one uses machine learning techniques MLP, Auto-encoder and Deep Auto-encoder. Also, one uses $1-gram$ method to classify sentiment in each word of tweets. The $1-gram$ method that considers one word to do the analysis of the sentiment involved in the classification. A Portuguese vocabulary of 1000 words labelled in positive sentiment equal one and negative sentiment equal zero, showed a sample in Table 1, generated by a work team and related with election, it was also used to verify if the sentiment in all the analyzed tweets were positive or negative for a given presidential candidate. In this vocabulary the most used slang in social networks [26] were considered, and information about emoticons were not considered. The next sections show details about the use of this learning techniques.

Table 1: Sample of Vocabulary with Words Classified in Sentiment Positive or Negative.

Word	Sentiment
abraçar	1
acabou	0
aceita	1
briga	0
brincadeira	0
bugada	0
fracasso	0
fraudes	0
furto	0
oportunistas	0
oposição	0
opressão	0
organizar	1
orgulho	1
sancionar	1
sangrou	0
sapão	1
saúde	1
segurança	1
votos	1
zelar	1
zoar	0

The MLP

In this work, the Keras API was used to generated the multilayer perceptron neural network to obtain the classification of the collected tweets. Keras is an open source API that works with high-level neural networks, written in Python and allows to develop applications together with the TensorFlow API (an end-to-end open source machine learning platform) [16]. It allows a quick experimentation, that is, to focus on the result in the shortest possible time [8]. This data set was divided in training, validation and test, considering the percentage of 60%, 20%, and 20%, respectively. As truth of the MLP the Portuguese vocabulary of 1000 words specific of the election domain was used. So each tweet is processed and its words are classified in positive or negative words considering the vocabulary created for this work. After, the twitter is classified according with how many positive or negative words were found in the message. The architecture of the MLP is presented in Table 2.

Table 2: The Architecture of MLP Neural Network applied in the work.

Activation Function	sigmoid
Epochs	100
Learning Rate	0.0001
Hidden Layers	2
Hidden Layers Neuron Numbers	3

The Auto-encoder - MLP

The Keras API again was used to constructed one auto-encoder to dimensionality reduce the collected tweets vector. The result of the auto-encoder application was a word vector with similar features. The word vector was applied in the MLP to classify the sentiment towards a given candidate. The MLP Network is the same one used in the classification with the original data. The architecture of the Auto-encoder is presented in Table 3.

obtida

Table 3: The Architecture of Auto-encoder - MLP Neural Network applied in the work

Activation Function	encoder=RELU decoder=sigmoid
Epochs	100
Learning Rate	0.0001
Hidden Layers	1
Hidden Layers Neuron Numbers	encoder=3000 decoder=35000

The Deep Auto-encoder - MLP

The Keras API was used to constructed one deep auto-encoder to dimensionality reduce the collected tweets vector. The intention of using deep auto-encoder was to present how the increase of layers in the auto-enconder can improve its performance. The result of the deep auto-encoder application was a word vector with similar features. The word vector also was applied in the MLP to classify the sentiment towards a given candidate. The MLP Network is also the same one used in the classification with the original data. The architecture of the deep auto-encoder is presented in Table 4.

Table 4: The Architecture of Deep Auto-encoder - MLP Neural Network applied in the work

Activation Function	encoder=RELU decoder=sigmoid
Epochs	100
Learning Rate	0.0001
Hidden Layers	encoder = 3 decoder = 3
Hidden Layers Neuron Numbers	encoder=3000,1000,500 decoder=1000,3000,35000

4.2 Phase 2: Methodology applied to analyze the armed conflict Social Media data

The methodology developed for analyzing the armed conflict Social Media data collected consists of the following steps: collect news about armed conflict between Syrian and the USA from media news, pre-processing of the news to eliminate characters and symbols without meanings, processing of these data

with the MLP network, Auto-encoder and Deep Auto-encoder techniques and and result degree of threats found in public media news exchanged between rulers of Syria and the USA. Figure 5 presents the scheme of this methodology.

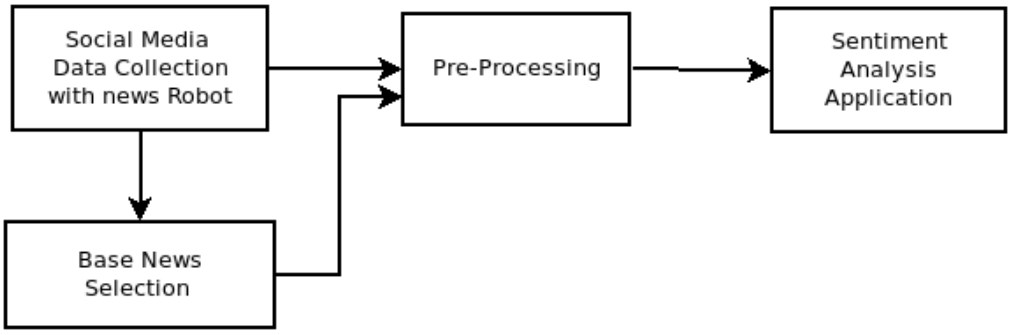


Figure 5: Methodology flow chart of the phases sequence followed for the development of the project.

4.2.1 Social Media Data collection with News Robot

The social media are responsible for publishing the events that occur daily in the world. Indeed, news are published of Heads of States referring to each other in an aggressive manner. Many of these verbal aggression may often culminate in some form of conflict. In this work, data is collected from the most relevant official social media in the world supported by news agencies (e.g. Reuters) in the English language. It is stored in a .csv file, generated for this work, considering the information of date and the internet access address (URL). For the speed-up of the collection process, news trackers are freely available, such as the News-bot [25]. It is considered to search for words related to the subject such as threats, conflicts, weapons, death, among others.The robot collected 40 news related to the domain of threat of armed conflict. All those related to the conflict between Syria and the USA in the period from January 2016 to April 2017. Thus, each url of news stored in the .csv file are read, and pre-process to calculate the threats level. The threat level is a value between 0 and 1, and corresponds to the similarity percent of the text with base news. Table 5 show one sample of .csv file structure of collected news in social media.

Table 5: Sample of .csv file structure with Date and URL of the Collected News.

Date	URL
01/03/16	reuters.com/article/us-mideast-crisis-syria-israel..
14/03/16	theguardian.com/world/2016/mar/14/syria-chemical-weapons ..
11/08/16	nytimes.com/2016/08/12/world/middleeast/syria-chlorine..
11/08/16	amnesty.org/en/latest/news/2016/08/syria-fresh-chemical..
25/08/16	csis.org/analysis/unpacking-syrias-chemical-weapons-problem..
13/09/16	bellingcat.com/news/mena/2016/09/13/chemical-attacks-syria..
16/09/16	foreignpolicy.com/2016/09/16/chemical-weapons-watchdog..
13/03/17	time.com/4699178/us-troop-increase-syria-raqqa-isis..
15/03/17	theguardian.com/world/2017/mar/15/syria-conflict-study..
05/04/17	politifact.com/truth-o-meter/article/2017/apr/05/revisiting..

4.2.2 The Base News Selection

In this work, a text file is generated from the news collected from social media by the chat bot. It is used to calculate the similarity. This domain is a base news selected by the News-bot robot, considering words that make some kind of reference to armed conflicts. Some of the words, like arms, attack, force, threat, weapons, tanks and conflict were presented as a reference for the robot news search. This text contains

terms that express the emotion of threat among the leaders of the countries and was constructed using some of the collected news. Following part of the text used in the Syria and the USA armed conflict work is shown.

```
Even though Moscow is in fact targeting Syria opposition and rebels â killing thousands
of civilians and destroying infrastructure such as hospitals, water plants, bakeries,
and schools it can justify hitting areas where Nusra is present because the
groups forces cannot be separated from other rebel factions. The US Secretary
of State John Kerry, at that time, pursued a joint command-and-control centre to
co-ordinate US and Russian aerial operations. Instead of working with long-established
groups inside Syria, the US military and CIA have tried a series of programmers of
vetting,training and equipping moderate fighters outside the country, all of which have
failed. Meanwhile, the US blocked any assistance to anti-Assad factions in
southern Syria, and detached from the critical battlefront in the north-west, where
Russia and the Assad regime have laid siege to opposition-held areas of Aleppo
city. The power of the blended rebel forces was made clear when a rebel offensive
turned the tide in the battle for Aleppo. The US could belatedly recognise the
folly of its artificial labels and establish lines of co-operation with the groups
inside Syria.
```

4.2.3 The Pre-Processing

In the pre-processing phase of data collection from social media on Syria/US armed conflict, the tokenization technique was applied to eliminate characters that have no meaning for the text, such as accent and punctuation characters (α, #, http,: ,?,!,.). The tokenization process is applied to the base text and social media news using NLTK [5] API.

4.2.4 The Sentiment Analysis Application

After the creation of the .csv file with the news collected and organized, the process of sentiment analysis the news begins. The URL information contained in the .csv file is used to read each news item in real time. As this reading is done via Web using the BS4 library of the Python 3.7 programming language, one access a .html (Hypertext Markup Language) file, which contains the news information within the $< p >< /p >$ paragraphs tags. The result of this process is a text containing the news information to be analyzed.

The treatment of the information contained in the text is processed using the concepts of Natural Language Processing using the tokenization step described in Section 3.1 for the elimination of symbols and characters that have no meaning representation for the text. After the tokenization of the information, the news analysis phase begins by applying the concepts of Sentiment Analysis through the library SpaCy (Industrial-Strength Natural Language). SpaCy is a free open source library for advanced Natural Language Processing in Python. SpaCy was designed specifically for use in production and helps create applications that process and understand large volumes of text. It can be used to build information extraction or natural language understanding systems, or to pre-process text for Deep Learning.

The project uses Sentiment Analysis based on topics or resources for extracting information from selected news. In this process of extracting information, a basic text was initially defined, which is considered to represent 100% of threats to be verified. This basic text is defined empirically, considering the knowledge of the people on the subject addressed. After the base news selection, an analysis is made of the percentage of similarity of the new data news to be analyzed within the base news using the SpaCy library. The result of this process is the degree of threat that each news item has in relation to the extreme event analyzed. This degree of threat is calculated and stored for each news item accessed and stored in the Similarity (%) field of the .csv file. The described process is illustrated in Figure 6.

The results this methodology's application are shown in the Section 5.2.

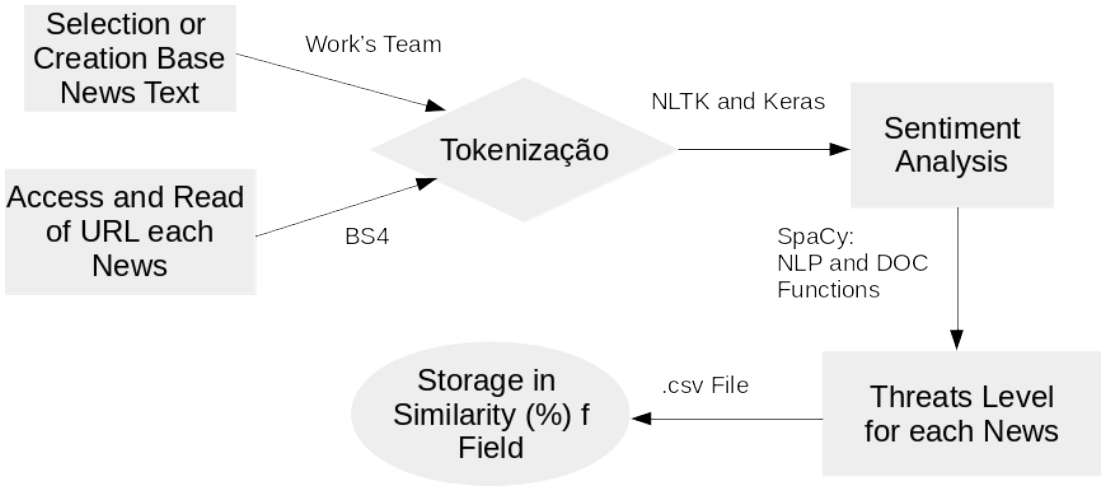


Figure 6: Methodology to Application of Sentiment Analysis for Calculating the Level of Threat.

5 Results

The results of the analysis of Twitter data for the 2018 presidential election are presented in the section 5.1. The results of the analysis of social media data on armed conflicts are presented in the section 5.2

5.1 Data analysis results of the Twitter for the 2018 presidential election

The results obtained in this work about the analysis of the unbalanced Twitter data were related to the application of MLP machine learning techniques, Auto-encoder and Deep Auto-encoder in the same database of the candidates Bolsonaro and Haddad, from the second round of the 2018 Brazilian presidential elections. The combination of these analysis made it possible to compare such techniques and confirm the population's sentiment in relation to the candidates in the election. The results of this comparison are presented separately by candidates and groups of techniques.

In Table 6 the results of the processing of collected data for the candidate Bolsonaro are shown. It is observed that the percentage of positive sentiment, which represents the number of words expressing positive information, in the text about the candidate Bolsonaro is always greater than the negative sentiment for the three techniques analyzed. It is also verified that the technique of Auto-encoder combined with MLP has presented a better classification of the positive sentiment.

Table 6: Comparison of the results of the application of machine learning techniques MLP, Auto-encoder-MLP and Deep Auto-encoder-MLP of the candidate Bolsonaro data.

	BOLSONARO	
Machine Learning Techniques	**Sentiment Positive (%)**	**Sentiment Negative (%)**
MLP Representation	50.199	49.801
Auto-encoder-MLP Representation	53.339	46.661
Deep Auto-encoder-MLP Representation	51.394	48.606

The Table 7 presents the results of the processing of the data collected for the Haddad candidate. It is observed that the percentage of negative sentiment, which represents the number of words expressing negative information, is always greater than the positive sentiment for the three techniques analyzed. It

is also verified that the technique of Auto-encoder combined with MLP was also the one that presented a better classification of the negative sentiment.

Table 7: Comparison of the results of the application of machine learning techniques MLP, Auto-encoder-MLP and Deep Auto-encoder-MLP of the candidate Haddad data.

	HADDAD	
Machine Learning Techniques	**Sentiment Positive (%)**	**Sentiment Negative (%)**
MLP Representation	46.614	53.386
Auto-encoder-MLP Representation	44.622	55.378
Deep Auto-encoder-MLP Representation	45.618	54.382

The results of the Table 6 demonstrate that the candidate Bolsonaro has a higher percentage of positive than negative sentiments in the analysis using the three machine learning techniques: MLP (50,199% positive and 49,801% negative), Auto-econder-MLP (53,339% positive and 46,661% negative) and Deep Auto-encoder-MLP (51,394% positive and 48,606% negative). The result of the Table 7 demonstrate that the candidate Haddad has a higher concentration of negative than positive sentiments in the analysis using the three machine learning techniques: MLP (46,614% positive and 53,386% negative), Auto-encoder-MLP (44,622% positive and 55,378% negative) and Deep Auto-encoder-MLP (45,618% positive and 54,382% negative). Thus, it is noteworthy that the results presented are consistent with the true result of the election, which had the candidate Bolsonaro as winner. In the actual analysis the sentiments of sarcasm and irony in the text were not considered. In the results, it was observed that the increase number of words in the vocabulary can be better for the classification of tweets either in positive or in negative. To verify the results obtained in the classification, the accuracy was calculated for data training of each neural network model used in this work. These results are presented in Table 8. It is observed that the Deep Auto-encoder-MLP technique had a better accuracy in the classification of the information with the value of 78.5 %. While the Auto-encoder-MLP techniques had 74.3 % and MLP had 59.2 % for unbalanced data inside of the presented context.

Table 8: Showing of the accuracy results of the application of machine learning techniques MLP, Auto-encoder-MLP and Deep Auto-encoder-MLP.

Machine Learning Techniques	**ACCURACY**
MLP Representation	0.592 (59.2%)
Auto-encoder-MLP Representation	0.743 (74.3%)
Deep Auto-encoder-MLP Representation	0.785 (78.5%)

5.2 Data analysis results of the Social media about Armed Conflict

This section presents the results obtained by analyzing the news from social media according to the methodology presented in the section 4.2. The news were collected from January 2016 to April 2017. The end date of the collection refers to the eve of the launch of the 59 Tomahawk missiles by the USA with destination to Syria. The launch day was not considered in the analysis to give an idea of the degree of threat that preceded the attack.

The Table 9 presents the results with values greater than or equal to 90% of the similarity. The Date column shows the publication of the news date, the URL column contains the information of which social

media the news was published in and the column Similarity(%) shows the similarity's percentage of the news analyzed with the base news with high level of threat.

Table 9: Sample of Data analysis of the result of the armed conflict between Syrian and the USA to determine the probability of threats.

Date	URL	Similarity (%)
01/03/16	reuters.com/article/us-mideast-crisis-syria-israel..	93.30
14/03/16	theguardian.com/world/2016/mar/14/syria-chemical-weapons ..	90.08
11/08/16	nytimes.com/2016/08/12/world/middleeast/syria-chlorine..	92.71
11/08/16	amnesty.org/en/latest/news/2016/08/syria-fresh-chemical..	93.82
25/08/16	csis.org/analysis/unpacking-syrias-chemical-weapons-problem..	94.62
13/09/16	bellingcat.com/news/mena/2016/09/13/chemical-attacks-syria..	91.78
16/09/16	foreignpolicy.com/2016/09/16/chemical-weapons-watchdog..	94.37
13/03/17	time.com/4699178/us-troop-increase-syria-raqqa-isis..	95.27
15/03/17	theguardian.com/world/2017/mar/15/syria-conflict-study..	90.93
05/04/17	politifact.com/truth-o-meter/article/2017/apr/05/revisiting..	93.62

It was analysed 40 news about threat between Syrian and the USA until the launch of Tomahawk missiles from the Mediterranean Sea to Syria. The Figure 7 presents the time evolution of the news threats during the analysed period. The spikes in the graphic represent the biggest values of similarity. Observe that graphic values of threats precede the missile attack by one week, in the period between March 30, 2017 and April 5, 2017 (dashed line), were increasing until the last analysed news. That is the signature for hostilities commence and the launching of the 59 Tomahawks.

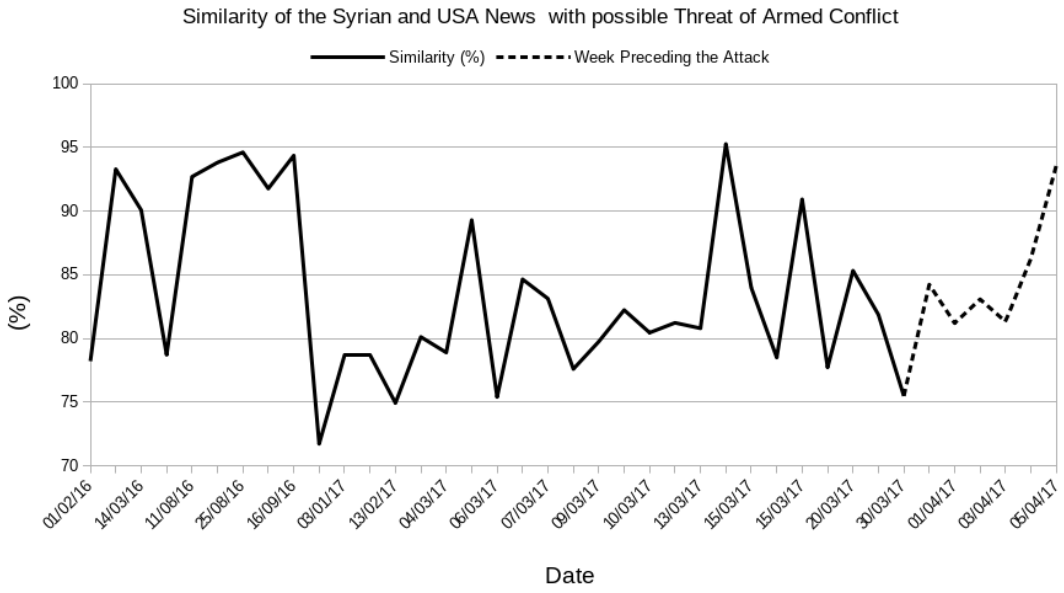


Figure 7: Time series of the threat level in the analyzed news of the conflict between Syria and the USA.

6 Conclusion

This work shows two approaches for sentiment analysis application to help understand emotional opinions of a population. The first approach data from the second round of the 2018 Brazilian presidential election collected from the social network Twitter was compared with analysis produced by the machine learning

techniques MLP, Auto-encoder and Deep Auto-encoder. The concept of polarized classification was used in this work to analyze the sentiments of the population in relation to the candidates Bolsonaro and Haddad. The results showed that although a direct analysis of the texts was carried out, without considering the sarcasm and irony commonly found in the Portuguese language, satisfactory results were obtained in relation to the real performance of candidates in the election. The use of the $1-gram$ method allowed an analysis of the context of the information used and not only of the isolated tweets. This first approach was applied in Portuguese language to verify presidential election results in Brazil using social network Twitter data and polarization technique of NLP and obtained the accuracy of 59,2% for MLP, 74,3% for Auto-encoder MLP and 78,5% for Deep Auto-encoder MLP.

The second approach, was applied in social media data in English language to verify the level of threats in the armed conflict between Syrian and the USA. This analysis use Sentimental Analysis Based on Topics or Features by choice of the one news defined as base news. The analysis of these data used a news base to compare with others news collected by a robot news collector and applying in spaCy NLP Python API. The result shows that the level of threats increased in the days before the attack of April 5, 2017 with a percent of similarity by sentiment threats between 70% and 95%.

For future work, we intend to extend the application of $N-gram$ concepts, to use the identification of sarcasms and irony in the Portuguese language texts, to increase vocabulary to improve the result of classification, and to use balanced Twitter data to better classification of information in social media. Also, it is intended to generated a public armed conflict portfolio with content of the .csv file, to apply citizen science to help the selection of news base, and to apply this methodology to predict future armed conflicts.

Acknowledgements

Marilyn Minicucci Ibañez acknowledge to the Federal Institute of São Paulo, campus of São José dos Campos, for financial support.
Reinaldo Roberto Rosa thanks partial support from FAPESP under grant number 2014/11156-4.

References

[1] A. E. Ali, T. C. Stratmann, S. Park, J. Schöning, W. Heuten, and S. CJ. Boll. Measuring, understanding, and classifying news media sympathy on twitter after crisis events. *arXiv.org e-Print archive*, 2018.

[2] M. N. Antunes, C. H. Silva, M. C. S. Guimarães, and M. H. L. Raba¸co. Monitoramento de informação em mídias sociais: o e-monitor dengue. *Scielo - Transinformação*, 26(1):1–11, 2014. $http://dx.doi.org/10.1590/S0103-37862014000100002$.

[3] R. Arunachalam and S. Sarkar. he new eye of government: Citizen sentiment analysis in social media. In *Proceedings...*, pages 23–28, Nagoya, Japan, 2013. in International Joint Conference on Natural Language Processing, IJCNLP.

[4] R C Balabantaray, Mudasir Mohammad, and Nibha Sharma. Article: Multi-class twitter emotion classification: A new approach. *International Journal of Applied Information Systems*, 4(1):48–53, September 2012. Published by Foundation of Computer Science, New York, USA.

[5] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python — Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, USA, 2009.

[6] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. pages 120–128, Sydney, Australia, July 2006. Association for Computational Linguistics.

[7] M. A. C. Cepik. *Espionagem e Democracia*. Editora FGV, Brasil, 1 edition, 2003. 936 p.

[8] F. Chollet. Keras: The python deep learning library, 2015. https://keras.io/ . Access in May 2020.

[9] CNN. Cnn - breaking news, latest news and videos, 2019. https://edition.cnn.com/. Accessed in November 2019.

[10] J. C. A. Cuadrado and D. P. G. Gómez-Navarro. Un modelo linguístico-semántico basado en emociones para la clasificación de textos según su polaridad e intensidad, 2011.

[11] V. Dhawan and N. Zanini. Big data and social media analytics. *Research Matters: A Cambridge Assessment*, 18:36 – 41, 2014.

[12] J. Dorsey, E. Williams, B. Stone, and N. Glass. Twitter developer documentation - api, 2017. *https : //dev.twitter.com/overview/api*. Accessed in 28 de abril de 2017.

[13] C. Garvey and C. Maskal. Sentiment analysis of the news media on artificial intelligence does not support claims of negative bias against artificial intelligence. *OMICS: A Journal of Integrative Biology*, 24(2):286–299, 2020.

[14] R. Ghosh, Davi, K., and V. Ravi. A novel deep learning architecture for sentiment classification. In *Proceedings...*, pages 1–5, Dublin, Ireland, 2016. in Recent Advances in Information Technology (RAIT) International Conference on Date of Conference.

[15] I. Goodfellow, Y. Bengio, and A. Courville. Deep learning. Book in preparation for MIT Press, 2016.

[16] Google. Tensorflow - an end-to-end open source machine learning platform, 2019. https://www.tensorflow.org/. Accessed March 2016.

[17] The Guardian. News, sport and opinion from the guardian's us edition the ..., 2019. https://www.theguardian.com/us. Accessed November 2019.

[18] J.-X. Hao, Y. Fu, C. Hsu, X. R. Li, and N. Chen. Introducing news media sentiment analytics to residents attitudes research. *Journal of Travel Research*, 0(0), 2019.

[19] S. O. Haykin. *Neural Networks and Learning Machines*. Pearson, New York, 3rd edition, 2008. 936 p.

[20] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *SCIENCE*, 313(1):504–507, july 2006. *https : //science.sciencemag.org/content/*313/5786/504.

[21] N. Indurkhya and F. J. Damerau. *Handbook of Natural Language Processing*. Taylor and Francis Group, LLC, USA, 2nd edition, 2010. 676 p.

[22] S. Kumari. impact of big data and social media on society. *Global Journal for research Analysis*, 5:437–438, 03 2016.

[23] Q. V. Le. A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks. Technical report, Google Brain, Mountain View, CA, 2015. 20p. Tutorial.

[24] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997. 421 p.

[25] Newsbot. Related news at the click of a button, 2019. https://getnewsbot.com/. Accessed in July 2019.

[26] A. Padilha. Gírias atuais mais usadas na internet, 2020. https://www.dicionariopopular.com/girias-atuais-internet/ . Accessed in February 2020.

[27] J. C. S. Reis, P. Gon¸calves, M. Araújo, A. C. M. Pereira, and F. Benevenuto. Uma abordagem multilíngue para análise de sentimentos. *Universidade Federal de Minas Gerais (UFMG)*, pages 1–12, 2015. Brasil.

[28] Reuters. Reuters news agency:world's largest news agency, 2019. https://www.reuters.com/. Accessed in November 2019.

[29] H. Sagha, N. Cumminsy, and B. Schuller. Stacked denoising autoencoders for sentiment analysis:a review. *WIREs Data Mining Knowl Discovy*, 7, 2017. doi: 10.1002/widm.1212.

[30] U. Sivarajah, Z. Irani, S. Gupta, and K Mahroof. Role of big data and social media analytics for business to business sustainability: A participatory web context. *Industrial Marketing Management*, 86:163 – 179, 2020.

[31] spaCy. Industrial-strength natural language processing, 2019. https://spacy.io/. Accessed in May 2019.

[32] S. M. Zavattaroa, P. E. French, and S. D. Mohanty. A sentiment analysis of u.s. local government tweets: The connection between tone and citizeninvolvement. *Government Information Quarterly*, 3(39):1–9, 2015.

[33] Y. Zhang, L. Shang, and X. Jia. Sentiment analysis on microblogging by integrating text and image features. In Springer, editor, *Advances in Knowledge Discovery and Data Mining*, pages 52–63. Springer, 2015.