Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Providing a greater precision of Situational Awareness of urban floods through Multimodal Fusion

Thiago Aparecido Gonçalves da Costa [*], Rodolfo Ipolito Meneguette, Jó Ueyama

Institute of Computer Sciences and Computational Mathematics (ICMC), University of São Paulo (USP), São Carlos, São Paulo, Brazil

ARTICLE INFO

Keywords:
Multimodal fusion
Text classification
Floods
Situational awareness
Disaster management
Machine learning

ABSTRACT

Floods are a source of anxiety for the people living in the city of São Paulo, Brazil. Every year, the city suffers a financial loss of more than US$ 35 million caused by damage to property, and countless lives are lost as a result of the flooding. Strategies such as Disaster Management can reduce and prevent flash floods and also assist their victims. Moreover, social networks such as Twitter can play a crucial role in offering assistance at the Disaster Management response stage because they produce a massive number of localized geo messages, which can help identify the flood victims. We argue that the mining of social network opinion raises a severe challenge since the Machine Learning (ML) algorithms cannot reflect the context of the messages in-depth, and thus this needs to be improved by combining textual data with contextual data. In this study, we combine multiple sources of weather data with the social network posts to obtain a Situational Awareness (SAW) of flash floods and hence be able to support the Disaster Management Response stage in São Paulo. We show that by combining meteorological with social network data, we can identify the flood victims with a greater degree of precision. The model that was designed for identifying the victims of flooding in São Paulo achieved a 87.69% rate of precision. Furthermore, contextual data inclusion led to a 22.8% increase in SAW of urban floods from tweets and contextual data, which shows that multimodal approaches are more promising than unimodal strategies. Finally, this work adopts a novel approach first by demonstrating that simply applying social network posts to ML strategies is not an efficient method of obtaining a SAW of flash floods. Second, we proved through this study that empirical strategies for establishing potential flood areas are more effective than the adoption of geostatistical approaches (Semivariogram) because the Semivariogram technique is more suitable for understanding scenarios that have not had any prior human interference or damage (e.g., when locating mineral reserves). Thus, particularly in the case of São Paulo, when the trash is disposed of close to drainage systems, this causes clogging of gutters and hence leads to floods.

1. Introduction

Currently, the amount of flooding that ravages Brazil is becoming exacerbated. As Tucci, Hespanhol, and Netto (2003) point out, these natural disasters cause a good deal of disturbance to the affected population, such as economic losses, suspension of daily activities, contamination by waterborne diseases, and deaths. According to Haddad and Teixeira (2015), the financial damage caused by the floods in the city of São Paulo is approximately US$ 35 million per year, and 674,329 people live in areas that are prone to natural disasters (IBGE, 2010). Thus, the floods impede the city's economic growth and impair the quality of life of the people while imposing constraints on local and international business competitiveness.

Thus, the purpose of strategies such as Disaster Management is to reduce or prevent damage to society and provide support for the communities affected by natural disasters by encouraging them to show resilience and an ability to adapt to the situation, Baharin, Shibghatullah, and Othman (2009), Mendiondo (2010), Norris, Stevens, Pfefferbaum, Wyche, and Pfefferbaum (2008), Poser and Dransch (2010). Furthermore, in Disaster Management, the use of heterogeneous data sources, such as sensors, radar devices, and social networks, are crucial for mitigating the impact of natural disasters, as they can enable authorities to detect the phenomenon in advance and issue early warnings to civil defense, emergency, and humanitarian agencies (Horita, de Albuquerque, Degrossi, Mendiondo, & Ueyama, 2015). However, it is not a trivial task to extract meaningful information from these data sources and combine it with the currently available Disaster Management models.

[*] Corresponding author.
E-mail addresses: thiago.gcosta@usp.br (T.A.G. da Costa), meneguette@icmc.usp.br (R.I. Meneguette), joueyama@icmc.usp.br (J. Ueyama).

[1] url: https://twitter.com.

https://doi.org/10.1016/j.eswa.2021.115923
Received 18 January 2021; Received in revised form 31 July 2021; Accepted 15 September 2021
Available online 12 October 2021

Social networks currently produce a vast amount of useful information on natural disasters (Win & Aung, 2017). There has been an increasing interest in analyzing messages obtained from social networks to assist Disaster Management in recent years (de Albuquerque, Herfort, Brenning, & Zipf, 2015). For example, researchers are increasingly using Twitter,[1] a social network that permits users to publish short messages of up to 280 characters, to detect events such as earthquakes (e.g., Avvenuti, Cresci, Marchetti, Meletti, and Tesconi (2014), Crooks, Croitoru, Stefanidis, and Radzikowski (2013), Earle, Bowden, and Guy (2012), Sakaki, Okazaki, and Matsuo (2010)), hurricanes (e.g., Kryvasheyeu et al. (2016)), floods (e.g., Brouwer et al. (2017)), and forest fires (e.g., De Longueville, Smith, and Luraschi (2009), Spinsanti and Ostermann (2013)). Twitter messages are publicly available and easily obtainable, which explains their popularity among researchers. Moreover, according to de Bruijn et al. (2020), Kwak, Lee, Park, and Moon (2010), news about disasters tends to appear first on Twitter rather than in traditional media, with more than 20,000 messages containing words related to natural phenomena. In this way, the capture, monitoring, and analysis of tweets can reveal possible natural disasters, as well as the whereabouts of possible victims, and other vital information needed for relief organizations (de Bruijn, de Moel, Jongman, Wagemaker, & Aerts, 2018).

However, the fact that users on Twitter, write their messages colloquially, makes it difficult to extract useful knowledge. Some words are related to natural phenomena such as *rain* and *flood*, but these are often used metaphorically by users. For example, a user may write messages such as a "shower of blessings" or a "heart flooded with joy", but these messages are unrelated to natural phenomena. Thus, an efficient filtering process is needed to improve the performance of algorithms when employed for Natural Language Processing (NLP) to obtain SAW data from the published messages.

This work combines social network messages with meteorological data by Multimodal Fusion Approaches to obtain a SAW of flash floods and support the response stages of Disaster Management. We show that this combination can bring about a high degree of accuracy. It should be noted that we are not deploying new sensors but are improving the degree of accuracy of the flash flood SAW from Twitter messages and data available at servers managed by the Brazilian National Institute of Meteorology. We are able to increase the flash flood SAW precision by only relying on existing data produced by social media and currently available sensors. This approach can be easily extended to identifying the possible victims of different natural disasters in different locations and sending messages written in other languages.

Unlike other works in the literature (e.g., de Bruijn et al. (2020), Feng and Sester (2018)), our research is novel as we use the Portuguese language and it is based on real data from towns with a high incidence of flash floods (e.g., São Paulo city). To the best of our knowledge, none of the studies in the literature adopt our approach or use real data from South America. Flash flooding occurs more frequently in South American developing countries where SAW of floods and assistance to Disaster Management are of crucial importance to reduce the risk of casualties and material losses, particularly for those living in shantytowns.

Another contribution made by this study is that we provide evidence to show that the empirical strategies outperform the Semivariogram technique in defining the radius of São Paulo city's flooded areas. It is worth noting that the creation of the flood-prone areas is essential because the probability of a social media message being related to the flash flood is proportional to the distance of these areas from the user that published the tweet. It should be emphasized that an article by Feng and Sester (2018) backs up this empirical approach. Following an article by Yin and Li (2001), we also argue that flooded areas in Brazil are, to a large extent, affected by human interference with the environment, such as changing the course of rivers or streams to construct shopping malls or freeways and trash disposal close to the drainage systems which causes clogging of gutters and hence leads to floods. Therefore, this suggests an empirical approach is more beneficial than a statistical approach.

This article is structured as follows. Section 2 contains information about the research background and related works. Section 3 outlines the methodology applied in this scientific work and describes in detail the chosen datasets. Section 4 describes the experiments that were carried out and the virtual machine configurations used to complete them. Section 5 examines the evaluation metrics used in the experiments. Section 6 analyzes and discusses the experimental results. Finally, Section 7 concludes this study and makes recommendations for future work.

2. Background

This section will define the main concepts of Disaster Management, Multimodal Fusion techniques and architecture, and Clustering. Finally, we make a comparison between the works related to the proposed solution.

2.1. Disaster management

In recent decades, natural disasters have increased in intensity and frequency, causing serious losses to society, such as high mortality rates, material damage, and socioeconomic damage (Cutter & Emrich, 2005; Klomp, 2016; Kobiyama et al., 2006). In parallel, according to Pouyanfar, Tao, Tian, Chen, and Shyu (2019), the number of users of applications (e.g., smartphones, social networks, sensors, and surveillance cameras) that produce both structured and unstructured multimedia data, has grown exponentially. Even today, this information forms a considerable amount of the daily Internet traffic. This means that, opportunities are created to design systems that can assist in Disaster Management. The objective of Disaster Management is to reduce or avoid damage to society and provide assistance to victims of extreme natural phenomena (Poser & Dransch, 2010)

Thus, designing systems that can support the Disaster Management stages, such as response or recovery, must be embedded functionalities so that SAW can be obtained in a dynamic way Pouyanfar et al. (2019). Moreover, during a natural disaster, the civil defense, non-governmental organizations, and firefighters can draw on the information of these systems to make decisions that give priority to people's health and lives. According to Endsley (1988), SAW is a human cognitive process that corresponds to the ability to observe, interpret, and understand one's environment. In addition, SAW involves perceiving the elements in an atmosphere of time and space, understanding their meaning, and predicting their state in the near future (Endsley, 1988). According to Li et al. (2017), the strategies necessary to build Disaster Management systems embedded in SAW mechanisms are information collection, data extraction, and useful data integration. Thus, the most advanced techniques of Data Mining, Artificial Intelligence, and NLP are required to design systems that are precise and robust.

A widely cited work in the Disaster Management literature that employs social networks is that of Sakaki et al. (2010). These authors state that the purpose of their paper is to monitor tweets to detect earthquakes and typhoons, which involves creating a mechanism that is capable of detecting disasters; they have also trained a model based on a Support Vector Machine (SVM) that uses historical and manually labeled data (Sakaki et al., 2010). As well as this, they designed a classifier that is capable of detecting typhoons and earthquakes and has performance superior to that of Japan's weather agency (Sakaki et al., 2010). With the aid of the information obtained from Twitter, the computational mechanism can detect these phenomena within 3 min (Sakaki et al., 2010).

Additionally, on the basis of research conducted by Yin, Lampert, Cameron, Robinson, and Power (2012), a computational system was devised that uses NLP and Data Mining to obtain SAW of natural disasters from tweets and assist Disaster Management. This computational system comprises five modules: data capture, emergency detection, data classification, information grouping, and geotagging (Yin et al., 2012). In its first stage, Yin et al. (2012) point out that the tweets extraction module collected 66 million geolocalized tweets from 2.51 million users from 2010 to 2011 through the Twitter Streaming API (Application Programming Interface). The emergency detection module is responsible for monitoring Twitter data to detect possible emergencies in the real world by filtering the keywords of natural disasters (Yin et al., 2012).

Thus the classification module was used by the authors to categorize the social media messages. The authors carried this out in the following stages: pre-processing text, textual transformation to numerical characters, training in a manually labeled database, and testing models using cross-validation in a validation database (Yin et al., 2012). The authors achieved results for the SVM classifier with a precision rate of 87.5% (Yin et al., 2012).

The module for compiling information had the function of arranging the Twitter messages in terms of the topic of interest through Online Incremental Clustering (Yin et al., 2012). The result of the Silhouette coefficient of this grouping strategy is 0.42 (Yin et al., 2012). Finally, the geotagging module has the capacity to visualize the information generated by the other modules (for example, textual groupings, emergency notifications, among others) (Yin et al., 2012).

As shown in the studies outlined above, ML algorithms must be applied to obtain the SAW of emergencies through messages published in the social networks. Thus some of the approaches adopted in the literature, use SVM (e.g., Caragea et al. (2011), Wang and Manning (2012)), Random Forest (RF) (e.g., Cobo, Parra, and Navón (2015)), Naive Bayes (NB) (e.g., Imran, Elbassuoni, Castillo, Diaz, and Meier (2013), Li, Caragea, Caragea, and Herndon (2018)), Decision Tree (DT) (e.g., Purohit et al. (2013)), Logistic Regression (LR) (e.g., Huang and Xiao (2015)), and Neural Networks (e.g., Avvenuti, Cresci, Del Vigna, Fagni, and Tesconi (2018), Devlin, Chang, Lee, and Toutanova (2018), Feng and Sester (2018), Joulin, Grave, Bojanowski, and Mikolov (2016), Kim (2014), Lai, Xu, Liu, and Zhao (2015), Lorini et al. (2019)). However, according to de de Bruijn et al. (2020), when these algorithms are applied for NLP, more discrepancies are found in the results than is the case with a human interpretation because people can refer to the context to understand the messages. According to Yin et al. (2012), unless NLP techniques are adapted to the context of social media messages, they cannot effectively handle the text. The exponential growth of works in the literature that are concerned with combining textual data with a context to solve this problem, include this strategy, which is called Multimodal Fusion. Moreover, Huang and Xiao (2015) point out that textual data must be combined with twitter data from meteorological sensors to improve the computational model designed by the authors, which can obtain SAW of natural disasters from the Twitter messages that are divided into the regions affected by Hurricane Sandy.

2.2. Multimodal fusion

Multimodal Fusion is the technique that combines various multimedia data (e.g., audio, video, sensors, texts, among others) to carry out specific tasks such as event detection, tracking of human beings, detection of semantic contexts, and identification of speakers from audiovisual data (Xie & Guan, 2013). The purpose of this merger is to improve the quality of multimedia content analysis because by combining data from different media sources, this technique can achieve more accurate results and ensure that decision-making is more assertive (Zhou, Leung, & Yao, 2013). For example, a data scientist can apply Multimodal Fusion analysis in a film because the textual resources usually found in subtitles can be combined with audiovisual data, and ensure that the decision-making of these cinematographic works is more accurate. Alqhtani, Luo, and Regan (2015) divided Multimodal Fusion into three categories: fusion at the resource level (Early Fusion), fusion at the decision level (Late Fusion), and a combination of two types of multimedia data which is called Hybrid Fusion.

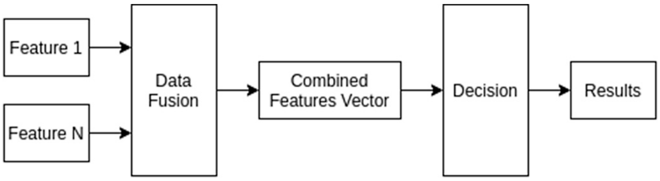

Fig. 1. Generic Early Fusion method based on Atrey et al. (2010).

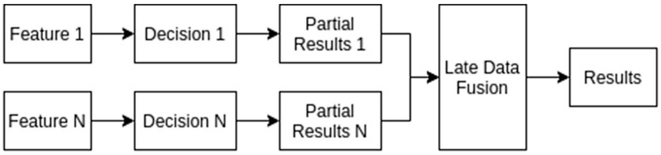

Fig. 2. Generic Late Fusion method based on Atrey et al. (2010).

Below is a detailed explanation of the types of multimedia data fusions:

- **Early Fusion** collates the information from the multimedia data at the characteristic vector level and then forwards it to the data analysis unit (Alqhtani et al., 2015). According to Atrey, Hossain, El Saddik, and Kankanhalli (2010), this strategy can be beneficial because it allows the correlation of the data features to be investigated at an initial level and ensures there will be only one resulting characteristic vector, which will make later tasks easier to carry out. However, in the opinion of de Bruijn et al. (2020) merging data early in a situation where intermodal dynamics is more complex, can cause overfitting problems. According to Lopes (2015), in this type of merging, there may be characteristic synchronization problems. Fig. 1 shows the generic Early Fusion method based on the recommendations of Atrey et al. (2010).
- **Late Fusion** involves processing the initial resources by means of their respective decision-making units and then forwarding the combined results to the data analysis unit (Alqhtani et al., 2015). Lopes (2015) states that in this strategy, each of the characteristics of media is processed independently and can thus ensure that the decision-making relies on more appropriate techniques and does not have problems with synchronization features. However, de Bruijn et al. (2020) argue that merging data late can result in an excessive simplification of intermodal dynamics, and according to Lopes (2015), the correlation between feature vectors is a question that has not been explored. Fig. 2 shows the generic Late Fusion method based on Atrey et al. (2010).
- **Hybrid Fusion** aims at reaping the benefits of a merger between the resource level and decision-level, by combining similar features based on the previous merger strategy and the discrepant and complex features of late fusion (Lopes, 2015). However, in the view of Lopes (2015), although this strategy is more expensive computationally than the others, it is able to solve the Late Fusion and Early Fusion strategic problems. Fig. 3 shows the generic Hybrid Fusion method based on the system of Lopes (2015).

In the literature on Multimodal Fusion, some works rely on textual data from Twitter combined with different types of media information to obtain SAW of natural phenomena for Disaster Management support (e.g., Liu, Jiang, and Zhao (2018), Poria, Cambria, Howard, Huang, and Hussain (2016), Zadeh, Chen, Poria, Cambria, and Morency (2017)). In addition, there is the research of Alqhtani et al. (2015) who employ a multimedia data fusion method based on the Dempster–Shafer Theory of Evidence applied to the fusion of textual and visual resources

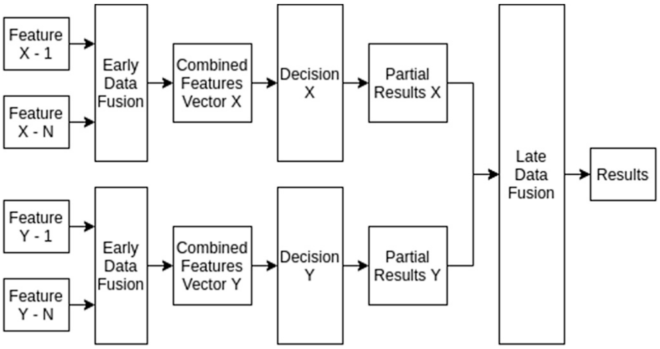

Fig. 3. Generic Hybrid Fusion method based on Lopes (2015).

obtained from Twitter. Even in Alqhtani et al. (2015), the results of textual information from the preprocessing of tweets are transformed into numerical characters through the Term Frequency–Inverse Document Frequency (TF–IDF) technique. Moreover, in the visual extraction of information, the authors employ the SIFT (Scale-invariant feature transform) technique and group the descriptors with the aid of the K-Means clustering algorithm (Alqhtani et al., 2015). Alqhtani et al. (2015) demonstrate that this system achieves a 97% rate of accuracy in classification, compared with the 93% accuracy of the strategy that only relies on textual information and 86% of the process that only uses visual information.

Furthermore, in the work of de Bruijn et al. (2020), a hybrid multimodal neural network model was designed with textual data and meteorological data to detect the SAW of natural disasters through tweets with the aim of support Disaster Management. In this work, the authors extracted information from Twitter by filtering messages that used keywords related to flooding in English, French, Spanish, and Indonesian (de Bruijn et al., 2020). Moreover, de Bruijn et al. (2020) obtained information from the weather sensors of a global precipitation database at the locations mentioned in the messages and date and time of the published tweets. de Bruijn et al. (2020), claimed that by using the neural network with heterogeneous data, they obtained a 91% rate of precision, which is higher than the accuracy of a neural network without meteorological information.

In general, the works found in the literature that have Multimodal Fusions of textual and contextual data (e.g., images, and sensors, among other devices), were able to obtain SAW of emergencies from tweets to support Disaster Management. These works adopt strategies at the textual classification stage: textual preprocessing; transformation of text into numeric characters from Bag Of Words (BOW) and TF–IDF; training and testing of ML models (SVM, NB, DT, LR, RF) in a manually labeled database. It was also noted in the literature, that in several cases, the use of the Hybrid Fusion method is preferable to the others because it combines the benefits of previous and late fusion. Furthermore, in work carried out by de Bruijn et al. (2020), the authors state that there is a lack of research in the literature on the combined use of geographical contextual data with textual data to improve the process of textual classification. Thus, our scheme has the advantage of obtaining SAW of flash floods from tweets that can assist Disaster Management with the aid of textual, meteorological, and geographic data.

2.3. Clustering

The purpose of the clustering technique is to group together datasets, as argued by Faceli, Lorena, Gama, Carvalho, et al. (2011), who state that the objects that are labeled in the same way, share characteristics that are relevant to the domain under study — that is, they are in some way, similar. As a result, several papers in the literature use clustering algorithms to group together unlabeled data, such as textual features (e.g., Jan (2020), Rosa, Shah, Lin, Gershman, and Frederking (2011)), and geographical information (e.g., Boettcher and Lee (2012), Feng and Sester (2018)).

The number of daily active flooding clusters in a city is an arduous task to determine. Yin and Li (2001) state that flooding is a natural phenomenon that is often caused by misguided alterations to river systems, such as the restriction of the flow of water through the construction of dams, the destruction of vegetation, and soil erosion near river beds that is responsible for rainwater runoff, among other examples of environmental damage. In addition, flooding in the urban perimeters depends on the weather conditions and infrastructural facilities of the cities. On a rainy day, flooding may occur in the town center of São Paulo, but not on another day. This means that clustering approaches that define a fixed amount of flood areas tend to be ineffective in this scenario (e.g., K-means). In contrast, strategies that are able to find an arbitrary number of clusters (e.g., Agglomerative clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify the Clustering Structure)) tend to be ideal for this scenario because the degree of grouping is volatile.

In the opinion of Feng and Sester (2018), clustering approaches that group data arbitrarily (e.g., DBSCAN) do not require a parameter for the resulting number of groups but require a maximum distance for group formation. For this reason, several studies, in the literature, that use clustering algorithms to collate information from a dataset, define the maximum distance between the samples so that the groups can be established either groups empirically (e.g., Boettcher and Lee (2012), Feng and Sester (2018)) or statistically (e.g., Trujillo and Izquierdo (2005)).

DBSCAN and OPTICS cluster algorithms on the basis of density. According to Faceli et al. (2011), these algorithms assume that clusters are areas with an object of high density, which means that the space between the groups is a region of low density. The OPTICS algorithm ensures the neighborhood radius of the groups is more flexible than that of DBSCAN (Ankerst, Breunig, Kriegel, & Sander, 1999). Furthermore, Faceli et al. (2011) add that the Agglomerative Clustering algorithm is based on hierarchical clustering, i.e., it generates a proximity matrix of a sequence of nested partitions. This approach starts with n groups with only one object and successively forms the partitions by ordering the rest of the data (Faceli et al., 2011).

2.4. Literature analysis

We have provided a summary of the related works that use heterogeneous data to detect natural disaster events. These studies seek to obtain SAW from these events by adopting unimodal computational strategies and Multimodal Fusion models of tweets and contextual information to assist Disaster Management.

Table 1 makes a comparison between the features present in related works and those of our work, including the following: types of data employed; which ML algorithms used; kind of Multimodal Fusion employed (e.g., previous, late or hybrid); the capacity to group historical flood events, and if so, which techniques were used by the author; which stages of Disaster Management were employed by the related works to offer assistance (e.g., mitigation, preparation, response, and recovery).

The related works outlined in Table 1 created computational mechanisms capable of classifying the messages obtained from Twitter. Thus, firstly, the researchers employed a manual classification strategy to generate labels for tweets. Later, in the pre-processing stage, the researchers applied the data to the cleaning and data transformation processes.

The researchers employed textual filtering techniques for the data cleaning, which included removing stop words, and separation of tokens (e.g., Alqhtani et al. (2015), de Bruijn et al. (2020), Sakaki et al. (2010), Yin et al. (2012), among others). In fact, during the data

Table 1
Summary of related works found in the literature.

Research	Data type	ML algorithm	Multimodal Fusion Type	Identification of flood-prone regions	The stages of assistance to GD
Sakaki et al. (2010)	Tweets	SVM	Not	Kalman Filter and Particle Filter	Preparation and Response
Yin et al. (2012)	Tweets	NB, SVM, Online Incremental Clustering	Not	Not	Response
Alqhtani et al. (2015)	Tweets and Images	K-Means	Early Fusion	Not	Response
de Bruijn et al. (2020)	Tweets and Meteorological Features	Convolutional Neural Network (CNN)	Hybrid Multimodal Fusion	Not	Response
Our work	Tweets, Meteorological Features, Historical Occurrences of Flooding	RF, DT, SVM, NB, LR, DBSCAN, OPTICS, Agglomerative Clustering	Early Fusion based on Atrey et al. (2010), Late Fusion based on Atrey et al. (2010), Tweet Hybrid Multimodal Fusion and Flood Hybrid Multimodal Fusion	Grouping of historical occurrences of flooding	Response

transformation stage, the authors of the related works applied TF–IDF (e.g., Alqhtani et al. (2015), Yin et al. (2012)) and Word Embeddings (e.g., de Bruijn et al. (2020)). In our work, we use the same strategies employed for the related works in the text pre-processing on the data cleaning stage. However, we allow for the possibility of correcting words that are written informally. We used BOW, TF–IDF, and Word Embeddings techniques (e.g., Fast Text and Word2Vec) in the data transformation phase. In this research, we evaluated how far these methods for converting symbolic data to numerical data can impact the performance of the ML algorithms in the multimodal sample space.

In the works of Alqhtani et al. (2015), Sakaki et al. (2010), Yin et al. (2012), the authors relied entirely on information obtained from Twitter for the development of computational approaches whose aims are to obtaining SAW of natural disasters. Contextual information (e.g., sensor data, radars, among others) must be included to design more accurate SAW models. In light of this, we have combined textual information obtained from Twitter with meteorological data and historical flood events in this work. Our goal is to obtain a model that can achieve a SAW of natural disasters and provides Disaster Management more effectively than the other unimodal versions in the literature.

In the work of de Bruijn et al. (2020), the authors devised a model of Hybrid Multimodal Fusion capable of obtaining a SAW of natural disasters and assisting Disaster Management. The information used by the authors to train and evaluate the model was based on textual data from Twitter and meteorological data (de Bruijn et al., 2020). Moreover, the authors used software designed by de Bruijn et al. (2018) to locate the messages through the places described in the tweets (de Bruijn et al., 2020). However, this strategy leads to areas of uncertainty that impair the accuracy of the Multimodal Fusion model proposed by these researchers. In view of this, we will only use geo-located information from Twitter in our work to mitigate the problem of erroneous space–time combinations between tweets and climate data.

Finally, in this work, we make use of the following meteorological information: precipitation, humidity, temperature, dew point temperature, and atmospheric pressure. This includes more climatic data for the Multimodal Fusion models to ensure that a SAW of natural disasters is obtained more precisely than was the case with de de Bruijn et al. (2020) where the authors only rely on precipitation information and the amount of accumulated rain.

3. STIGMA: a mechanism to SupporT dIsaster manaGement through Multimodal datA

This article sets out a computational mechanism capable of combining data from social networks with meteorological information through a Hybrid Mode Multimodal Fusion strategy, called STIGMA — SupporT dIsaster manaGement through Multimodal datA. We designed a Hybrid Multimodal Fusion model to obtain a SAW of flooding and assist in the response stage of Disaster Management. This approach will combine messages published on Twitter with meteorological data to provide a precise flood SAW that does not require new climate sensors.

STIGMA also solves a wide range of problems caused by the uncertainties present in the data (for example, informal textual data and erroneous voluntary geographic information). These data provide a SAW of floods imprecisely and cause both social and economic damage to the local communities.

The Hybrid Multimodal Fusion put forward in this article increases the precision of the flooding SAW through Twitter messages and data made available by servers managed by the Brazilian National Institute of Meteorology. Moreover, STIGMA improves the precision of the flooding SAW by only relying on existing data produced by social media and the currently available sensors. This makes it easy to extend it to determining the potential victims of different natural disasters in different locations and understanding wrote messages in other languages.

3.1. An overview of STIGMA

STIGMA is a methodology that combines seven key components (Fig. 4). These components are responsible for the acquisition of multimodal information for the classification and combination of data from different media. The goal of STIGMA is to provide more accurate flooding SAW from tweets and meteorological data. Each component is briefly described below.

- **Data Collection:** this component is responsible for extracting messages published by Twitter users, obtaining data from the meteorological sensors of INMET[2] (National Institute of Meteorology), and extracting the historical flood events of CGE-SP[3] (Center

[2] url: https://portal.inmet.gov.br/.

[3] url: https://www.cgesp.org/v3/.

for Climate Emergency Management of the City of São Paulo), which are described in detail in Subsubsection Section 3.2.

- **Data selection and Manual Classification:** this component involves the process of data selection and manual classification of information, which are described in detail in Section 3.3. It includes the manual classification of tweets into “relevant” (e.g., messages posted on Twitter related to floods or rainfall) and “irrelevant” (e.g., tweets not related to floods or rainfall). This stage is crucial as it maps flood-related tweets in a database written in Portuguese.
- **Textual Features:** this component is responsible for the pre-processing of the tweets, and includes the strategies of data cleaning and data transforms. The techniques adopted to the data cleaning are textual filtering, the removal of stop words, and the separation of tokens. With regard to the data transforms, we decided to change the information from symbolic to numerical. We provide details of these strategies in Section 3.4.1.
- **Characteristics of Historical Flood Events:** this component is responsible for pre-processing historical flood events. It employs strategies for (a) cleaning historical events of flooding (as reported by CGE-SP), (b) the conversion of historical flood data into geographical coordinates, and (c) the grouping of geographical coordinates. The aim of these strategies is to establish which areas are prone to flooding, as a means of supporting the Multimodal Fusion Models. The procedures followed by this component are set out in Section 3.4.2.
- **Climatic features:** this component involves pre-processing the meteorological information of the city of São Paulo, and in Section 3.4.3, it explains the procedure required to infer the missing climate data in the sample space.
- **Ground truth Creation:** this component establishes the training and testing dataset of the Multimodal Fusion models. We also describe the Spatio-temporal fusion of tweets, meteorological data, and historical flood events in Section 3.5.
- **Multimodal Fusion:** this component adopts four approaches to Multimodal Fusion. (i) Early Fusion — this aims at combining information at the resource level; (ii) Late Fusion — this seeks to combine information at the decision level; (iii) Tweet Hybrid Fusion — the purpose of this is to combine textual data at the decision level with other meteorological data at the vector resource level; (iv) Flood Hybrid Fusion — this seeks to combine textual information at the vector resource level with other meteorological data at the decision level. We describe the procedure followed by this component in Section 3.6.

In the next subsections, we will describe each component in more detail.

3.2. Data collection

The data used are from São Paulo city because it has a population of approximately 12 million people and a larger number of social media users than any other city of Brazil (IBGE, 2010). Additionally, São Paulo is the second city in Brazil in terms of the number of people who live in flood-prone areas (IBGE, 2010). It is hoped that the application of this model by the authorities responsible for Disaster Management will reduce or prevent the loss of life and damage to property in São Paulo.

As pointed out by de Andrade, Restrepo-Estrada, Delbem, Mendiondo, and de Albuquerque (2017), in São Paulo city, the tweets related to flooding in January 2016, tend to be clustered near meteorological stations. This provides an opportunity for the Multimodal Fusion of textual data with meteorological information to improve the SAW of the messages published on Twitter.

This module assisted the data collection procedure and the formation of our dataset because it meant that we were able to collect data from Twitter, which is a social network that allows free access to data through its API and also to obtain heterogeneous (e.g., “date”, “time”, “text”, “latitude”, and “longitude”). Moreover, we formed a partnership with a research group called AGORA[4] (A Geospatial Open collaboRative Architecture for building resilience against disasters and extreme events) at the ICMC[5] (Institute of Mathematics and Computer Science) - USP[6] (University of São Paulo). With the aid of the AGORA research team, we obtained 4.031 million geo-referenced tweets from the city of São Paulo during the period from November 7, 2016, to November 8, 2018. A crawler written in Python coded by the AGORA members returned these data by JavaScript Object Notation (JSON). The delimiting boxes of the north (−46.95, −23.62, −46.28, −23.33) and the south (−46.95, −23.91, −46.28, −23.62) cover the city of São Paulo.

Additionally, when obtaining the meteorological data of the city of São Paulo, it was necessary to draw on the data in the historical data section of the INMET website. In this way we found 33,576 occurrences of automatic measurements made every hour by the INMET weather radars. The type of information includes the “date”, “hour”, “temperature”, “humidity”, “dew point temperature”, “atmospheric pressure” and “precipitation”, and the data extraction period lasted from January 1, 2015 to October 30, 2018.

In addition, we designed a web crawler written in Python to obtain historical data about flooding. This picks up the information from the CGE-SP website by means of the BeautifulSoap library[7] and returns the date, time, address, and status of historical flood events, and retains this information in a MongoDB database. A computational mechanism for data extraction must be designed because there is no automating API provisioning of the historical datasets for flooding in São Paulo. We used MongoDB as a database in this software because it has a better performance than traditional databases in this area and includes a large volume of data and complex SQL queries (Aghi, Mehta, Chauhan, Chaudhary, & Bohra, 2015). We extracted 4,904 items of historical flooding data from January 1, 2015, to October 30, 2018, for this research study.

3.3. Data selection and manual classification

After we had collected the tweets, it was necessary to carry out a mapping of this textual information. This involves making a text filter to assist in extracting information about the flooding of the sample space. Thus, we made a computational engine written in Python to perform a substring search of keywords belonging to Table 2 and only return tweets that contain the keywords related to flooding. Table 2 shows the results of filtering tweets based on a set of keywords written in Brazilian Portuguese. These keywords are the various representations of the words “rainy” and “flooding” in Brazilian Portuguese and were selected and extended from articles found in the literature (e.g., Andrade (2020), de Andrade et al. (2017), de Assis, de Albuquerque, Herfort, Steiger, and Horita (2016)). We filtered 13,802 Twitter messages (0.34% of 4.031 million), and all these messages contained at least one of the predefined keywords selected. It should be added that the text of the tweets is not blank, and their coordinates are within the city of São Paulo.

Next, three assessors manually classified the 13,802 previously filtered tweets, 4,527 of which were “relevant” (i.e., Twitter posts related to flooding or rain) and 9,275 “irrelevant” (i.e., Twitter posts not related to flooding or rain) - these values are based on the definition of de Bruijn et al. (2020). During the manual classification, the assessors found several tweets that were written informally. These were classified as irrelevant because they included metaphors such as “a downpour of blessings”, “rain down curses”, “a flood of tears”. Table 3 shows examples of tweets that were manually classified as relevant and irrelevant, and that make up the sample data text of this work.

[4] url: http://www.agora.icmc.usp.br/site/language/en/.

[5] url: https://www.icmc.usp.br/.

[6] url: https://www.usp.br/.

[7] url: https://pypi.org/project/beautifulsoup4/.

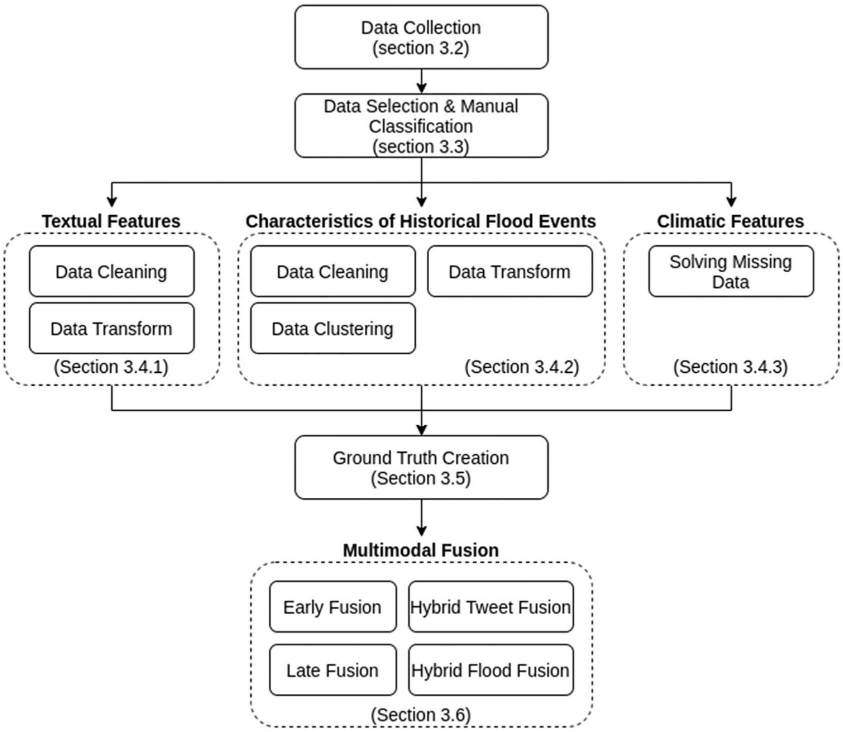

Fig. 4. Flowchart design of the models able to obtain SAW of flooding through tweets and contextual data automatically for the city of São Paulo.

Table 2
Keywords in Brazilian-Portuguese with their translation to English in brackets, based on Andrade (2020), de Andrade et al. (2017).

alagamento (flooding), alagado (flooded), alagada (flooded), alagando (flooding), alagou (flooded), alagar (flood), chove (it rains), chova (rain), chovia (it rained), chuva (rain), chuvarada (rain), chuvosa (rainy), chuvoso (rainy), chuvona (rainy), chuvinha (drizzle), chuvisco (drizzle), chovendo (raining), dilúvio (heavy rain), enchente (flood), enxurrada (flood), garoa (drizzle), inundação (inundation), inundada (flooded), inundado (flooded), inundar (flood), inundam (flood), inundou (flooded), temporal (storm), temporais (storms), tromba d'água (waterspout)

Table 3
Twitter messages that are either relevant or irrelevant to the context of rains and floods in Brazilian-Portuguese and English.

Date/Time	Tweet (PT-BR)	Tweet (ENG)	Rating
2016-10-11 / 20:24:40	Na hora que eu ia comprar o pastel começou chover	When I was going to buy the pastry, it started to rain	relevant
2017-07-08 / 00:48:25	Noite de musical "Cantando na chuva" @Teatro Santander	Night of the musical "Singing in the rain" @Santander Theater	irrelevant
2016-22-12 / 19:13:10	Forte temporal causa alagamentos na cidade de São Paulo	Stormy weather causes flooding in the city of São Paulo	relevant
2017-09-08 / 14:38:29	Baldes de chuva, baldes de lágrimas	Buckets of rain, buckets of tears	irrelevant

Following this, we measure the degree of agreement between the assessors through Krippendorff's Alpha coefficient, which is a statistical measure capable of computing the degree of agreement between two or more observers such as judges or evaluators (Krippendorff, 2004). This enabled us to reach a 0.66 agreement between the assessors who manually classified the tweets. According to Landis and Koch (1977), this result demonstrates that there was a substantial consensus between the evaluators. With Krippendorff's Alpha coefficient, the closer to zero (0), the more likely there is to be a lack of agreement between the judges (Landis & Koch, 1977). In contrast, the closer to one (1) means there is almost a perfect degree of consensus (Landis & Koch, 1977).

The reason we selected the historical data of floodings was that all their characteristics were present, i.e., date, time, address, and the status of the flood events. Thus, in the case of this dataset, it was necessary to exclude the missing data, because the application of ML techniques in this scenario can lead to errors in execution. In addition, historical flooding information from CGE-SP is sensitive, because the characteristic "address" which has to be completed is processed by computational mechanisms of geolocation inference (e.g., Google Geocoding API[8]). However, the "date" and "time" features are required for creating the truth set, as explained in Section 3.5.

Finally, for this research study, we selected the meteorological data from January 1, 2015, to October 30, 2018, which is an intersection of the dates of the published tweets and the weather data available. Thus, unlike the historical data on flooding, the meteorological data may contain a few cases of missing items, but these should not be excluded because this information does not need to be complete to be employed by mechanisms of Artificial Intelligence and Multimedia Data Fusion. This scenario includes an "inducer" for the missing data of

[8] url: https://developers.google.com/maps/documentation/geocoding/overview.

the meteorological features to estimate the missing attributes (e.g., by means of the linear interpolation technique). Additionally, the "date" and "time" of this dataset are required for the creation stage of the truth set outlined in Section 3.5.

3.4. Feature engineering

This section describes Feature Engineering, which is a strategy to extract more knowledge from the data belonging to the analyzed sample space (Jalal, 2018). Thus, in this section, we will describe the components responsible for processing the following data: text, meteorological, historical flood events in the city of São Paulo.

3.4.1. Textual features

The preprocessing stage is of crucial importance for improving the quality of the datasets since they may have different characteristics, dimensions, or formats (Faceli et al., 2011). This preprocessing is also essential because it rejects values that have unknown information, noise, attributes with low prediction value, and a disproportionate number of examples of each class. Thus, the inherent phases of preprocessing that are used before the ML algorithms to improve data quality are: handling imbalanced data, data cleaning, data integration, data transformation, and dimensionality reduction (Faceli et al., 2011).

The textual pre-processing of the tweets carried out in this study consisted of two stages, the first being called "data cleaning", where the purpose of this was to reduce the noise of the data. The second stage is called "data transformation", and involves converting data from symbolic values to numerical values. A Python algorithm capable of performing the data cleaning was designed for the first stage. In the case of this algorithm, the NLTK[9] (Natural Language Toolkit) and Hunspell[10] libraries were used to assist the NLP. Incidentally, the Hunspell library uses LibreOffice for the correction of erroneous words.

This algorithm makes use of a dictionary of colloquial keywords and a hashtag dictionary. Both contain conversions of words written in a colloquial language to the cultured norms of the Portuguese language. Many messages published on Twitter are informal and erroneous, and thus computational mechanisms must adapt to this context.

Moreover, the difference between both dictionaries is that in the hashtag dictionary, the meaning of some abbreviations and "foreignisms" are included (e.g., "sp4you" translated to Brazilian-Portuguese as "A cidade de São Paulo para você" and to English as "The city of São Paulo for you"). As well as this, the algorithm uses a Word Embeddings model designed by NILC[11] (Interinstitutional Center for Computational Linguistics) Hartmann et al. (2017) of the ICMC-USP that is trained from massive text corpus in Brazilian-Portuguese to check the existence of words contained in tweets and written in Portuguese. The following are some characteristics of the algorithm which we created:

1. Removal of email addresses, profiles, and duplicate characters from messages published on Twitter;
2. Conversion of tweets into tokens;
3. Replacement of hashtags because they include colloquial or foreign expressions;
4. Removal of special characters and stopwords from tweets;
5. Correction of features found in the Twitter messages that are not within the cultural norms of the Portuguese language.

Afterwards, we sent the results from the data cleaning stage to the data transformation stage, which is essential because as several ML algorithms have difficulty in using textual information, it is necessary to adopt approaches that transform symbolic values into numerical values.

Several conversion strategies were employed to convert textual data to numerical data during the data transformation stage of this study, such as that of de Aguiar, Faiçal, Ueyama, Silva, and Menolli (2018), de Bruijn et al. (2020). Additionally, we compared the impact of these approaches on the degree of precision of the Multimodal Fusion models by employing the selected techniques:

- BOW;
- TF–IDF;
- Word Embeddings of a Word2Vec and FastText type where both are the type of Skip-Gram and Continuous Bag of Words (CBOW) with 50 and 100 dimensions in Brazilian-Portuguese, trained by Hartmann et al. (2017).

The BOW consists of a technique in which each document represents a vector of the words present in the files of the dataset, i.e., each vector represents a tweet with the most frequent terms contained in the message and dataset (Matsubara, Martins, & Monard, 2003). In contrast, TF–IDF is a technique in which the weight of the most frequent terms contained in the textual set is in inverse proportion to the frequency of these words throughout the textual database (Zhang, Yoshida, & Tang, 2011). Furthermore, the data transformation approach that relies on Word Embeddings, provides the representation of each term in the text in latent vectors that allow the occurrence of words to be discovered semantically.

3.4.2. Characteristics of historical flood events

This preprocessing stage corresponds to two phases, the first being "data cleaning", and the second, "data transformation". The aim of "data cleaning" is to remove the inconsistencies in the information about addresses during flood events. Since there are several informally written and abbreviated addresses in the database (for example, "prof" stands for "professor" and "the old fepasa" stands for the "Hungarian community" address), it is difficult to obtain the coordinates from the geographic coordinate inference system. Additionally, "data transformation" converts historical flood events to geographic coordinates.

In view of this, during the "data cleaning" stage, we devised a computational mechanism written in Python to remove the inconsistencies from the addresses with the aid of the Pandas[12] library. This mechanism makes use of a dictionary of inconsistencies that has been prepared to translate the abbreviated and misspelled words into the cultural norms of the Portuguese language.

The "data transformation" phase uses the information obtained from the "data cleaning" phase. At this stage, we use the service provided by Google called Geocoding API, which is an NLP mechanism capable of geographical inference, i.e., interpreting the addresses of flood events and giving the geographical coordinates of natural phenomena. Table 4 shows some examples of the historical data on flooding and their respective geographical coordinates with the dates and periods of the occurrences.

Currently, there is no available tool that gives the geographical coordinates of the flood-prone areas in the city of São Paulo. The probability of a person being affected by the phenomenon is inversely proportional to their distance from the incident. Thus, it is necessary to group the historical occurrences of flooding to determine which areas are most prone to flooding, since the users affected by natural disasters tend to live close to the flooded areas.

In this paper, we applied and compared the performance of several clustering algorithms, such as DBSCAN, OPTICS, and Agglomerative Clustering. The aim at this stage is to define the flooding zones of the city of São Paulo, including the area studied in this phase, which was assisted by the library called Scikit-Learn[13] (Pedregosa et al., 2011).

[9] url: https://www.nltk.org/.

[10] url: https://pypi.org/project/cyhunspell/.

[11] url: http://www.nilc.icmc.usp.br/.

[12] url: https://pandas.pydata.org/.

[13] url: https://scikit-learn.org/stable/.

Table 4
Addresses and coordinates of historical flooding data.

Date / Period	Address	Latitude	Longitude
October 28, 2019 / 07:36 to 11:02	Celso Garcia Avenue with José de Alencar Street, São Paulo, Brazil	−23.54009	−46.61158
October 28, 2019 / 08:25 to 10:08	Guilherme Cotching Avenue number 16, São Paulo, Brazil	−23.53019	−46.59849

Two strategies are compared in this study to define the maximum distance of clustering formation, i.e., the radius between n samples to consider them part of the same group. The first is empirical (variation from 100 m to 5 km), and the second statistical (Semivariogram). The Semivariogram is a geostatistical technique that determines the spatial dependence between georeferenced sample points (Isaaks & Srivastava, 1989).

3.4.3. Climatic features

At times there are missing attributes in the weather dataset. Some ML techniques may have execution errors if the training set has one or more ineffectual features. The approach most often adopted in the literature to overcome this problem is to create an inductor that makes use of the other attributes in the dataset to estimate the value of the missing attribute. The interpolation consists of a technique that calculates intermediate values from a known dataset by means of mathematical formulas or graphic procedures (Shryock, Siegel, & Larmon, 1973). The linear interpolation method uses linear polynomials to build new data points within a known sample space (Meijering, 2002). In light of this, we devised a computational mechanism in Python that uses the Pandas library to infer missing data through Linear Interpolation of the other data in the historical flooding dataset.

3.5. Ground truth Creation

After completing the Feature Engineering stages, it is necessary to create the ground truth dataset of this work so that the Multimodal Fusion and Textual Classification models can be trained and tested. Fig. 5 shows the methodology for creating the ground truth dataset by combining the multimodal data through an augmented matrix strategy

Fig. 5 shows the process of heterogeneous data merging. First, we store the textual features of the pre-processed tweets in a dataset. After that, we filter the meteorological data that was published on the same date and time as the tweets and then store the results of this filtering process in another dataset.

Moreover, the results of the incidence of tweets in flood-prone areas verification are stored in a vector that is later transposed to aid the “combination” process of multimodal data. Thus, the probability of any message being related to a natural disaster is inversely proportional to the distance of the tweet from the natural phenomena. If the users who published the Twitter messages are inside a flooded area, then there is a more significant probability that it is related to the respective natural phenomenon. Additionally, we rely on the Haversine formula (1) to check if the tweets are in the flooded area (Winarno, Hadikurniawati, & Rosso, 2017). This is a mathematical formula used to calculate the geographical distance between two points, by taking into account the degree of curvature of the Earth.

Eq. (1) shows the math formula widely used in geographic information systems. Where “**d**” is the distance between the two geographic coordinates, “**r**” the radius of the earth (6,731 km), “**lat1**” and “**long1**” are the latitude and longitude of the first point, “**lat2**” and “**long2**” are the latitude and longitude of the second point.

$$d = 2.r.arcsin\left(\sqrt{sin^2\left(\frac{lat2 - lat1}{2}\right) + cos(lat1).cos(lat2).sin^2.\left(\frac{long2 - long1}{2}\right)}\right) \tag{1}$$

Furthermore, we combined the multimodal data through the augmented matrix strategy (2). This is a mathematical technique for adding columns to a matrix (Marcus & Minc, 1992), and the resulting dataset contains input data for the Artificial Intelligence mechanisms employed in this research.

$$data = \left(M|C|F^T\right) \tag{2}$$

Eq. (2) expresses the math formula devised in this research study, and this is based on the concept of multimodal data integration explained by Atrey et al. (2010), de Bruijn et al. (2020) and the augmented matrix theory exposed by Marcus and Minc (1992). Where “**data**” is the dataset resulting from the combination of heterogeneous information, “**M**” is the dataset of textual features (e.g., BOW, TF–IDF, and Word Embeddings) or the results of the classification model for textual data, “**C**” is the dataset of meteorological features (e.g., humidity, temperature, precipitation, atmospheric pressure, and dew point temperature) or the results of the classification model for meteorological information, and “**F**” is the vector based on the verification of the incidence of tweets in areas prone to flooding in the city of São Paulo”.

In addition, we labeled the samples as positive for the “ground truth” dataset according to the following conditions: tweets that were geographically located in flooded areas, the textual contents of Twitter messages related to “flooding” or “climate”, and the meteorological data belonging to timetables of flood events. At the same time, we labeled all the possible contrary combinations as negative to the “ground truth” dataset.

Finally, we balanced the database with 50 percent positive and 50 percent negative occurrences. This is because the performance of the ML algorithms will be impaired if the database has imbalanced data and will favor the class that has the most significant number of examples (Faceli et al., 2011). In addition, we defined 20% of this database for validation, and the remaining 80% of the data for the training stage of the Multimodal Fusion models.

3.6. Multimodal fusion

In this section, we set out four Multimodal Fusion techniques, which are shown in Fig. 6. Their objective is to obtain the SAW of flooding through tweets and support the response stage of Disaster Management. Fig. 7 shows the generic classification system used in this work, which was planned with the support of the Scikit-Learn library in Python (Pedregosa et al., 2011).

Thus, the four subcomponents for Multimodal Fusion are:

- **Early Fusion**: in this subcomponent which is based on Atrey et al. (2010), the model combines the heterogeneous information at a resource level through the augmented matrix strategy (2) and later trains a classification model to obtain the SAW of flash floods. In addition, the performance of the model was evaluated by the following metrics: accuracy, recall, f1-score. According to Atrey et al. (2010), this approach explores the correlation of characteristics at the resource level and only generates one resulting vector, which makes it easier to carry out tasks later.
- **Late Fusion:** in this subcomponent which is based on Atrey et al. (2010), one classification model is trained for textual data and another for meteorological data. After this, this information is combined with the results of the incidence of tweets in the

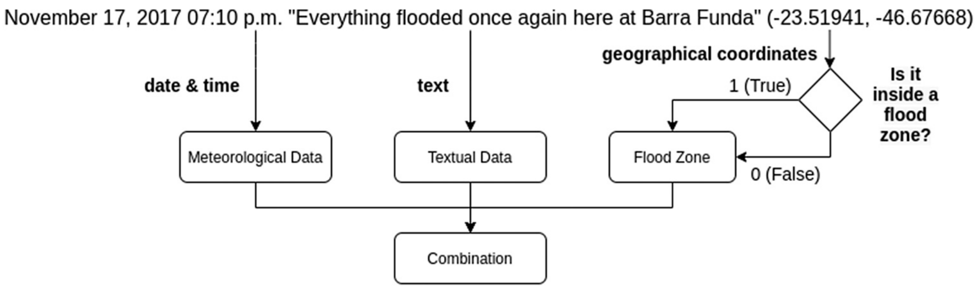
November 17, 2017 07:10 p.m. "Everything flooded once again here at Barra Funda" (-23.51941, -46.67668)
date & time
text
geographical coordinates
Is it inside a flood zone?
1 (True)
0 (False)
Meteorological Data
Textual Data
Flood Zone
Combination

Fig. 5. Ground truth Creation process.

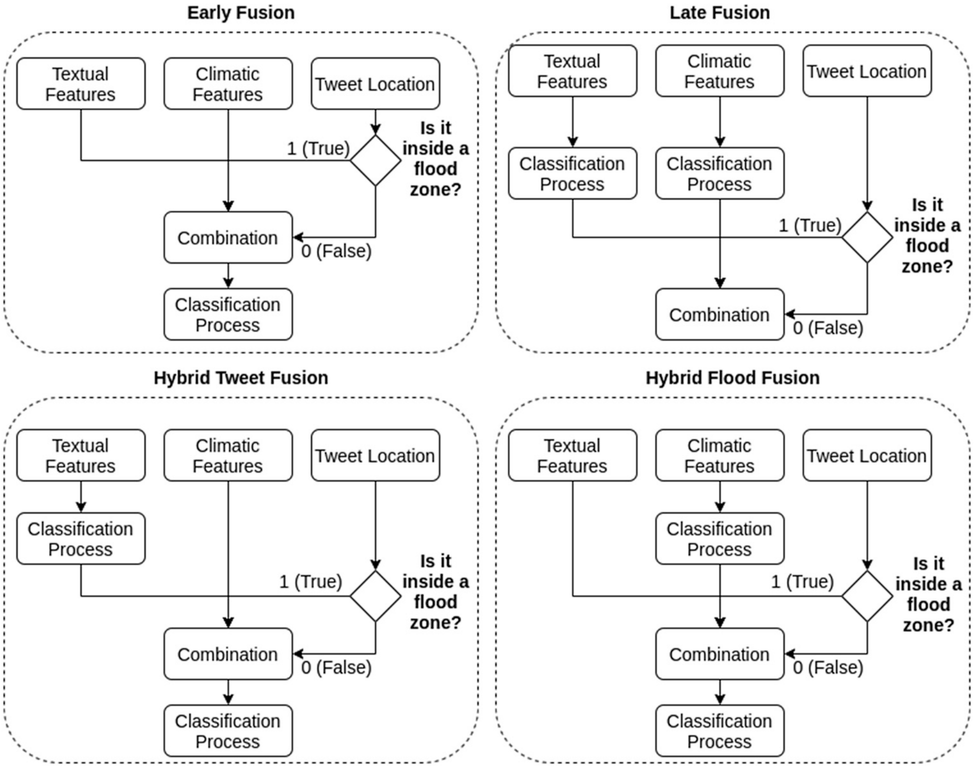

Fig. 6. Multimodal Fusion processes.

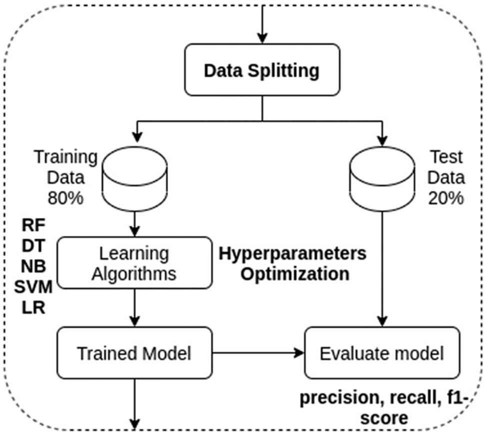

Fig. 7. Classification process.

flooded zones through the augmented matrix strategy (2). The Late Multimodal Fusion model labeled the events as positive according to the following conditions: the results of the flood forecasting mechanism show that the weather is prone to flooding, the results of the tweets classification model show that the content is related to “flooding” or “rain”, and the tweets are in the flooded areas of the city of São Paulo. Otherwise, the Late Multimodal Fusion model labeled the events as negative. Later the “Multimodal Late Fusion” model is evaluated in the validation dataset for precision, recall, and f1-score. It should be noted that this Multimodal Fusion strategy is adopted at this stage, because if the features are processed independently, more suitable ML techniques can be chosen for the datasets, and synchronization problems avoided (Lopes, 2015).

- **Hybrid Tweet Fusion:** in this subcomponent, we train a textual classification model. The results from the textual classification model are combined with the meteorological information and the results of the incidence of tweets in the flooded zones through the augmented matrix strategy (2). Following this, a classification model is trained with this combined heterogeneous data. Later the “Hybrid Tweet Fusion” model is evaluated in the validation dataset for precision, recall, and f1-score. What makes this approach superior to the others is that it provides the ML technique that is best suited to the textual dataset and produces a definitive classification model that generalizes the processed information from the tweets with the other data at the resource level.

- **Hybrid Flood Fusion:** in this subcomponent, a classification model is trained for meteorological data. The results of the flooding classification model are combined with the results of the incidence of tweets in the flooded zones and the pre-processed content of Twitter messages, through the augmented matrix strategy (2). After this, a classification model is trained with the combined heterogeneous data. Later the "Hybrid Flood Fusion" model is evaluated in the validation dataset for precision, recall, and f1-score. The reason why this approach is superior to the others is that it makes it possible to choose the most appropriate ML technique for the weather dataset and produces a definitive classification model that generalizes the processed weather data as well as the other resource-level information.

The **classification of all the methods** (Fig. 7) splits the "data training" into training (20%) and testing (80%). The training dataset is designed for the induction and adjustment of the model. In contrast, the test data simulate new samples for the prediction mechanism. The ML algorithms used in this training stage include the following: RF, DT, NB, SVM, and LR.

Additionally, in the classification stage, we employ the most widely used ML algorithms, together with different paradigms; these are based on works found in the literature that obtain SAW of emergencies through messages from social networks (e.g., de Bruijn et al. (2020), Huang and Xiao (2015), Li et al. (2018), Purohit et al. (2013)). When choosing the best ML parameters, it is necessary to adjust the hyper-parameters by means of optimization. Moreover, ML algorithms have parameters that require notification to be given to the user, so they can be adjusted to generate new models that can classify the samples with greater precision. Finally, when the parameter adjustment has been completed, the model undergoes a training and cross-validation process at the training base. The metrics for the performance evaluation are precision, recall, and f1-score.

4. Experimental setup

We conducted two experiments to obtain SAW of natural disaster through the messages published on Twitter and evaluate the value of the meteorological data for the Multimodal Fusion models. The computer settings used to run the experiments of this scientific research are: Operating System (Ubuntu 19.10); RAM Memory (8GB Single Channel DDR3 1600MHz); CPU (5th Generation Intel Core i5-5200U Processor); Graphics Card (NVIDIA(R) GeForce(R) 820M 2GB DDR3).

In the first experiment, we compared the performance of the clustering algorithms (DBSCAN, OPTICS, and Agglomerative Clustering) with the different linking criteria (e.g., Single, Complete, Average, and Ward) and examined the use of empirical and statistical strategies for the definition of the radius for the formation of clusters. The evaluation metric used for this is called the Silhouette Coefficient Rousseeuw (1987).

Table 5 shows the information of the dataset used in the experiment for the discovery of flooded areas, in which there are 1,433 unique geographical coordinates of flood events in the city of São Paulo. In the first experiment, all the clustering algorithms for the group formation of the empirical distance ranged from 0.1 to 5.0 km with an additional frequency of 0.1 km. The radius of the group formations varies within the maximum range of the intense rain cell size distribution in the region under study, and based on the strategy employed by Feng and Sester (2018). Thus, as the São Paulo city is located in a tropical region, according to Espejo (2016), the maximum size of an intense rain cell is 5 km. Furthermore, Scikit-Learn Pedregosa et al. (2011) was the ML library used in the first experiment. The other parameters required by the grouping algorithms used were in accordance with the standards of this library.

Table 5
Historical geographical coordinates of floods used in this experiment.

historical geographical coordinates of floods	11,798
historical geographical coordinates of unique floods	1,433
features	lat(float), long(float)
initial date	01/01/2015
final date	30/10/2018

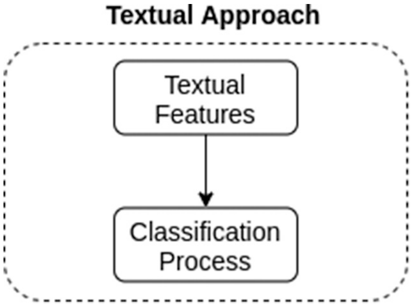

Fig. 8. Textual approach.

Table 6
Description of the database for the Multimodal Fusion and Textual models.

Data Models	Features	Amount of Data
Early Fusion	id (float), created_at (date), lat (float), long (float), text (string), inside_cluster (int), humidity (float), temperature (float), precipitation (float), atmospheric_pressure (float), dew_point_temperature (float), label (int)	910
Late Fusion — Climatic	id (float), created_at (date), lat (float), long (float), text (string), inside_cluster (int), humidity (float), temperature (float), precipitation (float), atmospheric_pressure (float), dew_point_temperature (float), label (int)	5,890
Late Fusion — Tweet	id (float), created_at (string), text (string), label (int)	5,890

In the second experiment, we carried out the training and testing required for the design of the following models contained in Fig. 6: "Early Fusion" and "Late Fusion" which are based on Atrey et al. (2010), as well as "Hybrid Flood Fusion" and "Hybrid Tweet Fusion" which are our own system. Furthermore, we compared the performance of these models with that of a text classification model in a validation database (as shown in Table 7). In addition, the evaluation metrics are precision, recall, and f1-score. Thus, the purpose of this experiment was to find the most precise model for obtaining SAW of flash floods to support Disaster Management. In addition, a comparison was made between a Multimodal fusion model and a textual classification model, to measure the effect of including contextual variables in Multimodal Fusion models, since this experiment was guided by an investigation carried out by de Bruijn et al. (2020), although it did not combine meteorological and geographical variables. Fig. 8 shows the textual approach adopted in this experiment to compare the effects of including contextual information based on Multimodal Fusion models.

The databases used in this experiment are described in Tables 6 and 7. The way these were set up is described in detail in Section 3.5, and

Table 7
Description of the validation database for Multimodal Fusion and Textual models.

Features	Amount of Data
id (float), text (string), humidity (float), temperature (float), precipitation (float), atmospheric_pressure (float), dew_point_temperature (float), is_alag (int), is_related (int), inside_cluster (int), label (int)	224

includes the heterogeneous data from November 1, 2016, to October 30, 2018, in the city of São Paulo.

All the processing in the textual stage of the multimodal model requires the use of the data cleaning algorithm (discussed in Section 3.4.1) to remove inconsistencies in the tweets. Later, we converted the symbolic characters to numerical ones through the techniques of BOW, TF–IDF, and Word Embeddings of the Fast Text type (50 and 100 dimensions) and Word2Vec (50 and 100 dimensions) Hartmann et al. (2017).

In the **"Early Fusion"** multimodal model, we trained the model in the "Early Fusion" dataset of Table 6 and evaluated the model in the dataset of 7. Even the generic classification used in this work follows the procedure outlined in Fig. 7. Moreover, in the **"Late Fusion"** multimodal model, we carried out the training for the flood detection mechanism in the "Late Fusion — Climatic" dataset of Table 6. The same procedure was carried out for the tweets of the classification model in the "Late Fusion — Tweet" dataset shown in Table 6. Both of these follow the classification process described in Fig. 7. As well as this, we evaluated the combination of heterogeneous data of the "Late Fusion" model in the dataset displayed in Table 7.

In the **"Hybrid Flood Fusion"** multimodal model, we first performed the training and testing of the flood identification model in the "Late Fusion — Climatic" database of Table 6 in accordance with the generic classification system (Fig. 7). Immediately after combining the predicted flooding values of the "Early Fusion" dataset with the respective textual data pre-processed from this database and geographical information, we created a new dataset for "Hybrid Flood Fusion". We trained a new model that was based on a generic classification pattern (Fig. 7) within the new "Hybrid Flood Fusion" dataset. We also evaluated, the combination of heterogeneous data in the dataset shown in Table 7.

As well as designing the **"Hybrid Tweet Fusion"** multimodal model, we first carried out the training and testing of the tweet classification model in the "Late Fusion — Tweet" dataset shown in Table 6 in accordance with the classification procedure outlined in Fig. 7. After this, we combined the predicted values for flooding of the "Early Fusion" dataset with the respective meteorological data pre-processed from this database and the geographical information. In addition, we created a new dataset for "Hybrid Tweet Fusion" and trained a new model that was based on a classification pattern (Fig. 7) within the new "Hybrid Tweet Fusion" dataset. We also evaluated the combination of heterogeneous data in the dataset shown in Table 7.

It should be noted that in all the models, the effects of the strategies adopted for the conversion of symbolic to numerical data are compared in terms of the precision, recall, and f1-score of the ML algorithms used in the models. We also used Scikit-Learn Pedregosa et al. (2011) and the ML library written in Python for these research experiments. The GridSearchCV[14] module was also found in the Scikit-Learn library Pedregosa et al. (2011) and this uses a full sample space of parameters and returns the parameters that provided the ML algorithms with the most precise classification.

Table 8
Matrix of confusion for the problem set out in this work, based on Faceli et al. (2011), where the columns show the predicted classes and the rows are the right classes.

	+	–
+	TP	FN
-	FP	TN

Finally, Table 7 shows that the **"is_related"** feature of this dataset results from manual classification (Section 3.3). The **"inside_cluster"** feature checks the incidence of tweets in the areas of flooding derived from the first experiment with the aid of the mathematical Haversine formula (Eq. (1)). The **"is_alag"** feature checks the occurrences of flooding in the period in which the meteorological sensors measured the climatic data, and CGE-SP reported the occurrences. Finally, the **"label"** feature is the result of manual fusion, as described in Section 3.5.

5. Evaluation metrics

The evaluation metrics of this study are widely used in the literature to classify messages published on Twitter (e.g., Ashktorab, Brown, Nandi, and Culotta (2014), Rosa et al. (2011)), grouping geographical data (e.g., Sparks, Thakur, Pasarkar, and Urban (2020)), and merging multimedia data (e.g., de Bruijn et al. (2020), Poria et al. (2016), Pouyanfar et al. (2019)). Thus, in this work, we use specific evaluation metrics for the experiments conducted, such as Silhouette metrics for evaluating the clustering of historical flood data, as well as precision, recall, and f1-score metrics for the generic classification systems and Multimodal Fusion.

The evaluation metric used to determine which clustering strategy has the best clustering performance is called Silhouette (Rousseeuw, 1987). In this approach, the closer to 1 the results are, the better formed the groups are (Sparks et al., 2020). At the same time, the closer to −1 the results are, the worse the formation of the groups is (Sparks et al., 2020). Moreover, the Silhouette values that are closer to 0 indicate the presence of an overlap in the groupings (Scikit-Learn, 2021).

The evaluation approaches used to determine which Multimodal Fusion model has the performance for combining the best media data are precision, recall, and the f1-score. These approaches even solve a two-class problem, about whether or not the tweets are related to flooding. Thus, for this type of work, one class is designated as positive (+) and another as negative (-) (Faceli et al., 2011).

According to the confusion matrix, which displays the possible answers to this problem and is shown in Table 8, TP (True Positive) represents the number of examples correctly classified as positive (Faceli et al., 2011). In contrast, TN (True Negative) represents the number of examples correctly classified as negative (Faceli et al., 2011). FN (False Negative) is the number of examples erroneously classified as negative because the right class is positive (Faceli et al., 2011). FP (False Positive) is the number of examples incorrectly classified as positive because the suitable class is negative (Faceli et al., 2011).

This means the performance metrics are derived from the matrix of confusion and responsible for the evaluation of models of Multimodal Fusion. According to Faceli et al. (2011) the precision corresponds to a set of truly positive examples classified as positive (Eq. (3)). On the other hand, the recall corresponds to the number of predicted hits in the positive class (Eq. (4)) (Faceli et al., 2011). In addition, according to Salas, Georgakis, and Petalas (2017) the result of the harmonic mean between the precision and the recall is called f1-score (Eq. (5)).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

[14] url: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

Table 9
Results of clustering procedures.

Ranking	Algorithm	Silhouette	Number of groups generated	Group formation distance (m)	formation distance type
1	Agglomerative Clustering (Average)	0.4788	271	900	empirical
9	Agglomerative Clustering (Single)	0.4692	5	3500	empirical
10	DBSCAN	0.4692	5	3500	empirical
54	Agglomerative Clustering (Complete)	0.4016	791	240	statistical
67	Agglomerative Clustering (Ward)	0.3839	873	5000	empirical
177	OPTICS	0.0931	103	3100	empirical

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

6. Results and discussion

In this section we describe the results obtained from the two exploratory scenarios of STIGMA.

6.1. Discover flooding areas

The first experiment conducted in this work was aimed at discovering flooded areas in the city of São Paulo on the basis of historical data of flooding from CGE-SP. This was because it is necessary to find which tweets are located in these areas for the Multimodal Fusion models and messages with content related to flooding tend to be located within flooded areas.

This involved employing a number of clustering algorithms (DBSCAN, OPTICS, Agglomerative Clustering), together with the linking criteria of Single, Complete, Average, and Ward type in the historical flood dataset. We compared the performance of these algorithms in terms of evaluating metrics for group formation called Silhouette (Rousseeuw, 1987). The parameter that defines the radius threshold for group formation in the clustering algorithms was defined empirically (ranging from 100 m to 5 km) and statistically (Semivariogram). Thus, 240 m was the value obtained from the Semivariogram function (Isaaks & Srivastava, 1989), where this result corresponds to the range of the function. The value of this range was calculated from the geographical coordinates of historical flood data (Table 5).

Table 9 shows the best performances of each clustering algorithm used in the descending order of Silhouette, the number of groups that each clustering approach produced, and the type of distance for group formation used by the unsupervised algorithms (empirical or statistical). This Table also shows the most significant results from each algorithm, the maximum radius of group formation, and the clustering quality of the formation (Silhouette). Additionally, we explored 286 clustering tests in this experiment.

Table 9 shows that the algorithm that obtained the best performance with Silhouette and benefited from an empirical strategy to define the maximum distance in group formation is Agglomerative Clustering with the Average type of binding criteria. In this approach, 271 groups were formed, and the value of Silhouette obtained is 0.4788. The algorithm that achieved the best performance with Silhouette and benefited from defining the maximum distance of the statistical group formation is Agglomerative Clustering with the Complete type binding criteria. In this approach, there were 791 groups, and the value of the Silhouette obtained is 0.4016. This meant that the first test in the ranking among all the tests carried out in this experiment had better-defined groups than the others because the value of Silhouette is closer to 1 than the others.

In January 2017, there were 324 occurrences of flooding in São Paulo city. In this same period, there were 1,640 geolocalized tweets containing keywords related to flooding, of which 1,231 tweets were manually labeled as related to flooding by the evaluators. Thus, the largest number of flood events in the city were in the central region, where there are neighborhoods such as Sé, República. Figs. 9 and 10 show a part of the central region of São Paulo in January 2017 with the areas of flooding and the geolocalized tweets. The flooding areas in Fig. 9 were generated by the Agglomerative Clustering algorithm and followed the criteria for linking groups of the Average type and the maximum formation radius of group formation defined as 900 m. The groupings in Fig. 10 were generated by the Agglomerative Clustering with the binding criteria of Complete type, and 240 m was defined as the maximum radius for group formation.

The groups shown in Fig. 9 are formed better than in Fig. 10 since the groups with a radius of 240 m overlap and are in an excessive number (791). On the other hand, the groups with a radius of 900 m do not overlap in the central region of São Paulo city and are lower (271) than the groups formed with a radius of 240 m. In addition, the tweets related to flooding tend to be located in flood-prone areas (in Fig. 9) to the detriment of the groupings in Fig. 10. Moreover, when designing the Multimodal Fusion models, it is necessary to determine the incidence of tweets in flooded areas of the city of São Paulo and then employ the groupings discovered in this experiment which have the best performance through Silhouette.

6.2. Comparison between the textual approach and multimodal fusion models

As described in Section 4, the second experiment conducted in this study focused on training the Multimodal models of “Early Fusion”, “Late Fusion”, “Hybrid Flood Fusion”, “Hybrid Tweet Fusion” and the unimodal textual model. It also made a comparison between the performance of these models in a validation database with the evaluation metrics defined in Section 5.

For this reason, we used data cleaning and data transformation strategies for the textual processing of the Multimodal Fusion models and unimodal classification model. In this experiment, we tested the effects of BOW, TF–IDF, Word Embeddings of Fast Text type, and Word2Vec with 50 and 100 dimensions on the performance of the models in terms of precision, recall, and f1-score.

As described in Section 4, in the case of the “Early Fusion” type model, the textual data, the meteorological data, and geographical information are combined at the resource level. Following this, they are used for the generic classification described in Fig. 7. The “Late Fusion” model also incorporates the textual data, the meteorological data, and geographic information at the decision level. In the “Hybrid Tweet Fusion” and “Hybrid Flood Fusion” type models, the first combines textual data at the decision level with the others at the resource level and then uses the generic classification system. The second combines the meteorological data at the decision level with the others at the resource level. It then creates a classification model with the resulting information. Finally, the unimodal classification model follows the generic classification described in Fig. 7 and only uses the texts of the tweets.

Table 10 shows the results of 10 cross-validation of the **Textual Approach** model in the dataset of training (Table 6). We carried out 50 tests and applied the RF, SVM, NB, DT, and LR algorithms to each type of conversion from symbolic to numerical characters, as described in Section 4. The results are in descending order of precision. The algorithm that obtained the best performance was the **DT** with **0.68937**

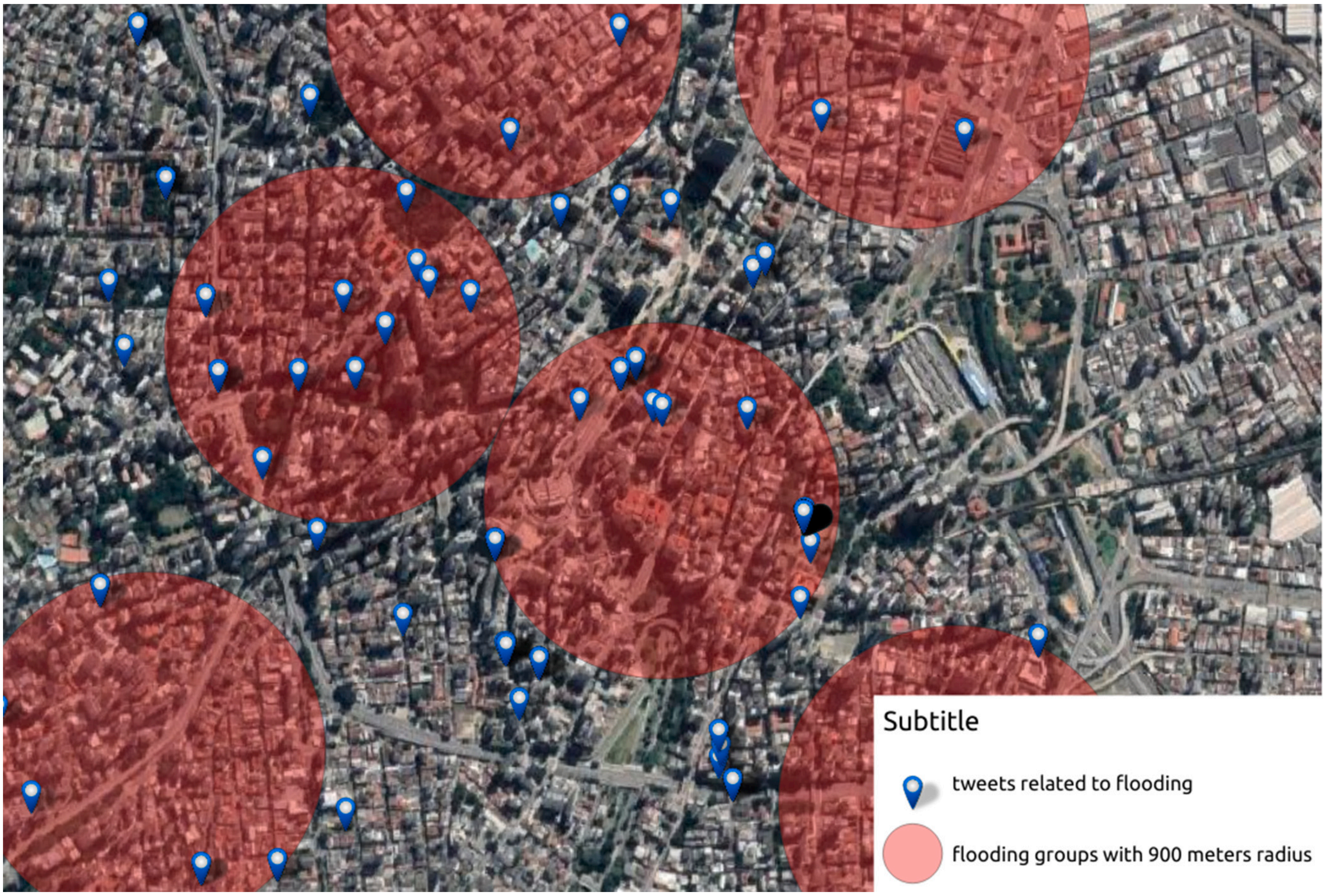
Subtitle
tweets related to flooding
flooding groups with 900 meters radius

Fig. 9. Flood groups with a radius of 900 m (empirical) and related geolocalized messages.

Table 10
Results of the 10 cross-validation techniques of the algorithms used in the Textual Classification System.

Algo	Transform Textual Data	Precision	Recall	F1-Score
DT	BOW	0.6893	0.6000	0.6375
RF	Fast Text type CBOW with 100 dimensions	0.6508	0.6500	0.6375
RF	Word2Vec type CBOW with 100 dimensions	0.6450	0.6277	0.6364
RF	FastText type CBOW with 50 dimensions	0.6442	0.6277	0.6443
RF	TF–IDF	0.6421	0.6333	0.6377

Table 11
Results of the 10 cross-validation techniques of the algorithms used in the Early Fusion stage.

Algo	Transform Textual Data	Precision	Recall	F1-Score
DT	BOW	0.8641	0.8555	0.8598
NB	Word2Vec type CBOW with 50 dimensions	0.8557	0.8388	0.8485
NB	Word2Vec type CBOW with 100 dimensions	0.8533	0.8323	0.8417
NB	Word2Vec type Skip-Gram with 100 dimensions	0.8501	0.8435	0.8448
DT	TF–IDF	0.8466	0.8267	0.8331

of precision and adopted the **BOW** type of data transformation strategy. It should be noted that since it included the data cleaning process, data transformation, and training of the ML algorithms, the operational time of the machine at this stage of the experiment was approximately **1 h and 17 min**.

Table 11 shows the results of 10 cross-validation techniques of the algorithms used in the **Early Fusion Multimodal** in the database of training (Table 6). We carried out 50 tests and applied the algorithms described in Section 4 to each type of conversion from symbolic to numerical characters (e.g., BOW, TF–IDF, among others). Thus, the algorithm that obtained the best performance is the one that occupies the first position in the ranking of the **DT** with **0.86415** of precision and with the data transformation strategy of the **BOW** type. The operational time of the machine, including the data cleaning process, data transformation, the combination of heterogeneous data, and the classification of training algorithms, was approximately **1 h and 9 min**.

Tables 12 and 13 show the results of the algorithms used in the "**Late Fusion**" approach which consist of the tweets classification model and flooding classification model. We also carried out 50 tests in the "Late Fusion — Textual" training database for the first model with the various ML algorithms and the data transformation types described in Section 4. According to Table 12, the best algorithm for this textual classification stage is **RF** with a precision rate of **0.80489**. The data transformation strategy for the tweets classification model is Word

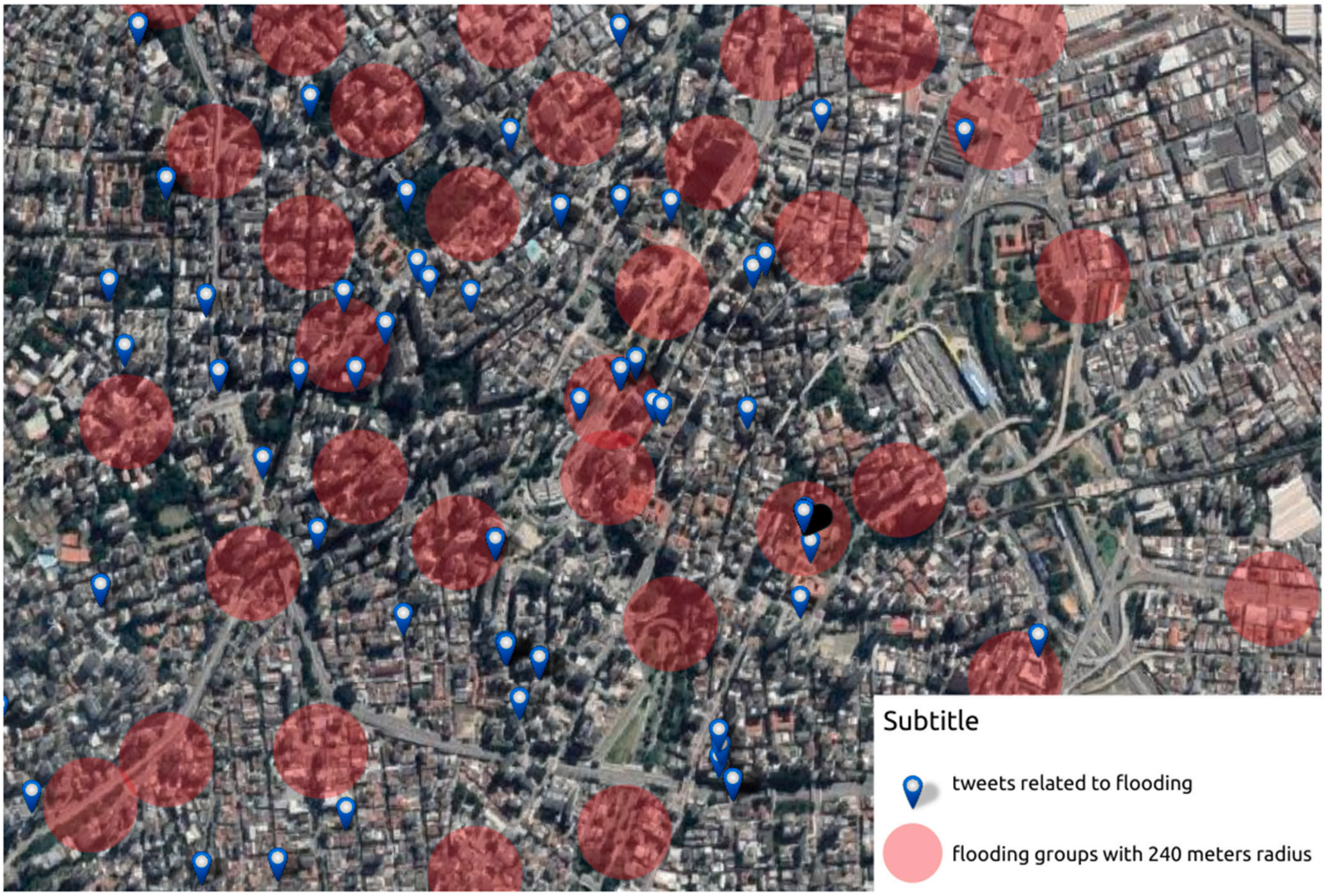
Subtitle
tweets related to flooding
flooding groups with 240 meters radius

Fig. 10. Flood groups with a radius of 240 m (statistical) and related geolocalized messages.

Table 12
Results of the 10 cross-validation techniques of the algorithms used in the tweets classification phase of the Late Fusion and Hybrid Tweet Fusion.

Algo	Transform Textual Data	Precision	Recall	F1-Score
RF	Fast Text type Skip-Gram with 100 dimensions	0.8048	0.8015	0.8036
RF	TF–IDF	0.7989	0.7926	0.7939
RF	BOW	0.7960	0.7926	0.7942
RF	Word2Vec type Skip-Gram with 100 dimensions	0.7951	0.7928	0.7939
SVM	BOW	0.7911	0.7761	0.7835

Table 13
Results of the 10 cross-validation techniques of the algorithms used in the flooding classification phase of the Late Fusion and Hybrid Flood Fusion.

Algo	Precision	Recall	F1-Score
DT	0.9310	0.9295	0.9302
RF	0.9173	0.9159	0.9166
SVM	0.7216	0.7071	0.7130
NB	0.7185	0.7011	0.7106
LR	0.6771	0.6646	0.6710

Embeddings of the **Fast Text Skip-Gram type with 100 dimensions** based on the Word Embeddings of NILC (Hartmann et al., 2017). In the second classification model, we conducted 5 tests in the “Late Fusion — Climatic” training database with the classification algorithms defined in Section 4. According to Table 13, the best algorithm in this stage was **DT** with a precision rate of **0.93102**. All the experiments of this phase lasted approximately **1 h and 38 min**, which was the processing time taken by the machine.

Table 14 shows the results of 10 cross-validation techniques of the algorithms used in the **“Hybrid Tweet Fusion”** in the database of training (Table 6). The “Hybrid Tweet Fusion” approach takes advantage of the results of the tweets classification mechanism that is trained in the design stage of the “Late Fusion” model, combining and training them with the contextual data from the “Early Fusion” training database. Thus in this experiment, we carried out ten-fold cross-validation tests in the “Early Fusion” dataset training with the aid of the classification algorithms defined in Section 4. According to Table 14, the best algorithm in this stage is **RF**, with a precision rate of **0.89806**. The operational time of the machine at this stage of the experiment was approximately **1 h and 13 min**.

Table 15 shows the results of 10 cross-validation techniques of the algorithms used in the **“Hybrid Flood Fusion”** in the database of training (Table 6). The Hybrid Flood Fusion approach takes advantage of the flooding classification mechanism trained at the design stage of the Late Fusion model and trains this model’s results with the contextual data. In light of this, we conducted an experiment involving 50 tests on the “Early Fusion” dataset with the classification algorithms and the numerical to symbolic data transformation approaches defined in Section 4. According to Table 15, the best algorithm in this stage is **DT** with a precision rate of **0.93403**. The data transformation strategy is Word Embeddings type **Word2Vec Skip-Gram with 100 dimensions** based on the Word Embeddings of NILC (Hartmann et al., 2017). All the

Table 14
Results of the 10 cross-validation techniques of the algorithms used in the Hybrid Tweet Fusion stage.

Algo	Transform Textual Data	Precision	Recall	F1-Score
RF	Fast Text type Skip-Gram with 100 dimensions	0.8980	0.8938	0.8959
DT	Fast Text type Skip-Gram with 100 dimensions	0.8783	0.8777	0.8780
NB	Fast Text type Skip-Gram with 100 dimensions	0.8724	0.8666	0.8684
LR	Fast Text type Skip-Gram with 100 dimensions	0.2363	0.4860	0.3180
SVM	Fast Text type Skip-Gram with 100 dimensions	0.2025	0.4499	0.2793

Table 15
Results of the 10 cross-validation techniques of the algorithms used during the Hybrid Flood Fusion stage.

Algo	Transform Textual Data	Precision	Recall	F1-Score
DT	Word2Vec type Skip-Gram with 100 dimensions	0.9340	0.9333	0.9336
LR	Word2Vec type CBOW with 100 dimensions	0.9302	0.9222	0.9262
DT	BOW	0.9283	0.9222	0.9252
LR	Fast Text type Skip-Gram with 50 dimensions	0.9278	0.9166	0.9219
LR	Fast Text type Skip-Gram with 100 dimensions	0.9264	0.9213	0.9240

Table 16
Results of the Validation of Multimodal and Textual Fusion Systems.

System	Precision	Recall	F1-Score	Execution Time (seconds)
Hybrid Flood Fusion	0.8769	0.8738	0.8736	0.234
Late Fusion	0.8654	0.8558	0.8549	0.945
Early Fusion	0.8249	0.8243	0.8242	0.170
Hybrid Tweet Fusion	0.8202	0.8198	0.8197	0.823
Textual	0.6488	0.6486	0.6485	0.196

experiments of this phase lasted approximately **1 h and 9 min**, which was the processing time taken by the machine.

Soon after we trained the Multimodal Fusion models and textual classification model, we compared the performance of these approaches in identifying the SAW of flash floods through tweets and contextual data in terms of precision, recall, and f1-score, in a validation database described in Table 7. It should be noted that in Table 16, the models are in descending order of precision. The model that obtained the best performance is the **Hybrid Flood Fusion** with a precision rate of **0.87693**.

Table 16 shows that the "Hybrid Flood Fusion" model performed better than the others. Since it merges heterogeneous features, the "Late Fusion" model oversimplifies intermodal dynamics (de Bruijn et al., 2020), and as a result, the training and execution processing times are too long. The "Early Fusion" model is also unsatisfactory since it causes problems with characteristic synchronization and overfitting (Lopes, 2015). In the case of the "Hybrid Tweet Fusion" model, the biases of the classification layer of messages published on Twitter are transferred to the merged layers, and thus the result of this hybrid merge technique tends to be inferior to the "Hybrid Flood Fusion" system. Hence "Hybrid Flood Fusion" is the best model because, in this process, the tweets are combined at the characteristic level with the results of the flooding classification mechanism and geographical information. In this way, the "Hybrid Flood Fusion" is able to exploit the correlation of the decision stage of the flooding classification model with the other heterogeneous data.

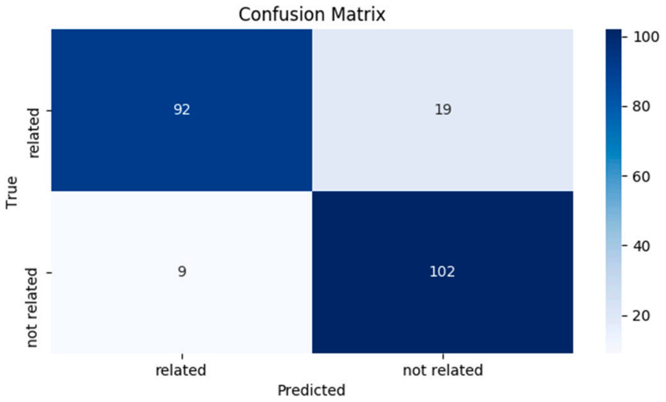

Fig. 11. Hybrid Flood Fusion confusion matrix.

Fig. 11 shows the confusion matrix of the "Hybrid Flood Fusion" model created from the labels of the validation dataset described in Table 7 and the values predicted by the Multimodal Fusion model when applied to this database. According to the confusion matrix, the model did not cause overfitting because it misclassifies some unpublished values, such as the identification of unrelated as related and related as unrelated.

Thus, the "Hybrid Flood Fusion" model is the most precise system among the compared multimodal fusion models because it did not cause overfitting. As well as this, the training and evaluation time is lower than the second rank shown in Table 16. The required execution time of the "Hybrid Flood Fusion" model by the machine in the validation dataset is four times less than is the case with the "Late Fusion" model because there are only two phases. The first model of this system can detect flooding, and the second is able to obtain the SAW of the flash floods through the tweets and contextual data. In comparison, the "Late Fusion" model has three phases: flood detection, classification of tweets, and combination of heterogeneous data.

The **inclusion of contextual variables** (meteorological data and geographic information) in the dataset of messages published on Twitter indicates a **22.8%** improvement in the **precision** rate of the Hybrid Multimodal Fusion model for detecting the SAW of flooding. This value is the difference between the precision of the "Hybrid Flood Fusion" and the "Textual Approach", where the former uses heterogeneous data. In comparison, the second approach only relies on the content of the tweets.

7. Conclusions

There are several works in the literature that adopt approaches aimed at obtaining a SAW of disasters from tweets. However, only a few of these make use of contextual features (weather data and geographic information) to improve the textual classification. This article demonstrates that by combining meteorological information with textual data and geographical information extracted from flooded areas, it is possible to make a significant increase in the precision required for identifying possible flood victims since models that rely only on textual features tend to perform worse than multimodal models. Furthermore, it was found that among the tested Multimodal Fusion models, the most precise model for obtaining the SAW of natural phenomena through tweets and contextual data in the validation dataset is "Hybrid Flood Fusion". This method combined the textual and geographic information with the results of the ML algorithm that adhered most closely to the meteorological data. For this reason, we decided to explore the correlation between the characteristic vectors of the messages published on Twitter with the processed geographic and meteorological information. This avoided overfitting and an excessive simplification of intermodal dynamics.

Moreover, in this study, the authors explored several strategies for grouping historical flood data and defining the maximum radius required for group formations. This meant that the hierarchical grouping algorithm used to determine flood-prone areas outperformed other clustering algorithms based on density. In addition, the maximum radius for defining groups that relied on empirical approaches outperformed statistical strategies (in our case, the Semivariogram technique). As Yin and Li (2001) points out, flooding is a natural phenomenon caused by human intervention in the environment, which makes it inappropriate to employ the Semivariogram technique.

Future work should concentrate on other strategies for defining the maximum distance for grouping formations (e.g., classical statistics and non-geostatistics) and different techniques for detecting flooded areas (e.g., neural networks). There is also a need for future studies to focus on identifying messages published on Twitter that are related to other natural disasters or other forms of Disaster Management (e.g., mitigation, preparedness, and recovery). We also recommend carrying out experiments that adopt other approaches for converting symbolic data into numeric data (e.g., DistilBERT Sanh, Debut, Chaumond, and Wolf (2019)). In addition, we also recommend adopting approaches based on CNN to obtain a SAW of the natural disasters from tweets and meteorological data and comparing the results of these studies with the Hybrid Multimodal Fusion strategies outlined in this study.

CRediT authorship contribution statement

Thiago Aparecido Gonçalves da Costa: Conceptualization, Methodology, Data curation, Software, Validation, Writing – original draft, Writing – review & editing. **Rodolfo Ipolito Meneguette:** Writing – original draft, Writing – review & editing. **Jó Ueyama:** Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank CAPES, Brazil (Coordination for the Improvement of Higher Education Personnel) for funding this research [concession number PROEX-10839814/M], the CEMEAI (Center for Mathematical Sciences Applied to Industry), and FAPESP (São Paulo State Research Support Foundation) [process number 2018/17335-9]. We are also grateful to the research group AGORA for providing the set of messages published on Twitter of the city of São Paulo from 11/2016 to 10/2018 that were essential for undertaking this study. Additionally, we are grateful for the hard work of Aysla C. G. Santos and Luan W. M. Alvares for helping us to label the 13,802 tweets manually. Finally, we would like to thank the ICMC-USP for providing the infrastructure and computational resources necessary for the planning and development of this study and carrying out the experiments.

Appendix. Computer code availability

To promote the reproducibility of the experiments carried out in this investigation, we present the codes developed in this research in the following Github link: https://github.com/thiagogcosta/multimodal-fusion-floods-tweets-meteorological. Thus, in this folder of the repository, it is possible to find the codes of the computational mechanism capable of assessing the level of agreement among the judges who manually classified the tweets (check it out on GitHub[15]) and the codes inherent in the pre-processing stages of the multimodal information (See it on GitHub[16],[17],[18]). Besides, this folder of the repository contains the codes written in Python of the mechanism for identifying flooding areas in the city of São Paulo (check it out on GitHub[19]). It is also possible to find the Python codes used to elaborate, train, and test the proposed Multimodal Fusion models (See it on GitHub[20],[21],[22],[23]). Furthermore, in this folder of the repository, the ground truth developed for this paper can be found, which was used to train and test the explored Multimodal Fusion models (check it out on GitHub[24]). Finally, due to the Twitter social network developer's contract policy, only the tweet identifiers can be found in the repository. Therefore, if you want to obtain the textual information of the messages of Twitter, it is necessary to use the Twitter Streaming API.

References

Aghi, R., Mehta, S., Chauhan, R., Chaudhary, S., & Bohra, N. (2015). A comprehensive comparison of SQL and mongodb databases. *International Journal of Scientific and Research Publications*, *5*(2), 1–3.

Alqhtani, S. M., Luo, S., & Regan, B. (2015). Multimedia data fusion for event detection in twitter by using dempster-shafer evidence theory. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, *9*(12), 2234–2238.

Andrade, S. C. d. (2020). *Mining of rainfall patterns from social media for supporting flood risk management* (Ph.D. thesis), Universidade de São Paulo.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. *ACM Sigmod Record*, *28*(2), 49–60.

Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response. In *ISCRAM* (pp. 269–272).

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, *16*(6), 345–379.

Avvenuti, M., Cresci, S., Del Vigna, F., Fagni, T., & Tesconi, M. (2018). CrisMap: a big data crisis mapping system based on damage detection and geoparsing. *Information Systems Frontiers*, *20*(5), 993–1011.

Avvenuti, M., Cresci, S., Marchetti, A., Meletti, C., & Tesconi, M. (2014). Ears (earthquake alert and report system) a real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1749–1758).

Baharin, S. S. K., Shibghatullah, A. S., & Othman, Z. (2009). Disaster management in malaysia: An application framework of integrated routing application for emergency response management system. In *2009 international conference of soft computing and pattern recognition* (pp. 716–719). IEEE.

[15] https://github.com/thiagogcosta/multimodal-fusion-floods-tweets-meteorological/tree/main/coefficient-of-agreement.

[16] https://github.com/thiagogcosta/multimodal-fusion-floods-tweets-meteorological/tree/main/textual-features.

[17] https://github.com/thiagogcosta/multimodal-fusion-floods-tweets-meteorological/tree/main/meteorological-features.

[18] https://github.com/thiagogcosta/multimodal-fusion-floods-tweets-meteorological/tree/main/flood-features.

[19] https://github.com/thiagogcosta/multimodal-fusion-floods-tweets-meteorological/tree/main/flood-features.

[20] https://github.com/thiagogcosta/multimodal-fusion-floods-tweets-meteorological/tree/main/train_test/early_fusion.

[21] https://github.com/thiagogcosta/multimodal-fusion-floods-tweets-meteorological/tree/main/train_test/late_fusion.

[22] https://github.com/thiagogcosta/multimodal-fusion-floods-tweets-meteorological/tree/main/train_test/hybrid_tweet_fusion.

[23] https://github.com/thiagogcosta/multimodal-fusion-floods-tweets-meteorological/tree/main/train_test/hybrid_flood_fusion.

[24] https://github.com/thiagogcosta/multimodal-fusion-floods-tweets-meteorological/tree/main/train_test/data.

Boettcher, A., & Lee, D. (2012). Eventradar: A real-time local event detection scheme using twitter stream. In *2012 IEEE international conference on green computing and communications* (pp. 358–367). IEEE.

Brouwer, T., Eilander, D., Van Loenen, A., Booij, M. J., Wijnberg, K. M., Verkade, J. S., et al. (2017). Probabilistic flood extent estimates from social media flood observations. *Natural Hazards & Earth System Sciences*, *17*(5).

Caragea, C., McNeese, N. J., Jaiswal, A. R., Traylor, G., Kim, H.-W., Mitra, P., et al. (2011). Classifying text messages for the haiti earthquake. In *ISCRAM*. Citeseer.

Cobo, A., Parra, D., & Navón, J. (2015). Identifying relevant messages in a twitter-based citizen channel for natural disaster situations. In *Proceedings of the 24th international conference on world wide web* (pp. 1189–1194).

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). # earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, *17*(1), 124–147.

Cutter, S. L., & Emrich, C. (2005). Are natural hazards and disaster losses in the US increasing? *EOS, Transactions American Geophysical Union*, *86*(41), 381–389.

de Aguiar, E. J., Faiçal, B. S., Ueyama, J., Silva, G. C., & Menolli, A. (2018). Análise de Sentimento em Redes Sociais para a Língua Portuguesa Utilizando Algoritmos de Classificação. In *Anais Do XXXVI Simpósio Brasileiro de Redes de Computadores E Sistemas DistribuíDos*. Porto Alegre, RS, Brasil: SBC, URL https://sol.sbc.org.br/index.php/sbrc/article/view/2430.

de Albuquerque, J. P., Herfort, B., Brenning, A., & Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, *29*(4), 667–689. http://dx.doi.org/10.1080/13658816.2014.996567, URL https://doi.org/10.1080/13658816.2014.996567.

de Andrade, S. C., Restrepo-Estrada, C., Delbem, A. C., Mendiondo, E. M., & de Albuquerque, J. a. P. (2017). Mining rainfall spatio-temporal patterns in Twitter: a temporal approach. In *The annual international conference on geographic information science* (pp. 19–37). Springer.

de Assis, L. F. F. G., de Albuquerque, J. a. P., Herfort, B., Steiger, E., & Horita, F. E. A. (2016). Geographical prioritization of social network messages in near real-time using sensor data streams: an application to floods. *Revista Brasileira de Cartografia*, *68*(6).

de Bruijn, J. A., de Moel, H., Jongman, B., Wagemaker, J., & Aerts, J. C. (2018). TAGGS: grouping tweets to improve global geoparsing for disaster response. *Journal of Geovisualization and Spatial Analysis*, *2*(1), 2.

de Bruijn, J. A., de Moel, H., Weerts, A. H., de Ruiter, M. C., Basar, E., Eilander, D., et al. (2020). Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network. *Computers & Geosciences*, Article 104485.

De Longueville, B., Smith, R. S., & Luraschi, G. (2009). "OMG, from here, I can see the flames!" a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 international workshop on location based social networks* (pp. 73–80).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Earle, P. S., Bowden, D. C., & Guy, M. (2012). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, *54*(6).

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the human factors society annual meeting, Vol. 32* (pp. 97–101). SAGE Publications Sage CA: Los Angeles, CA.

Espejo, T. M. S. (2016). *Interference due to rain in urban environments for millimeters waves* (Ph.D. thesis), Pontifícia Universidade Católica do Rio de Janeiro.

Faceli, K., Lorena, A. C., Gama, J. a., Carvalho, A., et al. (2011). Inteligência Artificial: Uma abordagem de aprendizado de máquina. *vol. 2*, (p. 192).

Feng, Y., & Sester, M. (2018). Extraction of pluvial flood relevant volunteered geographic information (VGI) by deep learning from user generated texts and photos. *ISPRS International Journal of Geo-Information*, *7*(2), 39. http://dx.doi.org/10.3390/ijgi7020039, URL https://doi.org/10.3390/ijgi7020039.

Haddad, E. A., & Teixeira, E. (2015). Economic impacts of natural disasters in megacities: The case of floods in São Paulo, Brazil. *Habitat International*, *45*, 106–113.

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., & Aluisio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. arXiv preprint arXiv:1708.06025.

Horita, F. E., de Albuquerque, J. a. P., Degrossi, L. C., Mendiondo, E. M., & Ueyama, J. (2015). Development of a spatial decision support system for flood risk management in Brazil that combines volunteered geographic information with wireless sensor networks. *Computers & Geosciences*, *80*, 84–94.

Huang, Q., & Xiao, Y. (2015). Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery. *ISPRS International Journal of Geo-Information*, *4*(3), 1549–1568.

IBGE (2010). Censo Demográfico.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). Extracting information nuggets from disaster-related messages in social media. In *Iscram*.

Isaaks, E. H., & Srivastava, R. M. (1989). *An introduction to applied geostatistics: Technical report*, Oxford university press.

Jalal, A. A. (2018). Big data and intelligent software systems. *International Journal of Knowledge-Based and Intelligent Engineering Systems*, *22*(3), 177–193.

Jan, T. G. (2020). Clustering of tweets: A novel approach to label the unlabelled tweets. In *Proceedings of ICRIC 2019* (pp. 671–685). Springer.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Klomp, J. (2016). Economic development and natural disasters: A satellite data analysis. *Global Environmental Change*, *36*, 67–88.

Kobiyama, M., Mendonça, M., Moreno, D. A., Marcelino, I., Marcelino, E. V., Gonçalves, E. F., et al. (2006). *Prevenção de desastres naturais: conceitos básicos*. Organic Trading Curitiba.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., et al. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances*, *2*(3), Article e1500779.

Kwak, H., Lee, C., Park, H., & Moon, S. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web* (pp. 591–600).

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Landis, J. R., & Koch, G. G. (1977). A one-way components of variance model for categorical data. *Biometrics*, 671–679.

Li, H., Caragea, D., Caragea, C., & Herndon, N. (2018). Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, *26*(1), 16–27.

Li, T., Xie, N., Zeng, C., Zhou, W., Zheng, L., Jiang, Y., et al. (2017). Data-driven techniques in disaster information management. *ACM Computing Surveys*, *50*(1), 1–45.

Liu, Y., Jiang, C., & Zhao, H. (2018). Using contextual features and multi-view ensemble learning in product defect identification from online discussion forums. *Decision Support Systems*, *105*, 1–12.

Lopes, B. L. (2015). *Detecção de cenas em segmentos semanticamente complexos*. (Ph.D. thesis), Universidade de São Paulo.

Lorini, V., Castillo, C., Dottori, F., Kalas, M., Nappo, D., & Salamon, P. (2019). Integrating social media into a pan-european flood awareness system: A multilingual approach. arXiv preprint arXiv:1904.10876.

Marcus, M., & Minc, H. (1992). *A survey of matrix theory and matrix inequalities, vol. 14*. Courier Corporation.

Matsubara, E. T., Martins, C. A., & Monard, M. C. (2003). *Pretext: Uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words: Techinical report, no. 4*, Universidade de São Paulo.

Meijering, E. (2002). A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proceedings of the IEEE*, *90*(3), 319–342.

Mendiondo, E. (2010). Reducing vulnerability ti water-related disasters in urban areas of the humid tropics. In *Integrated urban water management humid tropics, Paris, France* (pp. 109–127).

Norris, F. H., Stevens, S. P., Pfefferbaum, B., Wyche, K. F., & Pfefferbaum, R. L. (2008). Community resilience as a metaphor, theory, set of capacities, and strategy for disaster readiness. *American Journal of Community Psychology*, *41*(1–2), 127–150.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Poria, S., Cambria, E., Howard, N., Huang, G.-B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, *174*, 50–59.

Poser, K., & Dransch, D. (2010). Volunteered geographic information for disaster management with application to rapid flood damage estimation. *Geomatica*, *64*(1), 89–98.

Pouyanfar, S., Tao, Y., Tian, H., Chen, S.-C., & Shyu, M.-L. (2019). Multimodal deep learning based on multiple correspondence analysis for disaster management. *World Wide Web*, *22*(5), 1893–1911.

Purohit, H., Hampton, A., Shalin, V. L., Sheth, A. P., Flach, J., & Bhatt, S. (2013). What kind of# conversation is Twitter? Mining# psycholinguistic cues for emergency coordination. *Computers in Human Behavior*, *29*(6), 2438–2447.

Rosa, K. D., Shah, R., Lin, B., Gershman, A., & Frederking, R. (2011). Topical clustering of tweets. In *Proceedings of the ACM SIGIR: SWSM, Vol. 63*.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web* (pp. 851–860). ACM.

Salas, A., Georgakis, P., & Petalas, Y. (2017). Incident detection using data from social media. In *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)* (pp. 751–755). IEEE.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Scikit-Learn (2021). Sklearn.cluster. URL https://scikit-learn.org/stable/modules/clustering.html.

Shryock, H. S., Siegel, J. S., & Larmon, E. A. (1973). *The methods and materials of demography, Vol. 2*. US Bureau of the Census.

Sparks, K., Thakur, G., Pasarkar, A., & Urban, M. (2020). A global analysis of cities' geosocial temporal signatures for points of interest hours of operation. *International Journal of Geographical Information Science*, *34*(4), 759–776. http://dx.doi.org/10.1080/13658816.2019.1615069.

Spinsanti, L., & Ostermann, F. (2013). Automated geographic context analysis for volunteered information. *Applied Geography*, *43*, 36–44.

Trujillo, M., & Izquierdo, E. (2005). Combining k-means and semivariogram-based grid clustering. In *47th international symposium ELMAR, 2005* (pp. 9–12). IEEE.

Tucci, C. E., Hespanhol, I., & Netto, O. d. M. C. (2003). Cenários da gestão da água no Brasil: uma contribuição para a ''Visão Mundial da Água''. *Interações*, *1980*, 90.

Wang, S. I., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 90–94).

Win, S. S. M., & Aung, T. N. (2017). Target oriented tweets monitoring system during natural disasters. In *2017 IEEE/ACIS 16th international conference on computer and information science (ICIS)* (pp. 143–148). IEEE.

Winarno, E., Hadikurniawati, W., & Rosso, R. N. (2017). Location based service for presence system using haversine method. In *2017 international conference on innovative and creative information technology (ICITech)* (pp. 1–4). IEEE.

Xie, Z., & Guan, L. (2013). Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools. In *2013 IEEE international conference on multimedia and expo (ICME)* (pp. 1–6). IEEE.

Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012). Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, (6), 52–59.

Yin, H., & Li, C. (2001). Human impact on floods and flood disasters on the Yangtze River. *Geomorphology*, *41*(2–3), 105–109.

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv:1707.07250.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of tf* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758–2765.

Zhou, S., Leung, H., & Yao, F. (2013). Multimedia data fusion. *Mathematical Problems in Engineering*.