Towards Social Machine Learning for Natural Disasters

Jake Lever[1,2,3] and Rossella Arcucci[1,2,3]

[1] Department of Earth Science and Engineering, Imperial College London, Exhibition Rd, London, SW7 2BX, UK
[2] Data Science Institute, Imperial College London William Penney Laboratory, South Kensington, London, SW7 2AZ
[3] Leverhulme Centre for Wildfires, Environment and Society, Imperial College Department of Physics, South Kensington, London, SW7 2BW, UK

Abstract. We propose an approach for integrating social media data with physical data from satellites for the prediction of natural disasters. We show that this integration can improve accuracy in disaster management models, and propose a modular system for disaster instance and severity prediction using social media as a data source. The system is designed to be extensible to cover many disaster domains, social media platform streams, and machine learning methods. We additionally present a test case in the domain context of wildfires, using Twitter as a social data source and physical satellite data from the Global Fire Atlas. We show as a proof of concept for the system how this model can accurate predict wildfire attributes based on social media analysis, and also model social media sentiment dynamics over the course of the wildfire event. We outline how this system can be extended to cover wider disaster domains using different types of social media data as an input source, maximising the generalisability of the system.

Keywords: Natural Disasters · Machine Learning · Sentiment Analysis.

1 Introduction

As the climate begins to change, the severity and frequency of natural disasters is increasing yearly. Between 1998 and 2017, climate-related / geophysical disasters killed 1.3 million people, and left a further 4.4 billion injured, homeless, displaced or in need of emergency assistance [7]. In the last 30 years, the number of climate related disasters has tripled, and this increase is almost certain to continue into the future as climate records continue to be broken all over the world.

This increase in natural disaster activity due to climate change is additionally being compounded by human activity. Modern agricultural practices increase the risk of natural disaster through deforestation, and air pollution and the emissions of water soluble particles into the atmosphere also increase the risk of extreme weather. These disasters are also often highly linked, in ways which are still being studied; such as the recent arctic heatwave in 2020, which destabilized the polar

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62

vortex and allowed a cold front of air to move down over North America, causing sub-zero temperatures in Texas, freezing the power grid and leaving 210 people dead. Increasing world population and housing, water, food and health crises in many highly populated areas are forcing more people into living at the wildland urban interface, which is at a raised risk of disasters due to these reasons. As a result, more individuals & infrastructure will be increasingly exposed to more frequent and intense natural disasters, raising the overall potential cost to society of these destructive events. Since 1980, the US has sustained 310 weather events and climate disasters where the overall damages & costs reached or exceeded $1 billion USD (adjusted for CPI to 2021). 20 of these events occurred in 2021 alone, leading to the deaths of 688 people in this year. The total cost of these 310 events alone exceeds $2.155 trillion USD [17].

Advancements in computational models in the past 20 years have allowed humans to predict localised environmental conditions with increasing accuracy. This is due to increased computational resources, advancements in understanding of these systems, and the inclusion of better data sources. However, numerical computational models during natural disasters often suffer in terms of accuracy due to their extreme and unpredictable nature, meaning often not enough useful training data is available. One massive source of data which is increasingly being used in models is social media.

Social media plays an ever increasing role in society. In 2020, 23% of US adults reported getting their news from social media often [19], up from 18% in 2016 [19]. Additionally, 82% of US adults are using social media as of 2021 [6], representing a massive audience and huge amounts of shared information. This paper proposes a more simple, socially focused model which aims to avoid the common pitfalls of modern commonly used models by combining social media data as an input source. In the advent of social media in the last decade, more studies developing social and physical models are including these types of data sources as an input [22, 1, 9, 14]. These studies introduce the concept of the 'human sensor', where social media users are considered to be noisy sensors, posting a subjective account of their localised conditions. By analysing the response from these sensors, we can infer a model of the disaster as it unfolds in real time. We propose a system for collecting and analysing social media data to improve current natural disaster models, by monitoring online discussions and sentiments with the aim of ultimately identifying areas of actionable interest to disaster management teams. We show that online social discussions and sentiment are often linked to disaster activity, and that models can be trained to predict these shifts in sentiment over the course of the disaster. We also propose methods for information extraction & analysis of textual tweet data published during natural disasters, and discuss how this could be incorporated into a real-time model for public alerts.

The paper is structured as follows; Section 2 outlines the need for more socially conscious natural disaster models, and discusses the benefits of this. Following this, Section 3 defines the domain specific, modular architecture of the system we are proposing, describing the function of each of these modules. In Section 4

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62

we present a test case of the system in the context of north American wildfires from 2016. Finally, Section 5 summarises our contribution and further work.

2 Background

Extensive work has been put into climate modelling with increasing success, and managing and mitigating climactic effects has been achieved in the past with good results [13], such as flood mitigation with dams or drought management using reservoirs [16]. However, it has been shown that during natural disasters, the coordination of teams from a crisis management perspective becomes challenging [15]. Often emergency operations are given from a centre and coordinate multiple organisations including local government, police, fire, hospital, utility, and Red Cross representatives. These teams are often 'flown in' ad hoc, and are expected to collaborate to deliver optimal crisis management at often very short timescales. This can very easily lead to poor communication and coordination of disaster management teams at a time when quick, coordinated and effective action is often key to saving lives.

Inherent properties of social media have been shown to lend themselves towards crisis management during natural disasters from both sides of the public / disaster management coin [23]. On the one hand, social media allows people in different locations to post a subjective description of their surroundings & immediate dangers as a disaster unfolds, which again employs the concept of the human sensor. By in-taking and analysing this data during disasters, management teams could build up a geographic picture of these noisy accounts, and use predictions to quickly identify areas of interest/danger from public accounts posted in real-time.

Conversely, social media can also be effectively used by these management teams to relay operational messages, warnings, and updates back to the public once they have been authorized. An example of this has already been implemented by Google using satellite data for a number of crisis alerts including flood forecasting, wildfire boundary lines, earthquakes, and more [10]. Operational updates posted on social media by local authorities could also be used in models to coordinate information/response strategies between the different organisations, such as the ones mentioned previously, involved in crisis management. This could ease the communication strains between organisations and help coordinate a quicker, more direct response.

Social media data is increasingly being analysed using Natural Language Processing (NLP) methods. NLP is a set of broad analytical techniques for computationally interpreting human language [12]. The field has undergone rapid recent development, with increasingly more information such as topics, intent, themes, entities, and sentiments being able to be inferred by models from text. Due to the vast amounts of text information generated by social media posts, it has been shown that analysis of these posts show insights [22, 9, 1, 5, 2] into processes and events. This paper aims to yield similar insights in the context of natural disasters by performing information extraction & sentiment analy-

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62

sis (SA). APIs exist for accessing data from all main social media platforms, including Twitter [20], Facebook [8], Instagram [8], and Reddit [18].

By implementing a machine learning (ML) based information system such as the one described in this paper, we can improve the streamlining of information at emergency operations centres, hereby allowing disaster management teams to make real-time, bottom-up decision as an event unfolds. Previous work has shown that there is a link between social media expression/sentiment, and natural disaster activity in the context of hurricanes [22].

This paper proposes a system, shown in Fig 1, for collecting and analysing social media data in real time, and uses models trained on historical data to predict instances and characteristics of natural disasters. We show that gathering information on social media given the current state of the art of language models is feasible and efficient, and discuss the domain specific training using a test case in the context of wildfires for forecasting. We outline a system which adds a social aspect to disaster models, and contributes to the advancement of their capabilities by providing additional information to the overall system.

3 System Overview

The main concept behind the system is the live extraction of important information and sentiments from social media channels regarding natural disasters as they unfold. This system could be implemented initially in a predictive manner, predicting whether there is a natural disaster unfolding, and subsequently monitor and track the disaster by analysing social media discussions and extracting information from this live text data.

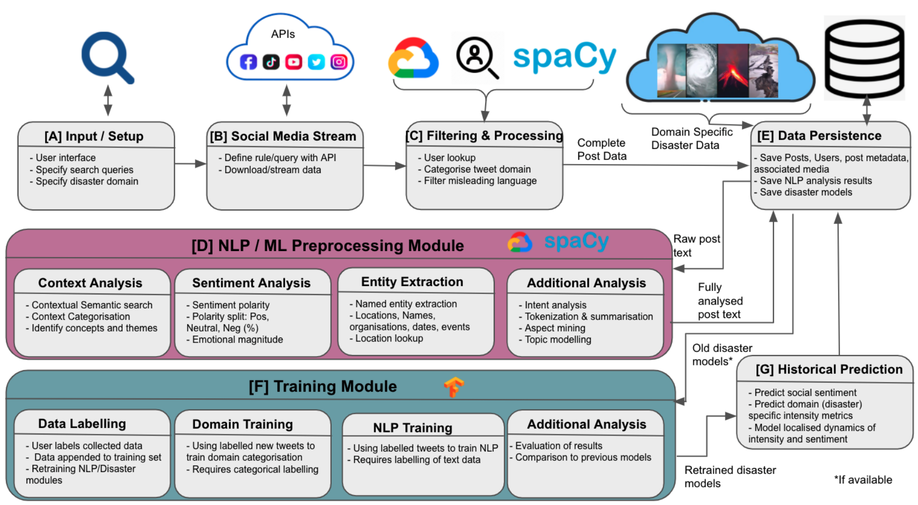


Fig. 1: Modular system diagram outlining flow of processes.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62

The system implements a live data stream based on search queries and using a prediction model which is specific to the disaster domain for the model. Here we mean the domain to be the category of natural disaster in the scope of this study, for example; Wildfires (this domain will be the demonstration case presented in Section 4), Earthquakes, Flood forecasting, or Hurricanes. With more work, the system could also further still be extended to include modelling and prediction on a wider range of types of events such as protests and political unrest, as well as political events and elections, health and food crises, and crime. This would require further development, primarily a more complex language model.

As mentioned, the prediction model would need to be domain specific to the type of disaster it is making predictions on. A model is made to be domain specific by training on a historical social media dataset which is again domain / disaster specific. For example, a model specific to the domain of wildfires would need to be trained on wildfire specific social media data. The entire system is designed to be modular and so can be adapted to be domain specific for each type of natural disaster, as well as different platforms of social media. This section will outline the systems component modules, and how they differ in terms of domains. The operational modular system is shown in Fig 1. Each modules function is now briefly explained.

3.1 [A] System Input / Setup

The system is designed to take a series of input queries, or rules on which to set up a live real time stream for social media posts which are published satisfying one of the given rules. This is shown in Module A of Fig 1. The rules must be formatted conforming to the social media platforms query syntax, e.g. Twitter's syntax rules [20]. The set of queries must be domain specific to the disaster, and include keywords specifically used in language specific to this domain.

3.2 [B] Social Media Data Streaming

This system implements real time social media streaming for the use of this live data in predictions. The function of this module is the implementation of a social media platform specific streaming function to download live data from the given website, as shown in Module B of Fig 1. A number of different APIs exist to facilitate this for Twitter [20], Facebook [8], Instagram [8], and Reddit [18]. The function sends a set of queries, outlined with the domain specific language in the syntax outlined by the individual APIs documentation, as outlined in Section 3.1. This starts the live data stream. Here, we can also search on accounts which are publishing posts, and part of the training phase for this system will be the creation of a list of reputable accounts to monitor. These accounts will have been shown to be useful in providing information about disaster instance and development. These may include local authority service accounts, local and environmental journalists, rescue workers, and local government and authorities.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62

3.3 [C] Filtering & Text Pre-processing

Following collection of the social media data, filtering of results removes false positives (posts which do not mention the natural disaster domain when the model detects that they have) from the collected data. Filtering by misuse/misleading phrases, e.g. 'Corruption is spreading like wildfire' makes the dataset cleaner, resulting in more accurate models. This is represented by Module C of Fig 1. Text is also pre-processed in this module, removing punctuation, hyperlinks, correcting spelling mistakes and formalising words etc., reducing noise in the text.

3.4 [D] Information Extraction & Sentimental Analysis

The function of this module is for the generation of social sentiment data from posts on natural disasters, and the extraction of published information which may be of help to the prediction module of the system. This is shown in Module D of Fig 1. Sentiment Analysis is the process of computationally extracting a numerical value corresponding to the overall emotional leaning of the text. This is achieved using analysis of words used, word patterns and part-of-speech (POS). This is useful because it allows us to convert the qualitative text data into quantitative metrics which can be used for further analysis. The system can then be trained to predict the aggregated sentimental values for each natural disaster in Section 3.7. SA was performed using Google's NLP API for SA [11], and yields two metrics:

- Firstly, the **Sentiment Score** ranges between -1.0 (negative) and 1.0 (positive) and corresponds to the overall emotional leaning of the text.
- Secondly, tweet **Magnitude** indicates the overall strength of emotion (both positive and negative) within the given text, between 0.0 and +inf. [11]

This defines two numerical sentimental variables; S and M, which are the social sentiment variables for each natural disaster, observing the constraints $S \in \mathbb{R} : 0 \leq S \leq 1$ and $M \in \mathbb{R} : 0 \leq M$ respectively.

A domain specific language model could be used to infer greater insight into the text data collected by the collection module outlined in Section 3.2. Models like this can be trained using domain specific language datasets, which can be the textual social media data previously collected as part of the system setup in Section 3.6. The NLP model will be similar to the one implemented for hurricanes in [22], and will additionally aim to extract domain specific disaster characteristics from the noisy text data. These may include but are not limited to; disaster duration, total area damage, and number of people displaced, in need of emergency assistance, injured or killed. This is achieved through the sentimental & linguistic analysis of the words used in the posts, for example; contextual, entity, and intent analysis, results of which are also saved to the database. Tokenisation of the text is also stored.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62

3.5 [E] Data Persistence

Data persistence and model iteration & integration are a key aspect of the evolutionary aspect of this system. Due to the systems online data collection module outlined in Section 3.2, data is constantly collected and analysed surrounding different types of natural disaster domains. The function of this module is to save analysed social media data, including metadata on posts, users etc and from Section 3.2, NLP and sentimental analysis results from Section 3.4. Additionally stored will be the models trained in Section 3.6 and the downloaded natural disaster data, as shown in Module E of Fig 1.

3.6 [F] System Training (Pre-Operational)

Before the system can be used in real time for operational use, a model needs to be trained in order to make the model domain specific for the type of natural disaster it will be used for. Each type of disaster discussed in Section 3 will have different types of expression on social media and the dynamics and relationship between social media activity and disaster activity will vary from disaster to disaster. Thus, it is necessary to implement individual models for each type of disaster.

To train a domain specific model, we need a similarly domain specific historical social media dataset related to this type of disaster. The system is set up using historical natural disaster data for the domain specific disaster data in order to implement a base model.

This module also supports the labelling of training data gathered from the data collection and analysis modules of the system. The reason for this is for the iterative improvement of the ML models in Modules D & G, described in Sections 3.4 & 3.7 respectively.

An important aspect of this training phase is that it can be achieved offline. That is, models can be iteratively trained while the system is online, and swapped out as they are improved. Due to the system constantly collecting and saving tweet data surrounding these types of natural disasters, the database used to train these models will only get larger and more diverse, both geographically and between disaster domains, as the system is used more. This new data can be labelled and then used to iteratively train improved systems in a supervised manner, or remain unlabelled for models to be trained using unsupervised methods. To summarise, the system is designed to have an offline training programme which iteratively retrains and improves the model as it runs, leading to evolutionary improvement and in turn greater accuracy, and the rapid development & deployment of improved predictive models.

3.7 [G] ML Prediction Model

This part of the system utilises the model trained historical data outlined in Section 3.6. We take the model trained on historical, domain specific disaster data, and use this to make predictions about the occurrences of new disaster instances. The aims of this module are as follows;

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62

- To detect instances and categorise types of new disasters, and predict the times and locations of these new events.
- To predict the severity of disasters over the course of the crisis period. By severity we mean the physical domain specific disaster variables from the historical data supplied in the training dataset outlined in Section 3.6, e.g. size of area affected, number of injuries etc. We hypothesise that there exists some function of f which allows us to predict disaster intensity from social media post data.

The ML method which we utilise for this system is the Gradient Boosted Random Forest, implemented in python with the *XGBoost* package [21]. Random Forests are an ensemble method extension of Classification and Regression Trees (CART) [4]. In the training phase, these methods start with a simple model, often a single tree (or weak learner), and then additive training occurs where trees are generated and added to build up a forest of trees which is used for the final prediction. For new data, predictions are then made by averaging the majority vote of all trees in the forest. The GBRFs are built and evaluated using the MAE, and the Gini coefficient [4]. The Gini coefficient is a scoring metric which is a measure of the degree to which a particular element is wrongly classified when randomly chosen and it is expressed by the formula:

$$\sum_{i=1}^{n} p_i(1-p_i) \tag{1}$$

where p_i denotes that the physical event i happens (or that the sentiment $i \in S$ is evaluated) and n is the number of possible events (or possible sentiments). Gradient boosting Random Forest algorithms are particular Random Forests which begin with a base (weak) learner (tree), and consecutively add more weak learners to the ensemble with the goal of minimising a loss function [4]. Given a training sample $\{\mathbf{x}_i, \mathbf{y}_i\}_1^n$, the goal is to find function of $F'(\mathbf{x})$ such that the expected value of the loss function $\psi(\mathbf{y}, F(\mathbf{x}))$ is minimised [4]. The gradient boosting problem is then expressed as:

$$F'(\mathbf{x}) = \arg\min_{F(\mathbf{x})} \mathbb{E}_{\mathbf{x},\mathbf{y}} \psi(\mathbf{y}, F(\mathbf{x})) \tag{2}$$

where $\mathbb{E}$ denotes the expected value. The boosting algorithm then approximates $F'(\mathbf{x})$ by additive expansion, which can be summarised as:

$$F(\mathbf{x}) = \sum_{m=0}^{s} \beta_m h(\mathbf{x}, a_m) \tag{3}$$

where the functions $h(\mathbf{x}, a)$ are known as the ‘weak learners’.

Over a series of s steps, the weak learners are sequentially added, and the expansion coefficients $a = \{a_1, a_2, ...\}$ and $\{\beta\}_0^s$ are jointly fit to the current models pseudo-residuals. For $m = 1, 2, \ldots, s$, this gives

$$(\beta_m, a_m) = \arg\min_{\beta,a} \sum_{i=1}^{n} \psi(\mathbf{y}_i, F_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i; a)) \tag{4}$$

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62

and

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \beta_m h(\mathbf{x}; a_m) \tag{5}$$

which shows the step-wise optimisation of F . Equation (4) is then optimised to solve the loss function ψ by fitting least squares on $h(\mathbf{x}, a)$. This replaces the optimisation problem presented in Equation (2) with one based on reducing least squares in Equation (4).
The models were trained on a wide grid search covering the hyper-parameters denoted by:

- ETA (Learning Rate): Step size shrinkage used in update.
- $maxDepth$: Maximum depth of a tree.
- $minChildWeight$: Minimum sum of instance weight needed in a child.
- $subsample$: Subsample ratio of the training instances.
- $colSampleByTree$: Subsample ratio of columns when constructing each tree.
- $nEstimators$: Number of trees in the ensemble.

The output of this module is the overall system output: predictions of domain, instances and severity of the natural disaster. This represents a socialised model of the disaster, from detection to evolution and finally resolution. Once a disaster has been detected and registered in the application database, it follows the process flow shown in Fig 1, being continuously updated as more information on this suspected disaster is streamed through by the data collection stream. A disaster is monitored until information is passed through that this crisis is resolved. We now demonstrate in Section 4 a prototype of the model described in Section 3, applied with a wildfire domain.

4 Experimental Test Case: Wildfires - Satellites & Twitter

We now present a prototype version of the domain specific model applied to satellite wildfire data using twitter as the social media data source. We choose wildfire events occurring in Australia and North America (United States and Canada) in 2016 for the creation of our training dataset. We chose to use this type of disaster domain in this geographic area due to the abundance of both wildfire activity in this area and active Twitter users. We now outline the implementation of the system for this particular context.

Satellite & Twitter Data Collection: For the combination of social media data with wildfire data in our model component of the system, we chose to use twitter for the historical social media source, and wildfire data from the Global Fire Atlas [3] for the 2016 wildfire data. The physical wildfire characteristics taken from this data are: Latitude ($\circ$), Longitude ($\circ$), Size (km2), Perimeter (Per) (km), Duration (d) (days), Speed(km/day), Expansion (Exp) (km2/day), and Start (S_{DOY}) and End (E_{DOY}) . This defines in the physical vector $\mathbf{x}$ of our model:

$$\mathbf{x} = [Lat, Lon, Size, Per, d, Speed, Exp, S_{DOY}, E_{DOY}, Pop_{Density}] \tag{6}$$

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62

Twitter's V2 API [20] with an academic research product track was used for the collection of the twitter data associated with historical wildfires, using a query implementation in line with Twitter's query syntax [20]. Queries were designed to search for tweets mentioning locations of the burn as well as a set of wildfire domain specific keywords to search on; Fires OR Wildfires OR Bushfires OR "Landscape Burn" OR "Wildland Burn", as well as generated hashtags included in the query. Tweets which contained certain misuse phrases such as "like wildfire" were removed to reduce noise. Tweets were saved to a database with meta and user data and media, along with a Fire ID. The result is a dataset of US wildfires in 2016 and tweets associated with these events.

Twitter Data Analysis: We are now able to analyse this text data for wildfires and generate social sentiment variables from this data using SA. Recalling Section 3.4 where the two numerical sentimental variables S and M are defined, Tweets are grouped by day for each fire and social sentiment values from the S and M are averaged and summed to generate the following social sentiment variables for each wildfire:

- S_{mean} : Average Daily Sentiment Score
- M_{mean} : Average Daily Magnitude Score
- S_{ovr}: Overall Sentiment Score
- M_{ovr}: Overall Magnitude Score
- Tot_{tweets}: Total number of Tweets for Wildfire

These variables constitute the sentimental vector

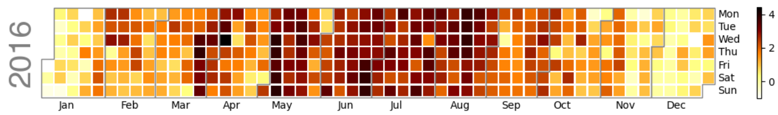
$$\mathbf{y} = [S_{mean}, M_{mean}, S_{ovr}, M_{ovr}, Tot_{tweets}]. \tag{7}$$

The generation of these social sentiment variables for each wildfire represents the completion of our two datasets in both the Australian (AUS) and North American (US) domains.

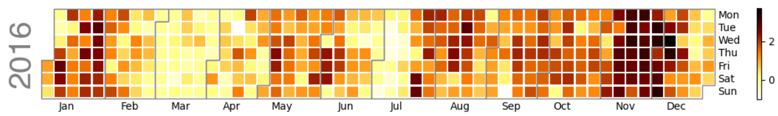
The data can now be viewed from a temporal perspective, by plotting heatmaps for the online social sentiment across the year of 2016. Fig 2 shows how online sentimental activity matches the fire seasons in the two geographic domains, which is a positive indication of the quality of the data.

Wildfire & Social Modelling Results: Two types of ML models were implemented using the *XGBoost* package [21] in python as outlined in Section 3.7. The first type of model takes as an input the physical vector $\mathbf{x}$, and is trained using this data on the target vector $\mathbf{y}$. The 5 variables which were described in this section were used to train 5 models, one predicting each social sentiment variable. These models are called the sentimental prediction models. The second type of model predicted the 10 physical wildfire characteristics vector $\mathbf{x}$, using the social sentiment vector $\mathbf{y}$, predicting in the opposite direction to the sentimental prediction model. This type of model was called the physics prediction model, and there were 10 of these models trained, one for each variable, meaning total of 15 models were trained on the combined AUS + US dataset. The results are

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62



(a) Heatmap of US online social sentiment for 2016



(b) Heatmap of AUS online social sentiment for 2016

Fig. 2: US (a) and AUS (b) online social Sentiment heatmaps for 2016

shown in Tables 1 & 2 below. As the last column in Tables 1 & 2 shows, the execution time (Exe_T) for running the model on a 2020 Macbook Pro 2.3GHz 8 core i7 Intel processor is very low, whch is an important condition for real time operational predictions.

Table 1: Results from predicting social sentiment variables from physical wildfire characteristics.

Variable	MAE	RMSE	Gini	$Gini_N$	Score	Exe_T (msecs)
S_{mean}	6.34	17.83	0.367	0.846	38.7%	2.94
M_{mean}	10.88	40.85	0.316	0.822	38.3%	5.75
M_{ovr}	107.27	407.36	0.338	0.872	24.4%	3.37
S_{ovr}	54.75	364.76	0.381	0.865	15.3%	2.28
tot_{tweets}	458.927	2777.28	0.323	0.863	10.6%	2.94

Table 1 demonstrates that the models for predicting average Sentiment and Magnitude (S_{mean} and M_{mean}) show good results when attempting to predict these variables, with low MAEs of 6-10. Overall Sentiment and Magnitude (S_{ovr} and M_{ovr}) models also preformed well. The most notable result shown in Table 2 is the model predicting fire Duration (d). This model showed very positive results, with an MAE of 0.84. This shows that the resulting model was able to predict wildfire duration to within one day from social sentiment values/social media data alone. Additionally, *Speed* and *Exp* both performed well.

5 Conclusion and future work

This paper outlines a system to facilitate the integration of social media data into physical models, discusses the benefit of this, and demonstrates a prototype test

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62

Table 2: Results from predicting physical variables from social sentiment data.

Variable	MAE	RMSE	Gini	$Gini_N$	Score	Exe_T (msecs)
Lat	3.72	5.2	0.344	0.868	0.75	4.90
Lon	6.185	7.40	1.73	0.929	0.79	6.11
$Size$	9.66	33.89	0.354	0.835	0.14	1.95
Per	7.22	15.89	0.227	0.764	0.24	3.69
d	0.84	2.01	0.207	0.961	0.87	4.22
$Speed$	0.533	0.86	0.143	0.659	0.25	2.32
Exp	0.88	3.24	0.226	0.652	0.16	2.09
$Pop_{Density}$	69.05	575.18	0.397	0.424	0.02	2.22
S_{DOY}	40.73	60.90	0.100	0.736	0.56	2.83
E_{DOY}	39.62	61.15	0.095	0.734	0.56	5.48

case with current wildfire models. As shown, social media data is being adopted increasingly in scientific studies and specifically disaster management, yielding benefits which could be transferred to natural disaster relief efforts. This work attempts to bridge this gap by developing a modular social media alarm system for natural disasters which predicts instances and localised severity of the event based on analysis of social media posts.

Having successfully trained and implemented a retrospective historical social media wildfire model using Twitter as a data source, the next step for testing this type of model would be the integration / coupling of the system with a real time wildfire model. The modular design of the system allows for rapid updating of the system as future work allows. This will primarily be focused on the development of ML methods for information extraction of the post text data and modelling disaster activity. The benefit of this will be improved accuracy of these models which will allow localised modelling of disaster conditions.

Social media data represents a near limitless supply of real time data on almost any large event. If disaster models do not consider this data, then this represents a loss of information, as there is data available via these networks which is not accessible elsewhere. Systems which analyse this data in real time are able to recoup this loss, which ultimately leads to the development of improved models.

References

1. Alkouz, B., Aghbari, Z.A., Abawajy, J.H.: Tweetluenza: Predicting flu trends from twitter data. Big Data Mining and Analytics **2**(4), 273–287 (2019). https://doi.org/10.26599/BDMA.2019.9020012
2. Alrashdi, R., O'Keefe, S.: Automatic Labeling of Tweets for Crisis Response Using Distant Supervision, p. 418–425. Association for Computing Machinery, New York, NY, USA (2020), https://doi.org/10.1145/3366424.3383757
3. Andela, N., Morton, D.C., Giglio, L., Paugam, R., Chen, Y., Hantson, S., van der Werf, G.R., Randerson, J.T.: The global fire atlas of

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62

individual fire size, duration, speed and direction. Earth System Science Data **11**(2), 529–552 (2019). https://doi.org/10.5194/essd-11-529-2019, https://essd.copernicus.org/articles/11/529/2019/
4. Bishop, C.M.: Pattern recognition and machine learning (2006)
5. Cui, R., Gallino, S., Moreno, A., Zhang, D.J.: The operational value of social media information. Production and Operations Management **27**(10), 1749–1769 (2018). https://doi.org/https://doi.org/10.1111/poms.12707, https://onlinelibrary.wiley.com/doi/abs/10.1111/poms.12707
6. Department, S.R.: Social media usage in u.s. (November 2021), https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/
7. for Research on the Epidemiology of Disasters & The UN Office for Disaster Risk Reduction, T.C.: Economic Losses,Poverty & Disasters 1998-2017. Tech. rep., The Centre for Research on the Epidemiology of Disasters & The UN Office for Disaster Risk Reduction (2018)
8. Facebook: Facebook graph api, https://developers.facebook.com/docs/graph-api/
9. Gallagher, R.J., Reagan, A.J., Danforth, C.M., Dodds, P.S.: Divergent discourse between protests and counter-protests: #blacklivesmatter and #alllivesmatter. PLOS ONE **13**, 1–23 (04 2018). https://doi.org/10.1371/journal.pone.0195644, https://doi.org/10.1371/journal.pone.0195644
10. Google: Forecasting & alerts: Google's crisis alerts provide access to trusted safety information across search, maps, and android. https://crisisresponse.google/forecasting-and-alerts/ (2017), accessed: 2021-01-17
11. Google: Analyzing sentiment (2021), https://cloud.google.com/natural-language/docs/analyzing-sentiment
12. Hirschberg, J., Manning, C.D.: Advances in natural language processing. Science **349**(6245), 261–266 (2015). https://doi.org/10.1126/science.aaa8685, https://www.science.org/doi/abs/10.1126/science.aaa8685
13. Job C. M. Dullaart, Sanne Muis, N.B..J.C.J.H.A.: Advancing global storm surge modelling using the new era5 climate reanalysis. Climate Dynamics **54**(1), 1007–1021 (2020). https://doi.org/10.1007/s00382-019-05044-0, https://doi.org/10.1007/s00382-019-05044-0
14. Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., Cebrian, M.: Rapid assessment of disaster damage using social media activity. Science Advances **2**(3) (2016). https://doi.org/10.1126/sciadv.1500779, https://advances.sciencemag.org/content/2/3/e1500779
15. Laura G. Militello, Emily S. Patterson, L.B..R.W.: Information flow during crisis management: challenges to coordination in the emergency operations center. Cognition, Technology & Work **9**(1), 25–31 (2007). https://doi.org/10.1007/s10111-006-0059-3, https://doi.org/10.1007/s10111-006-0059-3
16. Lempérière, F.: Dams and floods. Engineering **3**(1), 144–149 (2017)
17. National Centers for Environmental Information: U.s. billion-dollar weather and climate disasters, 1980 - present (ncei accession 0209268). https://www.ncdc.noaa.gov/billions/overview (2021), accessed: 2021-01-17
18. Reddit: Reddit api documentation, https://www.reddit.com/dev/api/
19. Shearer, E.: 86% of americans get news online from smartphone, computer or tablet (Jan 2021), https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/
20. Twitter: Twitter api v2, https://developer.twitter.com/en/docs/twitter-api

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62

21. XGBoost: Dart booster (2020), https://xgboost.readthedocs.io/en/latest/tutorials/dart.html
22. Yao, F., Wang, Y.: Domain-specific sentiment analysis for tweets during hurricanes (dssa-h): A domain-adversarial neural-network-based approach. Computers, Environment and Urban Systems **83**, 101522 (2020). https://doi.org/https://doi.org/10.1016/j.compenvurbsys.2020.101522, https://www.sciencedirect.com/science/article/pii/S0198971520302556
23. Young, C., Kuligowski, E., Pradhan, A.: A review of social media use during disaster response and recovery phases (2020-01-31 2020). https://doi.org/https://doi.org/10.6028/NIST.TN.2086

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_62