Sentiment Analysis in Social Medias for Threats Prediction of Natural Extreme Events

S

Marilyn Minicucci Ibañez
National Instituto of Spatial Research, Federal Institute of São Paulo, Brazil

Reinado Roberto Rosa
Lab for Computing and Applied Mathematics, Brazil

Lamartine Nogueira Frutuoso Guimarães
National Instituto of Spatial Research, Brazil

INTRODUCTION

The end of the twentieth century was marked by the advent of the internet and consequently this favored those massive volumes of information, from the most different fields of knowledge, to circulate through the World Wide Web – WWW (Castells, 2003). The disclosure and sharing of this information by society introduced new understandings of how this volume of data could be used to generate value for the most diverse areas of knowledge and thus bring benefits to society. Among the most varied domains of information circulating on the WWW, natural extreme events deserve attention for a more detailed study and understanding of their causes, consequences, and possible prevention.

An extreme event is characterized by a sequence of small events generated by human emotions or some reaction of nature that can evolve into a larger event reaching up to a catastrophic event (Rosa et al., 2019) (Clauset, 2018) (Ibanez et al. 2022). The natural extreme event model considered is deforestation, as it has a great influence on the life of society (Santos et al., 2017). Due to its complexity, a multidisciplinary solution would assist in understanding its evolution and possible prevention of this natural extreme event model.

Based on the context presented, this work proposes a multidisciplinary solution that considers the threats of droughts and fires in the Brazilian Amazon region as the evolution of deforestation. To carry out the threat analysis, used as case studies for the natural extreme event data collected from social media, such as newspapers and magazines. Considered the social media of large national circulation, about the occurrence of droughts, fires, and deforestation in the years 2015, 2016, 2017, 2018, 2019 and 2020. The collection of this information is carried out using Google's web search engine (Google, 2016) which performs a search for news related to the topics addressed about drought, burning, and deforestation threats. Each collected news is stored and grouped considering the increasing order of its publication date (Ibanez et al., 2022).

The news collected from social media are processed using data science and machine learning techniques that allow identifying some nature reaction present in a text document. As per the context of the chapter, the reaction identified in the analyzed news texts is the threat of the natural extreme event addressed. The machine learning technique used for the identification of the threat in the news is sentiment analysis, being applied in the chapter with the Natural Language Processing technique, which performs

DOI: 10.4018/978-1-6684-7366-5.ch046

text analysis (Ibanez et al., 2022). Sentiment analysis makes it possible to identify how similar a text is to a given context, using a base text with words referring to a domain (Bird et al., 2009), in this case, the threat of the extreme event. Thus, for each news story collected, the percentage of threat existing in its text is analyzed and identified (Ibanez et al., 2022).

The result of this collection is a threat portfolio with time series with the threat degree referring to these extreme events. This portfolio is used as input for the P-model algorithm (Rosa et al., 2019) to generate a time series with endogenous characteristics. This time series model is characterized by considering only elements that exert some internal influence on the analyzed domain (Rosa et al., 2019) (Sornette, 2006). The generated endogenous series is used as input to a system developed with machine learning, more specifically deep learning, for the creation of the extreme event prediction model. The prediction of the evolution of the threat of occurrence of drought, fire, and deforestation is being carried out for a future period of three months (Ibanez et al., 2022).

The deep learning technique is being used due to its properties of automatically extracting features and nonlinear correlations existing in the data (Goodfellow, et al., 2016). Deep learning concepts are applied using a neural network framework that uses Long Short-Term Memory – LSTM recursive neural networks in a system developed using the TensorFlow Keras deep learning API (Chollet, 2015). LSTMs are suitable for classifying, processing, and predicting time series due to the unknown length delays of some series (Hochreiter, Schmidhuber, 1997).

Validation of the prediction is performed using the statistical tool, DTW (Dynamic time warping) that compares the predicted period with the first three months of the subsequent year. Proof of the endogeneity pattern is performed by an analysis of the endogenous time series generated by the threats using the statistical tools of mean, variance, standard deviation, skewness, and kurtosis.

BACKGROUND

This section presents the study of state of the art related to sentiment analysis and prediction of natural extreme events considering the droughts, fires, and deforestation.

The article Rainfall prediction for Manaus, Amazonas with artificial neural networks (Lima; Guedes, 2015) addressed the problem of rainfall prediction in Manaus using multilayer artificial neural networks. The input data were obtained from an automatic weather station during the years from 1970 to 2015. The performance factor considered was the normalized root-mean-square error. According to the observed results, a feedforward neural network with 2 hidden layers with 10 neurons each performed best in solving the issue. In the work, it was also observed that the use of recurrent neural networks had no influence on the performance gain of the problem addressed.

In Predicting amazon fires for policymaking (Morello et al., 2016) presented contributions with public policies for fighting fires in the Amazon. The work focused on identifying at the municipal scale the main variables for predicting fire occurrences. As a result, a data panel unprecedented in the literature was built from satellite images and socioeconomic data, covering the years 2008, 2010 and 2012. The work concluded that of the 41 potential fire predictors evaluated, only 9 were significant at a tolerable level of uncertainty, comprising deforested areas, pasture and forest areas, indigenous lands, temperature, and soil texture.

The paper Integrating remotely sensed fires for predicting deforestation for redd (Armenteras et al., 2017) presented work in which it addresses that the United Nations Reducing Emissions from Deforestation and Forest Degradation (REDD+) program leaves a gap in decision-making about REDD+ interventions.