

**Medical Severity Assessment of Natural Disasters**

**Table of Contents**

[***Business Understanding***](#page2)[***1***](#page2)

[***Project objectives and scope***](#page3)[***1***](#page3)

[***Data understanding***](#page4)[***2***](#page4)

[***Methodology***](#page5)[***3***](#page5)

[***Modelling Results and Findings***](#page10)[***8***](#page10)

[***Results evaluation***](#page12)[***10***](#page12)

[***Evaluation of Project delivery outcomes***](#page13)[***11***](#page13)

[***Ethical Consideration***](#page14)[***12***](#page14)

[***Limitations***](#page15)[***13***](#page15)

[***Key improvements for future steps***](#page15)[***13***](#page15)

[***Project progress timeline with milestones achieved***](#page16)[***14***](#page16)

[***References***](#page17)[***15***](#page17)

[***Appendix***](#page18)[***16***](#page18)

**Business Understanding**

Mortality rates of Natural disasters have significantly declined over the years from millions of deaths to an average of 60,000 per year. Different natural disasters often include different types of events such as starvation, malnutrition, fire outbreak, medical emergencies, infrastructure damage. These events share a commonality in that they all impact society in a negative way. With the rise of technology and contingency practices, the designated stakeholders responsible at maintaining the outcomes of a natural disaster have manage to limit the significant impact. Despite leveraging advanced data science techniques such as Machine Learning, Deep learning, and Natural Language Processing (NLP), the stakeholders of Natural disasters are unable to consistently reduce the cost accumulated as a result, with year-on-year changes and fluctuations. (See appendix 1) The main problem faced by majority of the stakeholders surrounds the severity assessment of events from a natural disaster. There are two major issues facing severity assessment of natural disaster: Firstly, there is no gold standard universal scale that can assess different types of events equally; secondly, the overlap between events and stakeholders makes it difficult to associate disaster events to a particular stakeholder.

**Project objectives and scope**

Given the broad aspect of severity assessment, this report will focus on the following project objectives within the specified scope:

1. Developing a severity assessment guideline for Medical related events during a Natural disaster
2. Creation of a new feature from noisy textual data based on severity assessment achieved.
3. Building a Machine Learning Model that can accurately classify medical severe events from none-severe medical events.

Severity assessment changes across different industries and is entirely based on the field, hence this report will narrow the scope of the NLP project to the medical sector involved with emergency responses to medical. Events during a Natural disaster.

1

**Data understanding**

This section of the report focuses on discussing about the data used for this NLP project, mainly discussing the techniques, findings and insights derived from data acquisition, data cleaning, data exploration.

1. Data Acquisition

Most disaster messages acquired through methods like web-scraping on twitter are relatively noisy, including multiple types of data and special characters that are insignificant for machine learning models. Hence, this project acquired a dataset from the Hugging face platform called “[**disaster\_response\_messages**](https://huggingface.co/datasets/disaster_response_messages)”.

2. Data cleaning

From a preliminary inspection conducted by the researching team, this dataset contains

30,000 messages drawn from Natural disaster events in multiple language. The dataset was

translated, and new column was created with the translation. The creators of the dataset

encoded the data with 36 different categories related to disaster response and stripped the

messages of any sensitive information in their entirely. For this research, the column

“**medical\_help**” was of importance to the scope. The column was filtered of all events that

related to medical help, eliminated all other 35 features. Furthermore, the columns for “**ID**”

were pruned as the values were all distinct and useless for any Machine learning model only

contributing additional noise. A new dataset was created called “**Medical\_severity\_dataset”**

and a new feature name was prepared called “**Medical\_severity**”, ready to perform

annotations on.

3. Data description summary

After the data cleaning implementation, we have resulted in a dataset that contains **2084** rows of medical related events from the natural disasters. Hence demonstrating a massive skew of events towards other types of events during a natural disaster such as infrastructure damage, transportation issues and fire related events.

2

**Methodology**

This section of the report overlooks the different NLP based methods utilized by the researchers to achieve the objectives throughout the NLP pipeline project. It will review the methodologies’ use case of the method, limitations, and suggested modifications. Moreover, The NLP section of the methodologies will review the data exploration after any transformation has been conducted by the researchers.

1. NLP methods

NLP is a vast and relatively mature aspect of Artificial Intelligence, hence the methodologies that are used for processing textual data tend to be redundant. In this research, common methods of feature extraction were utilized including TF-IDF, Bag-of-words and Word vectorization. Moreover, other common NLP techniques were deployed for analysis of the disaster messages in the dataset including word frequency distribution, n-grams, and word cloud visualization. These powerful text extractors and analysers were deployed after a crucial step of severity annotations were implemented by the researchers.

This following section will demonstrate a breakdown of the steps from severity annotations to a full complete textual pre-processing technique implementation:

* Step 1: Severity Annotations

The research focuses on assessing the severity of medical events found in natural disasters from the dataset. However, the dataset does not contain any label that can assess the severity of the events, hence the researchers opted to create the annotations. A severity annotation guideline was prepared by the researchers using a severity assessment scale that is currently used in Australia for clinical purposes and therefore adapted and modified for natural disasters alongside the guidelines used for natural disaster severity assessment. (See appendix

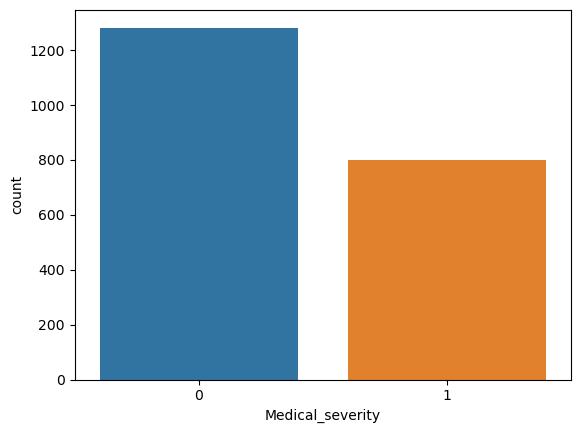
1. Once the guidelines were developed the following steps were taken to achieve fair, unbiased, and accurate results:
2. 4 annotators were used to increase the level of human judgement.
3. A validation set of the same rows was tested (200 rows validated and a 96% similarity was achieved)

3

* Step 2: Data exploration

Given the less significant amount of data cleaning required for the NLP project, the researchers decided on conducted an exploration of the data after annotations were made, given the significant change of the data structure and scope.

**Figure 1: The frequency of annotations created by the researchers.**



The annotations reflected that majority of the medical events expressed in the dataset expressed no severity (Label=0) but not so significant as to deem the annotation as a class imbalance as there was a significant amount of severity events (Label=1).

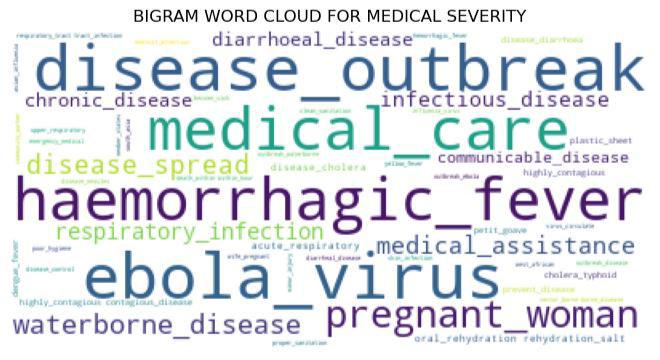
After annotations were made, the researchers noticed that there could be additional noise in the dataset that could contribute to failure of the machine learning process. Hence, the n-grams technique was used to determine whether topic words were commonly associated with words that were considered unimportant to the scope or translated words that do not have any relevance. The researchers found significant words that do not contribute to the scope such as “United Nations”, “International”, “South Africa”. These words were mainly proper nouns either reflected popular NGOs which is understandable as this is a disaster dataset or countries in which the natural disaster occurred in. Therefore, these words were pruned by adding them to the stop words list. (See appendix 3 and 4)

4

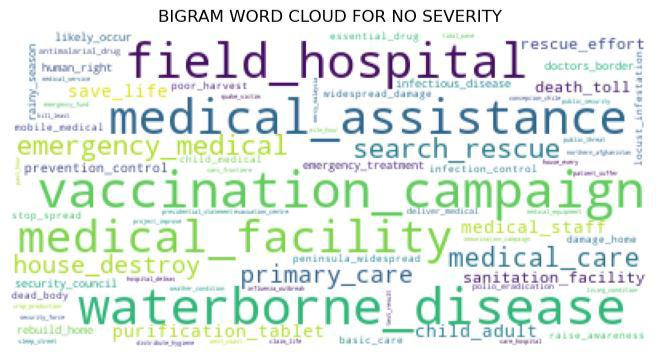
**Overlapping issues**

During the annotation process, it was determined by the researchers that a word overlapping issue may arise between the class 0 and 1. Hence, a validation process for refining the annotation was added by visualizing the word clouds of the bigrams from both class 0 and 1 of medical severity. After multiple trials of refining this process and re-validating the annotation methods the significant bigrams between each class were unique which will contribute to the model being able to distinguish between both classes despite there being some overlap which is unavoidable such as events are classified 0 containing similar words with events classified as 1, due to contextual differences.

**Figure 2: The bigram word cloud for the positive class**



**Figure 2: The bigram word cloud for the negative class**



5

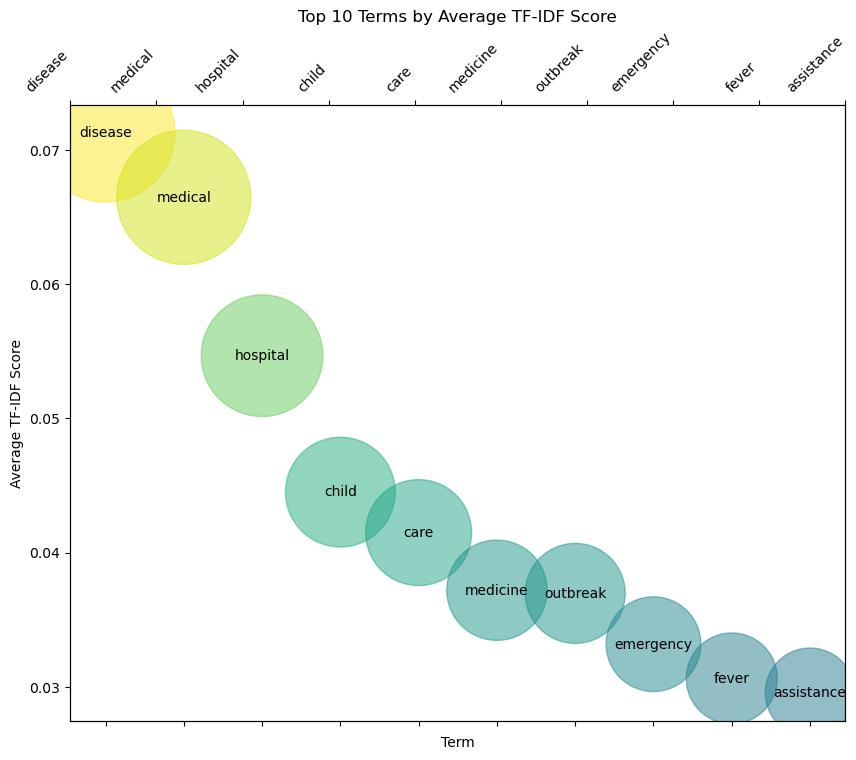
* Step 3: Data preparation

Before the Machine Learning process is conducted, the features are supposed to be determined and given the only feature used by the researchers is the disaster message, appropriate NLP techniques are supposed to be applied to extract features from the text.

**TF-IDF**

The TF-IDF method is the most appropriate method of vectorizing and extracting features from the text as any semantic issues were pre-resolved and considered by the researchers during the data exploration phase. Moreover, with the shape and state of the dataset pre-determined as single words carrying more weight than word associations, the limitations of TF-IDF can be bypassed. After weighting this against the user-friendly and easy to use principles of TF-IDF, this method was selected for vectoring and feature extraction.

**Figure 4: Demonstrates the TF-IDF scores for the top terms.**



6

2. Machine Learning Models

While NLP is still a mature field of AI, Machine learning has been around for a while. We have been equipped by packages in the python language that can run machine learning models efficiently and effectively. However, it is important to choose these models appropriately as each have their own use case. Given our scope is to build a predictive model separating classes we will use models that are designed for a classification issue and more specifically, a **binary classification.**

**Testing Models**

The models used for testing in this research are listed below:

1. Logistic Regression
2. RandomForest Classifier
3. DesisionTree Classifier
4. Multinomial NB
5. Support Vector Machines (SVM)

**Testing techniques**

As mentioned in the data description section, there is lesser sample that has been extracted for testing hence we have to deploy techniques to ensure that the machine learning section is carried out effectively.

1. Splitting ratio

The dataset was split by a ratio of 0.75:0.25 as this was determined the best for increasing the number of test data available for final testing.

2. Cross-validation

A technique known as **k-fold** was adopted for validating the training results before feeding the algorithm to the unseen testing dataset. This allows for validating across multiple splits and determining which model presents the bets result.

3. Hyperparameter tuning

There are several hyperparameters that differ according to the model that can be adjusted to improve the results achieve or fix any issues of overfitting or underfitting.

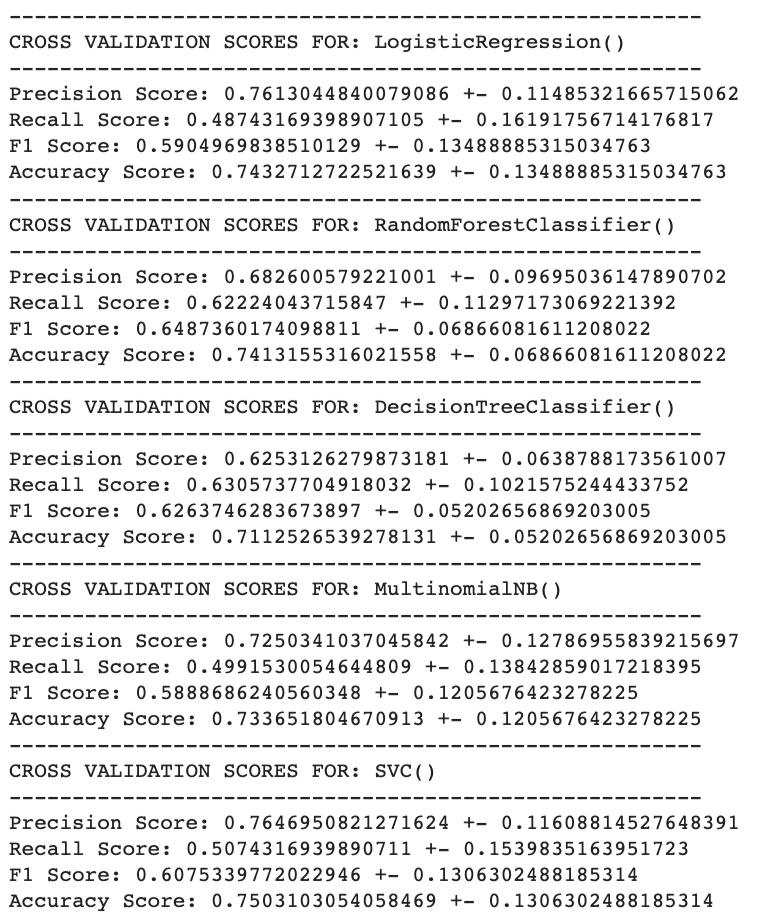
7

**Modelling Results and Findings**

This section demonstrates the results achieved by testing the different models and the insights derived from those results. For classification-based problems, it is important to determine the correct performance metrics for evaluating the results. The researchers determined that given there is a relative balance of class that it is safe to use accuracy as a measure of score.

Moreover, other metrics were used to determine the level of overfitting, underfitting or misclassifications including precision, recall, F1-measure, and confusion matrix.

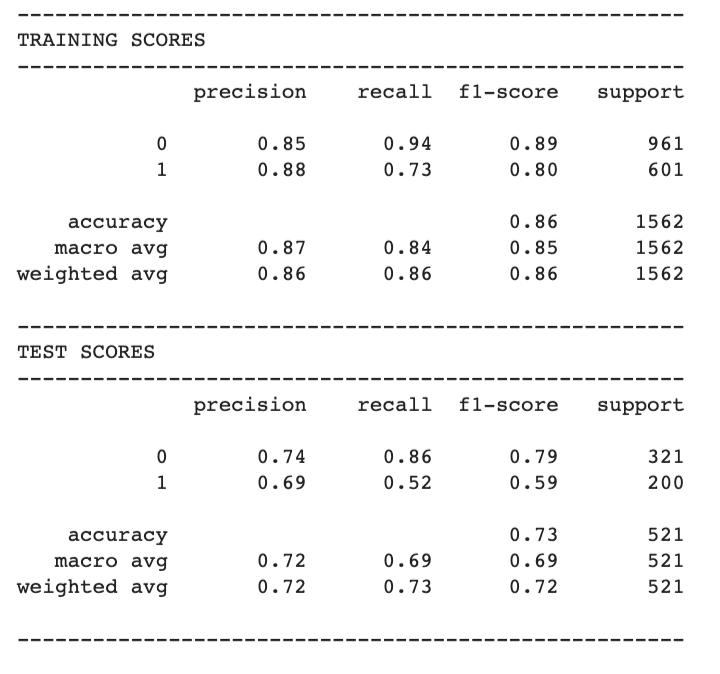
**Figure 5: Cross-validation results**



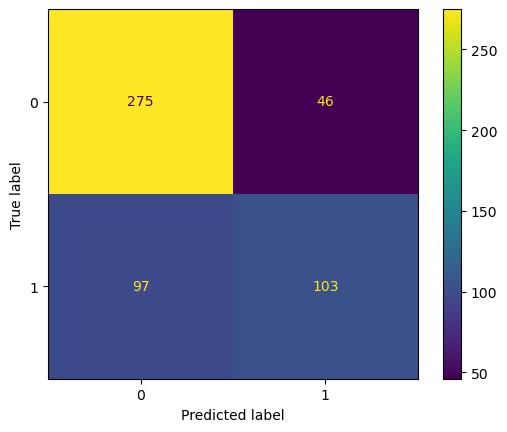
8

**Final testing**

**Figure 6: Final test results of the Random Forest Classifier**



**Figure 7: Confusion Matrix of the test results**



9

**Results evaluation**

There are several figures presented above, this section will breakdown the importance of these figures as they display the results of the model testing and proceed onto determining and evaluating the significance of the model to the scope of the study.

As presented from figure 5, the results from the cross validation show each performance metric results and the potential for fluctuations. This aided in determining which model would perform the best and it was concluded by the researchers that overall, the RandomForest Classifier presented the best results and least likely to vary from each corresponding result.

Figure 6 showcases the results achieved between the test and the training sets. This was to display if there was a chance the model was overfitting towards the training. The results show that the RandomForest classifier achieved a 73% accuracy score on the test data, which underperforms as compared to the training, but does not show any signs of overfitting. The performance decrease could be a result of the sample size being relatively small. While the result achieved is relatively good, the sample size does make it difficult to attribute any generalisation to the results, hence it cannot be generalised to any dataset and is specifically tailored to the dataset used in this experiment.

Another key metric to spotlight is the 52% recall achieved on the testing set, showcasing a relatively low recall by the model. Thus, the model completeness of the true positives (severe events) is being misclassified significantly as (non-severe events). This is further backed by figure 7 as we can see from the confusion matrix displayed that there is a minor difference between what was accurately classified as severe events and what was misclassified as a non-severe event but was a sever event. From a Machine learning perspective, this does potential ramifications as the numbers do not indicate a perfect model, but they are not unexplainable within the scope of the study.

10

**Evaluation of Project delivery outcomes**

In a nutshell, the project aimed to achieve the following objectives:

1. Developing a severity assessment guideline for Medical related events during a Natural disaster
2. Creation of a new feature from noisy textual data based on severity assessment achieved.
3. Building a Machine Learning Model that can accurately classify medical severe events from none-severe medical events.

The researchers successfully developed a severity annotation guideline based on medical severity assessment currently used in the field within Australia. The annotation scale was adapted for a binary classification, which eliminated any potential for a vague scale, creating an environment of a two-pole spectrum. Moreover, the method was adapted for a completely NLP basis whereby the features used included disaster messages from people in the form of text. This was conducted by using several NLP techniques outlined in appendix (). Extracting features from noisy data achieved after annotations was successful through the NLP techniques adopted by the researchers including TF-IDF, Word cloud visuals and N-grams. While the pre-processing and data-preparation were achieved with relative success, building a predictive model that can be adapted into an application used by medical professionals when determining the severity of an event in a natural disaster for operational improvements is not so straightforward. Despite successfully building a predictive model, the lack of any deployment application is shared in the limitations and future steps section of the report.

11

**Value Added**

As mentioned previously, there is an uncertainty around severity, its definition and how its assessment is taken place for natural disaster, let alone a specific stakeholders’ interests within a natural disaster like the medical field. With the annotations achieved, it is determined that a severity assessment purely based on NLP has been created without any combination of numerical values or statistics on the natural disaster events, opening the doors to fields that deal mostly with textual data but without a common scale for severity.

**Ethical Consideration**

The dataset explored a limited amount of potential for ethical ramifications, potential scoping within data privacy, data sensitivity and accountability. But the potential use case of a model used to predict the severity does exhibit some significance in terms of morality in determining the severity of a medical event.

* **Data privacy**

The creators of the dataset eliminated any form of private data found which improves any issue surrounding the use of the dataset.

* **Data Sensitivity**

The dataset contains disaster message that are sensitive to world issues including murder, poverty and death and the handling of such data must be conducted in a respectful and considerate manner.

* **Morality and Accountability**

The application of this study is in the medical field, by subjectively determining who is worthy of medical attention and who is not, this raising concerns about the future use of case of this study and the morality of its application deployment.

12

**Limitations**

This section outlines the key limitations of this study that have significant impact towards the scope while also outlining the future steps to be taken for progress.

Key limitations:

* Shortage of data samples

The dataset sample is significantly low to generalise to a whole population of people and exhibits some form of sample bias as it included natural disasters from specific places.

* Annotation research limitation

The annotations guidelines were built and developed based on limited existing research and multiple adaptations based on existing scales within medical field which are used directly in clinical practices but may not apply to natural disaster response.

**Key improvements for future steps**

There are some key improvements that can be considered for future application or replication of this research paper. These following are the key improvements:

* Revisiting the annotations by cross validating with a larger group
* Consulting an expert in the medical field for further improvement on the validation layer
* Adapting a technique to combined multiple types of data with textual data such as natural disaster statistics for severity assessment such as existing scientific scale to determine the extent of a natural disaster event.

13

**Project progress timeline with milestones achieved**

This section gives an overview of the NLP project through the NLP pipeline process, demonstrating the timeline of the project.

**Figure 8: The timeline of the NLP project**

The timeline demonstrates each milestone achieved during the NLP project and outlines the complete pipeline.



Reshaping



of dataset

with



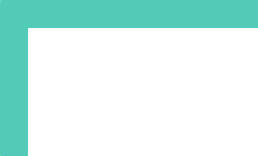
annotations



Model

testing

Developing



Annotation

Guidelines

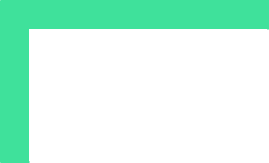


Final testing

Data



exploration



Evaluation

and

Derviation of

insights

Data

acqusition



Completion of Report

14

**References**

This section documents the referencing of the research paper in Harvard Style formatting.

community, T.H.D. (no date) Disaster\_response\_messages · datasets at hugging face, disaster\_response\_messages · Datasets at Hugging Face. Available at: https://huggingface.co/datasets/disaster\_response\_messages (Accessed: 15 May 2023).

Pontiki, M. et al. (2014) ‘Semeval-2014 task 4: Aspect based sentiment analysis’, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) [Preprint]. doi:10.3115/v1/s14-2004.

Caldera, H.J. and Wirasinghe, S.C. (2022) A universal severity classification for natural disasters, Natural hazards (Dordrecht, Netherlands). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8630994/ (Accessed: 15 May 2023).

Zong, S. et al. (2019) Analyzing the perceived severity of cybersecurity threats reported on social media, arXiv.org. Available at: https://arxiv.org/abs/1902.10680 (Accessed: 15 May 2023).

Severity assessment code (SAC) 1 clinical incidents (no date) WA Health, Government of Western Australia. Available at: https://www.health.wa.gov.au/Articles/S\_T/Severity-assessment-code-SAC-1-clinical-incidents (Accessed: 15 May 2023).

JACOBSSON, R. (no date) *Extraction of severity information from clinical narratives using*

*...* Available at: https://publications.lib.chalmers.se/records/fulltext/246132/246132.pdf

(Accessed: 14 May 2023).

Ritchie, H., Rosado, P. and Roser, M. (2022) *Natural disasters*, *Our World in Data*.

Available at: https://ourworldindata.org/natural-disasters (Accessed: 15 May 2023).

Rudden, J. (2023) *Cost of natural disaster losses worldwide 2000-2022*, *Statista*. Available

at: https://www.statista.com/statistics/612561/natural-disaster-losses-cost-worldwide-

by-type-of-

loss/#:~:text=In%202022%2C%20the%20estimated%20economic,of%20181%20billio

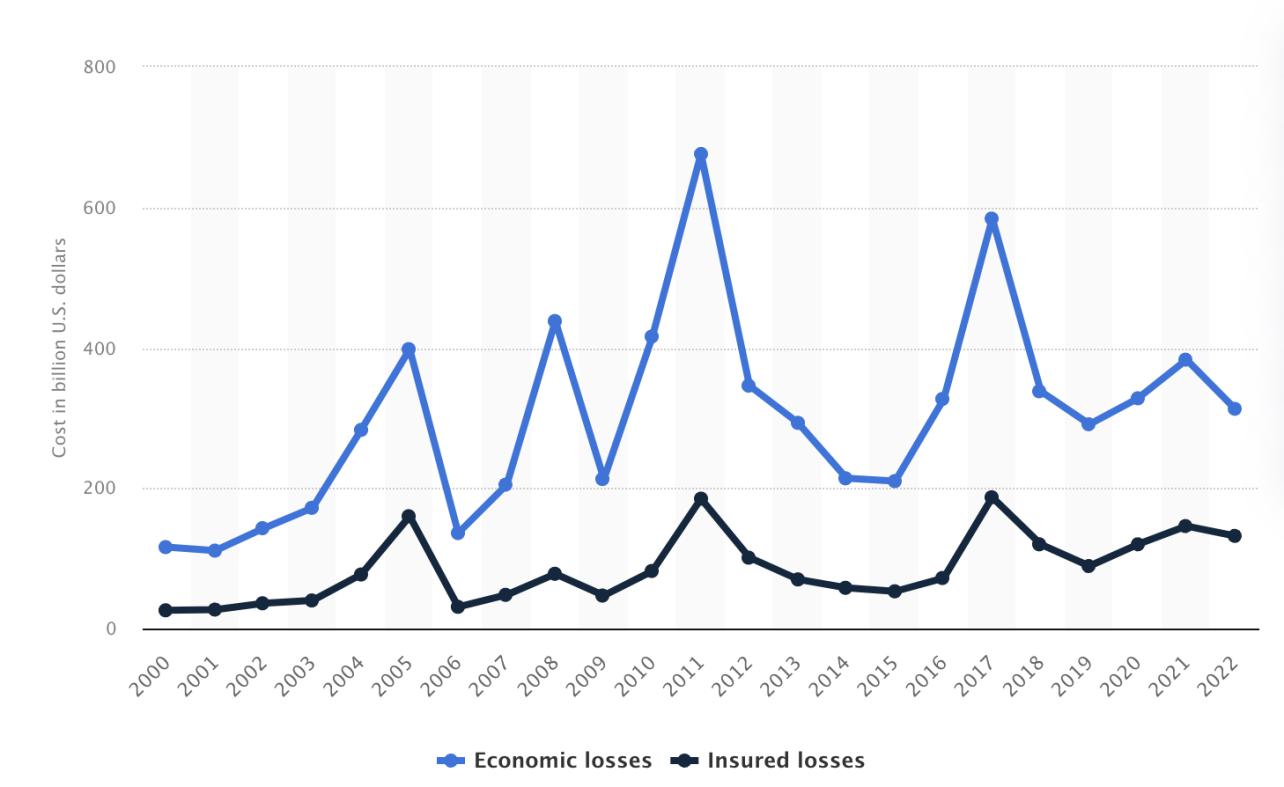
n%20U.S.%20dollars. (Accessed: 15 May 2023).

15

**Appendix**

This section showcases all the material used during this NLP project that has not been included in the main body of the report.

**Appendix 1: Statistics of industrial cost due to Natural Disasters**



Source: <https://www.statista.com/>

16

**Appendix 2: Severity Annotation Guidelines**

**Annotations Guidelines for Severity Ranking for Medical Events during a Natural Disaster**

1. **Introduction**

The purpose of this annotation is to detect aspect terms, intensifiers, and contextual cues from the sentences. Based on Severity Assessment in the Medical Field, different severity rankings will be assigned depending on detections. This will be implemented on the translated disaster messages found in the dataset with our target entity being the classification of severity.

1. **Scope of study**

This study is focused on implementing severity rankings to medical events identified in the dataset, which has been annotated by the authors as ‘**Medical\_help**’. The study will annotate the messages with severity rankings based on the definitions, terminology and context provided by the clinical incident management guidelines published by the department of Health in Australia.

1. **Annotation Technique**

Most severity annotations are based on guidelines that use a combination of multiple techniques including Natural Language Processing (NLP), statistics, human judgement, and expert advice. In this study, the focus is on leveraging NLP techniques like word associations and basic concepts of linguistics. Hence, this study uses the keyword method, whereby the scale of severity changes dependent on the combination of medical type found with potential severity triggers (medical impact and contextual cue).

3.1 Key definitions

Definition of Severity:

The ShARe guidelines define severity as “the relative intensity of a process or the relative intensity or amount of a quality or attribute”.

3.2 Annotation milestones

The task of the annotator is to identify the following types of information:

* **Aspect Term, which will be known as Medical Incidents.**

17

These are single or multi-word terms that are specific types of medical situations. For example, “Chest pain” or “Breathing problem” are considered a medical incident. Since this dataset contains translated messages, we can assume that the word “pain” should be placed under a medical incident. Medical types that should be considered in the annotation:

* Medical condition: symptoms, injuries, diseases
* Common symptoms include fever, headache, cough, pain, dizziness

Medical types that should NOT be included in the annotation:

* Pre-acquired medical conditions should not be considered as we are dealing with situations that occur not pre-exist. For example, diabetes or cancer are serious illnesses but do not contribute towards being severe unless there is another severity trigger is mentioned.
* Fatalities are not considered a medical condition, therefore should not be considered during the annotations.
* **Intensifiers, which will be known as medical impact.**

Verbs, Adjectives, or Adverbs known as Intensifiers determine the impact of a medical incident, such as bleeding, shortening of, high, low, mild, much, severe.

Examples of comm

* **Contextual cues, which will be known as medical extent.**

This severity trigger demonstrates the extent of a medical incident, determining potential for escalation or further harm.

Types of contextual cues to consider:

**Time-related:** increasing, decreasing, quickly, may, rather, could be, need, quick, urgent.

* These types of terms are connective words to the medical impact contributing to its urgency. Terminology that does not specify the high or low spectrum of time such as maybe, could be, might are considered to be uncertain extents. While terminology that specifies the spectrum such as rather could imply “to some degree” can considered significant.

**Numerical stats:** “100 people dead and 25 injured.”

* These are numerical states that are supplied to a medical incident to demonstrate the urgency of a situation.

18

1. **Severity Classification**

The Clinical Severity Assessment guideline was adapted for this ranking system which uses the term Natural Disaster Medical Severity Assessment (NDMSA) as the annotation. Each ranking contains a definition of the type of information that would correspond to a specific severity. Annotators should read and follow this scale that has been specifically adapted for the dataset used.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Severity Ranking** | | | **Definition** | | | |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  | There is a medical incident and no signs | | | |  |
|  |  |  |  |  |  |  |  |  |
|  | **NDMSA 0** | **= No Severeity** |  | of symptoms or intensified situation | | | |  |
|  |  |  |  | with no determined extent | | | |  |
|  |  |  |  | There is a one or more medical incident | | | |  |
|  |  |  |  |  |  | |  |  |
|  | **NDMSA 1** | **= Severity event** |  | contain injuries and or symptoms that | | | |  |
|  |  | have significant intensifiers with certain | | | |  |
|  |  |  |  |  |
|  |  |  |  | extents. | | | |  |
|  |  |  |  |  |  |  |  |  |

**Annotation Guideline for NDMSA 0:**

* 1. When a medical event is expressed through a medical condition:
* “Where can I get help, I am wounded?”

A medical event is defined here as *wounded* as injuries are included as part of a medical event, but the intensity or extent of the incident is not mentioned, therefore classified as mild.

1. Messages that are making a factual statement should NOT be annotated unless they include indications of intensifiers.

 “Thailand: The Bureau of Epidemiology is investigating dengue hemorrhagic fever cases in Phuket following increased case reports.”

While there is a medical condition mentioned here, it does not involve and event, rather a stating of facts.

**Annotation Guideline for NDMSA 1:**

* 1. A medical event expressed through a medical condition with specified intensifier and or the extent of the medical condition.
* “Levels of chronic malnutrition and anaemia continue to worsen.”

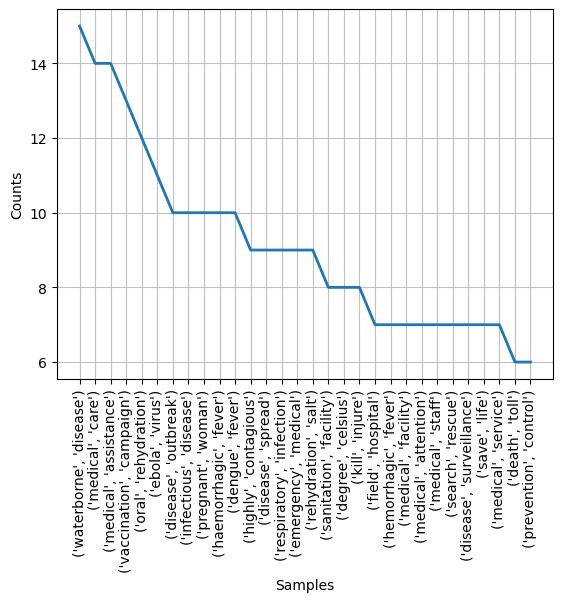
19

The medical condition is *Malnutrition and anaemia* defined by an intensity as *chronic* and the extent of the incident is *worsen*

1. **Human Judgement**

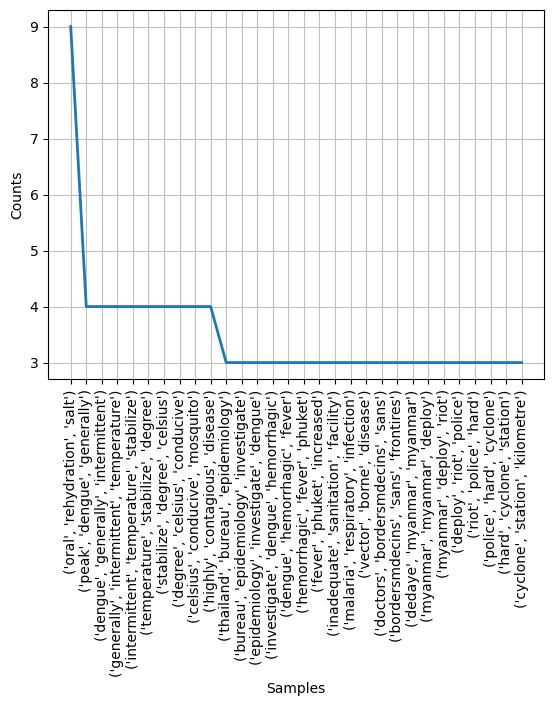
It must be noted that the keyword method with word association is paired with human judgement and intuition as to what is deemed a sever context. The annotators are given the space to use their logic, intuition and experience in grading different scenarios while maintaining the integrity of the scaling definition.

**Appendix 3: Bigrams of the dataset**



20

**Appendix 4: Trigrams of the dataset**



21