Extracting and Organizing Disaster-Related Philippine Community Responses for Aiding Nationwide Risk Reduction Planning and Response

Nicco Louis S. Nocon and Charibeth K. Cheng, PhD
De La Salle University
nicco_louis_nocon@dlsu.edu.ph, charibeth.cheng@dlsu.edu.ph

ABSTRACT

Philippines is one of the most disaster-prone countries in the world. In every experience, spike of data in all mediums are evident and it granted researchers the opportunity to study these data. Attempting to amend the current situation of the country, disaster risk reduction strategies were directly taken from local communities by an online participatory platform called *Malasakit*. In light of the specific needs of people, insights were gathered which provided numerous points on how to prevent and mitigate disasters; however, these insights come in an unstructured form. Hence, this research aims to give structure to the data by extracting disaster-related Filipino community responses, organizing the extracted information, and generating a report for decision makers to address. It was mainly implemented through Part-of-Speech-based Information Extraction with Statistical and Semantical Clustering techniques. As support, a Normalizer and Language Identifier were added. Results for extraction achieved 70.09 Precision, 80.87 Recall, 72.57 Accuracy, and 75.09 F-Measure; while in clustering, all tested approaches successfully joined similar ideas and provided various results in terms of relatedness, clarity, and coverage. Final products of this research exhibit an open-source API and the organized reports. Future directions involve exploring more novel approaches, including other report formats such as infographics or visualizations, commercializing and integrating the products on other applications, and studying the effects of applying the report in disaster strategies.

Keywords

Information Extraction, Clustering, Word Embeddings, Disaster Data.

1. Overview

Disasters are one of the primary reasons for disrupting the world's social and economic status. From 2006 to 2015, the average occurrence of natural disasters is at 376.4, resulting in 69,827 deaths at the average and about $137.6 billion worth of damages. Around the world, China, USA, India, Indonesia, and Philippines are the most disaster-prone countries – where an average of 18.1 natural disasters are annually experienced by the Philippines [6].

In these occurrences, efforts have been given by numerous people and organizations to provide support for the experienced losses. There are those that provided relief programs, money, and goods, and some were inspired to address the problem through research, even using technology for practical use. In fact, these types of situations produce a large amount of data across different sources that are usable in contributing knowledge about disasters.

One work, called *FILIET* [17], took opportunity in acquiring tweets to classify and extract disaster-related contents. Since *FILIET's* data came from an online platform, finding out relevant information was the priority. Its result extracted, presented, and stored information which are details about certain experienced disasters such as casualty, damages, and donations, rather than finding solutions that could help in prevention and mitigation.

A more direct approach was conducted by *Malasakit* [14], an online participatory tool. It collected responses from local communities with ideas on how to make the country better handle disasters. Existing works analyzed these responses through classification [3] and modeling [5] techniques, providing a general representation and understanding of the responses given. Even though hints to what people want or need for their communities were presented, more can still be exploited by capturing specific points in responses that can directly help disaster risk reduction. Doing so would present many ideas that would need certain methods for arranging. Nevertheless, when these ideas are brought up, they can be further used by not only researchers, but also organizations that handles disasters in the country.

Hence, this research attempts to fill what is missing from *Malasakit* and its related works: the extraction of key insights or actionable points in community responses regarding disaster prevention or mitigation, the organization of these thoughts, and a medium that can connect local communities with respective decision makers.

Since *Malasakit* responses are unstructured, Information Extraction (IE), a method focused on automatically extracting and providing structure to unstructured text [7] was implemented. With a structured information, it can be used to disseminate information to an intended target, may it be people or technology (a different application). In relation to *FILIET*, this study adopted the idea of using Part-of-Speech for IE. It is then supplemented by introducing a grouping and ranking technique that organizes the extracted ideas. Furthermore, dissemination in a form of a list report was made as medium to relay and communicate ideas with corresponding decision makers. Accordingly, culminating these ideas produced a text analysis tool that is publicly accessible[1].

Following this, a discussion about the research's methodology is included, followed by showcasing the experiments, results, analysis, and lastly providing a recap of the research and listing its future directions.

[1] https://github.com/noconoccin/Filipino-Text-Analysis-Tool
Google Drive (with model): https://bit.ly/3f56oAC

2. Research Methodology

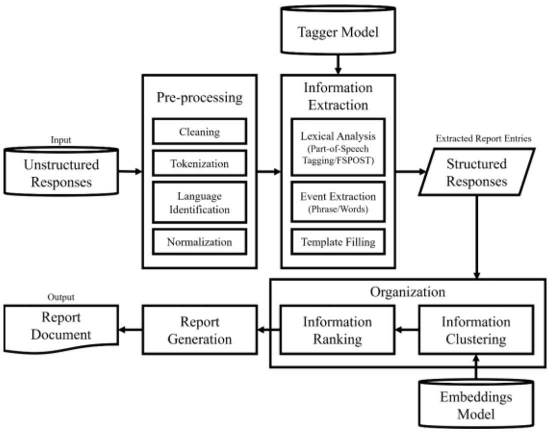


Figure 1. Architectural Diagram of the Tool.

The following are tasks involved in developing the tool (see Figure 1). It was packaged as an open-source Python API.

2.1 Data Acquisition

Data gathered came from *Malasakit's* Local Community Responses [14]. It captures insights to an open-ended question, "How could your Barangay help you better prepare for a disaster". It contains 934 qualitative responses, gathered in Filipino and English, with each comprised of *response* and *tag* features. The *response* feature contains ideas and the main target for extracting and organizing information, while *tag* (or response category) represents a general view of the response given used for ranking.

Along this data, a gold standard was produced by an annotator involved in the *Malasakit* project for evaluating the extraction process. There are two kinds of gold standard: one in the form of insight phrases with sequence of words ranging from proposed actions to its target subject/s and word sets containing word tuples that groups actions and target/s keywords.

2.2 Data Preprocessing

As an attempt to correct misspelled (e.g., bagy0 and maayus) and shortened (e.g., brgy, LGU, and kpag) words in *Malasakit* responses, preprocessing tasks such as tokenization, cleaning, and normalization (word standardization) were applied.

Under normalization, an extendable prefix list [15] to find and join all unmerged prefixes in the text, and a *Filipino Normalizer* [13] were utilized to automatically process or overwrite texts. It covers *textspeak*, *swardspeak*, *conyo* (Tagalog-English mix), and *datkilab* (metathesis or reversed words) styles, built from social media sources and StatMT's Moses engine.

In addition to these, a language identifier was implemented, with the intention of directing appropriate natural language to the Part-of-Speech tagger. An off-the-shelf tool called *LangID* [9] was used, that is built through NumPy's Naïve Bayes classifier and trained on a multi-domain corpus comprised of various documents (i.e., news, encyclopedia, and internet crawled texts). Applying it, the model was filtered to focus on English (en) and Tagalog (tl).

2.3 Information Extraction

Responses were automatically extracted with insights through Part-of-speech-based approach. In detail, responses were labeled by Filipino Stanford Part-of-Speech Tagger (FSPOST) [4]. It was built from Wikipedia data using Maximum Entropy Cyclic Dependency Network, which makes use of features such as word and tag contexts, word affixes, and word shapes in Filipino morphology to determine the appropriate tag for a word.

A Python implementation was done through Natural Language Toolkit (NLTK) [8], which facilitates switching of languages in tagging once the identifier labels a response. Switching involves using NLTK Tagger for English and FSPOST-NLTK for Tagalog.

Equipped with FSPOST, "events" wherein described as verb phrases or verb to noun word sequences that contain actionable suggestions given by the community were extracted. It is done through a pattern of finding an action/verb and traversing through words until a target/noun is found (extended if there is a comma or conjunction *at* 'and' that is followed by another noun). After extraction, template filling provides structure to texts. It works by plugging a set of information into corresponding fields.

2.4 Information Organization

This task is comprised of Clustering and Ranking. There are two ways that it can be set up: one is by organizing all entries, where similar entries are clustered and ranked solely on frequency; Another is by response categories, where each is clustered under a category, ranked by category priority and locally by frequency.

2.4.1 Information Clustering

Entries with similar contents were collated through string and semantic (word embeddings) clustering techniques. For string similarity, an online available Python library called *Strsim*[2] was used, in which the collection contains Sørensen-Dice Coefficient. For word embeddings, pre-trained Wikipedia Tagalog models[3] were used as resources and *Gensim* library [18] for implementing *Word2Vec* [11] and *FastText* [1].

The process starts by collecting all extracted insights and then a set of insights are compared to one another to find words that are similar orthographically or semantically. The process compares two types of string pairs. The first one compares proposed actions or verbs (to join similar actions) and second compares targets or nouns (to remove duplicates within clusters).

Applying a technique results into a computed number representing their similarity. This value is then checked with a similarity threshold. A value lower than or equal to the threshold means the pair are not similar with each other, thus there would be no clustering involved. On the other hand, a higher value means the pair are similar and should be joined into a single cluster, appending their information such as IDs, frequencies, actions, and targets. In cases that a pair is exactly similar, only one instance of the word will remain, but their information will still be appended.

2.4.2 Information Ranking

Since organization accommodates two options, ranking can use frequency counts in decreasing order, the given response categories as basis for prioritization, or a combination of both. Moreover, prioritization for categories was arranged based on its characteristic of being actionable by decision makers specific to handling disasters. Having said that, arranging the categories to prioritize those that can be done before, after, and during disasters will make words tackled (or extracted) be in line with concrete actions to be considered or implemented by decision makers.

[2] https://github.com/luozhouyang/python-string-similarity

[3] https://github.com/Kyubyong/wordvectors

2.5 Report Generation

The ranked information will be transformed into a report, generated in a list format through a template-based approach. The report can be written in Microsoft Word (designed for reading and marking items) or Excel (for analysis). Report generation's main task is to add details and design the entries on those formats, with consideration to setups toggled through information organization. It produces a readable and digital report document which can be relayed to or used by decision makers in handling disasters.

3. Experiments, Results and Analysis

In this part of the paper, experiments were discussed; specifically, changes to utilized data and software modules. Evaluation was done through quantitative and qualitative analysis.

3.1 Quantitative Results

IE's performance was measured through standard, quantitative metrics such as Precision, Recall, Accuracy and F-Measure (see Table 1). Moreover, there were two tests, comparing insight phrases to know the performance in extracting insights and comparing word sets to know the performance in extracting action/verbs and target/nouns.

3.1.1 Insight Phrases

Complete Matches (CM) and True Positives (TP) are statistical measures that counts the number of extractions that matches the gold standard (GS) or are considered as actual insights. In addition, there are partial matches that are also considered as insights which are the following: Over-extractions (OE), Under-extractions (UE), and Overlapping-extractions (OVE).

OE is a type of partial matches with more words in extractions than GS contents. Sample of OE in the results are instances of GS entries without Auxiliary verbs (VBS) at the start of the insight such as *dapat*, *kailangan*, and *be*. Another, OE includes a word or few words after conjunctions *at* or *and* which made it longer. Unlike in GS, some ideas were separated into two entries. An example for this is in GS, "linisin ang kanal" and "itapon ang basura" are separate entries, while "linisin ang kanal at itapon ang basura" 'clean the canal and throw the garbage' is a single extraction entry. Adding to the causes, forcing to search for a noun to end an insight added unnecessary words that could have stopped on an adjective or pronoun.

UE are partial matches with less words in extractions than GS contents. With the same cause in ending a noun for an insight, the effect for UE was different. Instead of having more words, UE lacked necessary details in the insight. The phrase "make the barangay gym" is a sample of UE which should be "make the barangay gym ready for evacuation" to complete the insight. Despite this, there were some instances that can stand alone without the additional details such as "ayusin ang drainage" and "do announcements", which could have also been "ayusin ang drainage ng barangay" 'fix the barangay's drainage' and "do announcements via megaphones".

OVE are partial matches that are equal in length or with certain words overlapping with GS. Sample for this is "kaylangan maging aware" 'have to be aware' and "maging aware sa balita" 'be aware with the news', where they overlap with "maging aware" 'be aware'. There are instances that OVE only differ in spelling, in which a normalizer may have failed to overwrite while the annotator enforced the correct spelling in GS, or the other way around (e.g., provide first aid *kita* / provide first aid *kits*).

Table 1. Results of the Information Extraction Task. [4]

	Insight Phrases %			**Word Sets %**	
	Orig.	**Norm.**		**Orig.**	**Norm.**
CM	**19.59**	18.23	**EM**	**18.75**	18.22
OE	8.35	**8.99**	**PM**	25.6	**26.43**
UE	**25.48**	24.98	**AM**	8.22	**9.32**
OVE	0.3	**0.66**	**TM**	**17.47**	16.08
CMM	**46.27**	47.13	**COM**	**4.02**	2.48
			NMGS	**25.94**	27.46
			NMT	**37.86**	37.97
TP	**53.73 (41.35)**	52.87 (40.24)	**TP**	**51**	50.2
FP	29.69 (17.65)	**29.39 (17.46)**	**FP**	31.09	**30.75**
FN	**16.59 (9.78)**	17.74 (10.6)	**FN**	**17.88**	19.02
TN	0 (31.22)	**0 (31.70)**	**TN**	0	0
P	**64.41 (70.09)**	64.27 (69.74)	**P**	**62.14**	62.03
R	**76.41 (80.87)**	74.87 (79.15)	**R**	**74.06**	72.54
A	**53.73 (72.57)**	52.87 (71.94)	**A**	**51.03**	50.24
F	**69.90 (75.09)**	69.17 (74.14)	**F**	**67.58**	66.88

Complete Mismatch (CMM) are entries that does not match with either the tool's extractions or GS. Under this measure belong the sum of False Positives (FP) which are extractions that are not actual insights nor in GS, and False Negatives (FN) which are not extracted that are actual insights or in GS.

In FP, majority of the instances counted were deemed to be either unusable or insufficient suggestions (as they were missing details) but were still extracted as it indicates an action/verb and a target/noun. Instances of FP are the following: "mentioned in the survey [what was mentioned?]", "allow the residents [allow to?]", and "magkaroon ang mga tao [have what?]" 'people have'.

In FN, results were similar to OE. As some instances with multiple ideas and conjunctions such as *at* were joined by the tool; unlike in GS, the entries were separated. In this tool's extraction, "magkaroon ng komunikasyon at ikutin ng mga council members" 'have communication and rotation of council members', only the first insight was counted as a match with the GS counterpart "magkaroon ng komunikasyon", while the second could not be matched with the same extracted entry and was counted as FN. The restriction to match only once was placed so that the automated evaluation would not be prone to partial match error.

In addition, there were a great number of annotated insights that started with a noun, adjective, or adverb. Examples include "paggamit ng 'use of' public address system", "announcement of a disaster", "proper information dissemination", and "regularly clean canals". Moreover, there were also insights with typographical errors in the extraction's end that did not match with normalized GS as annotations in it are correctly spelled.

[4] For the metrics, values in parenthesis are based on word counts, while those outside of it are based on insight counts.

True Negatives (TN) are instances wherein the tool did not extract as it is not an actual insight nor in GS. Regarding insight counts, the value for this measure is zero, as those that were not extracted and not existent in GS were not included in this type of evaluation. However, for word counts, the value can be computed by counting the total number of words in the data and subtracting word counts in TP, FP, and FN.

Precision (P) is the percentage of extractions that are insights and Recall (R) is the percentage of insights that were extracted. Currently, P and R of IE is 64.41 and 76.41, respectively. With word counts, values are higher with 70.09 and 80.87, respectively. To increase the values of these two, FP and FN were aimed to be low in value, while TP should be higher.

The 53.73 (or 72.57 in word count) Accuracy (A) is another value that represents the percentage of correct extraction, meaning more than half were considered to match with GS. The 69.90 (or 75.09 in word count) F-Measure (F) score is the harmonic mean of P and R which can also be interpreted as the overall performance for extracting insight phrases.

A normalized version of the insight phrases was also evaluated. Generally, the original version gained better scores compared to the normalized one. The cause of this was mostly on spelling mismatches, specifically over-normalization (e.g., *kits* to *kita*, *kapwa* to *kawpa*, and "ayusin ang mga" to "ayusin ang Metro Rail mga"), annotation normalizations (e.g., *kagamitin-kagamitan*, *pundo-pondo*, and publicsoundnotifsystem-public sound notif system), and unnormalized typographical errors (e.g., *plansan*, *taraining*, and *alertoat*). Despite this, numerous words in the data were still standardized; examples are *di* to *hindi* 'no', *san* to *saan* 'where', *anu* to *ano* 'what', *pls* to *please*, and more.

3.1.2 Word Sets

Exact matches (EM) pertain to matches that has the same action/verb and target/noun with GS. Similar to CM and TP, it is ideal to have higher value for this measure. Partial matches (PM) label pertains to matches that has the same action and almost the same target noun with GS.

Instances of PM mostly differ in a few words, more or less than actual GS annotations like [magkaroon, early, warning] / [magkaroon, early, warning, system] and [pagbibigay, assistance, goods], [pagbibigay, humanitarian, assistance, goods]. There are some instances that could have been EM, only to differ in spelling like [improving, imformation, dissemenation] / [improving, information, dissemenation].

Under partial matches, action matches (AM) pertain to only the action field matches with GS. Another is target matches (TM) which only the target field matches with GS. Both are comprised of entries with either insufficient or incorrect action/target, mismatches with GS' typographical corrections (e.g., *pgbaha-pagbaha 'flooding'*), and GS' missed corrections (e.g., *paguusap-pag uusap 'talking'*). Similar to OE, included entries in TM were recorded auxiliary verbs that differed with GS' verbs (e.g., [kailangan, ka-barangay] / [magtulungan, ka-barangay]).

Adding to types of partial matches, when the action matches the target field of GS or vice-versa, it is considered as crossover matches (COM). Examples for this are the pairs: [mabilis, pagbibigay] / [pagbibigay, pagkain] and [do, train, deaf] / [train, deaf, emergency], where the action is designated on the first tuple while the rest indicates targets/nouns.

No Matches for Tool (NMT) is when an extraction does not match with any of the GS entries. No Matches for GS (NMGS) is when a GS entry does not match with any extractions. NMT and NMGS' contents are generally the same with FP and FN. Examples for entries that were not considered as a suggestion by GS are the following: [maiwasan 'avoid', pagbaha 'flooding'], [pagpapadala 'shipping', kunting 'little'], and [putting, pockets]. For reasons that they contain verbs but were used as a justification to the real suggestion, failed to include a proper target, or deemed unusable to act as a solution. On the other hand, NMGS samples mostly have actions that were tagged with a different Part-of-Speech, which could not be extracted due to the verb to noun pattern rule.

Based on the results, less than 20% are exact matches, while PM has the highest number of correct instances with more than 25%. Represented with a combined value of 74.06%, extracting action and target fields were effective provided with a straightforward Part-of-Speech pattern-based design. Perfectly extracting both fields, however, still needs work.

Overall, word sets' P, R, A, and F metric values exceeded half, garnering 62.14, 74.06, 51.03, and 67.58, respectively. Comparing with its normalized counterpart, PM, AM, and FP values improved, while everything else were lower than the original.

3.2 Qualitative Analysis Setup and Results

Qualitative analysis was conducted on organized information to bring out the positive and negative characteristics, as well as the coverage of clustering approaches and collated information. Each of the experiments was clustered through Dice Coefficient, Word2Vec, and FastText, with applied lexicalization on insights.

Experiments were generated, analyzing Malasakit Responses that were organized with all entries under one level, and another grouped by response categories. Extending the tests, a normalized version was evaluated. Furthermore, to determine the impact of the 50% default clustering threshold – a value indicating if words are considered similar or not – adjustments such as decreasing and increasing its value were done.

In this type of analysis, the following was discovered: the highly suggested insights, set of similar words, effectivity of applying a normalizer, more appropriate configuration, and many more.

3.2.1 Community Suggestions

The entirety of the extraction process produced 1,392 insights. Two perspectives can be taken away from this: (1) by organizing all entries regardless of which categories they are under, insights focused mainly on actions provided; and (2) by organizing per category, it is asserted that insights are already under a particular theme, which made clusters/actions under it more focused into suggesting steps that contributes to that category.

Top ideas collated from organizing all entries are the following:

- **Logistics** mentioned items in need such as early warning system, medical kits, flashlight, garbage cans, shelter, medicine, and grocery/supply.
- **Dissemination** suggested information about disasters such as typhoons, floods, storms, and consequences that comes with them.
- **Sanitation** points mostly towards cleaning the surroundings, specifically sewers and areas near households.

- **Community and self-preparedness** through alarms and designating places such as schools for evacuations were suggested, mainly to be able to avert or avoid ramifications such as floods or clogging, spread of tragedies, and getting trapped.
- **Solidarity** recommended families and the Barangay to continuously help or support each other.

For reference, these topics were formed from generalizing verbs found in clusters. Based on this, *Malasakit's* predefined categories adequately covered the dataset. No new categories were discovered as all are under its scope.

Top ideas collated from organizing by response categories are the following:

- **Information Campaign and Capacity Building** focused on actions that the community must have such as items or programs, preparations, support, and logistics for information dissemination. It also enumerated intended contents, medium of information, and receiving end of the information. Its key insights are: (1) conducting programs such as seminars, drills, and assemblies, (2) using infographic materials such as signages, posters, and leaflets, and (3) suggested contents for this include reminders or tips on what to do before, during, and after the calamity, while the target audience are family members.
- **Disaster Relief** pertained to the same ideas which mentioned receiving or providing assistance. Its key insights are receiving goods, food or grocery, and medicine, whilst providing evacuation options.
- **Community-wide Logistic Support for Disaster Response** provided ideas that involves safety and support. Its key insights are: (1) to have safety gears, sirens, and shelter for the operations, (2) to add more budget, equipment, and volunteers, (3) to build boats and storage facilities, (4) to build or buy an area or place for disaster response, and (5) to communicate weather predictions to the community.
- **Infrastructure Maintenance and Management** were about sanitation and repairs. Cleaning the mentioned areas in the minds of the community would ensure prevention of floods and clogs. Its key insights are: (1) clean surroundings such as streets, sewers, rivers, and garbage waste, (2) to avoid throwing trash everywhere, and (3) there must be a proper place or containers for garbage.
- **Early Warning System** involved information dissemination. Its key insights are: (1) having proper communication, alert, news, updates, and radio, (2) information should be about disaster and announced to the public, specifically people, citizens, and residents, and (3) alerts with regards to calamities, catastrophe, disaster, typhoon, communication, support, and assembly must be provided.
- **Preparedness for Emergency** showed ideas in a form of a reminder to people. Its key insights are: (1) to alert and be aware or attentive with nature, news, and officials, (2) medium in reminding people include watching weather forecasts or news in television, and (3) preparation is needed on the following: news, evacuation plan, officials, management, unit, and the community.
- **Local Government Accountability** indicated expectations of the community with the government. Its key insights are: (1) the government should be accountable for subjects such as disasters, evacuations, posters, cases, and the society, (2) it is expected for them to be ready, cooperative, and able to show and send help to the community, and (3) they should be active and focused on helping people.
- **Filipino values** contained traits that the whole community must have to get through disasters. Its key insights are: (1) encouraging solidarity, that is to keep everyone united and participative in helping each other, (2) help should be observable in preparing families and community, (3) people must be cooperative with duties, plans, and preparation, and (4) traits mentioned tidiness, readiness, equality, kindheartedness, concern, and order.
- **Others** contained mixed ideas to improve disaster prevention and mitigation. Its key insights are: (1) calling out corruption, which specifically mentioned the act of putting valuables in pockets, (2) endorsing their satisfaction with decision makers, encouraging to continue their work or activities, and (3) minor and major entities in disasters are mentioned, specifically suggesting communication between deaf, citizens, and responders.

3.2.2 Lexicalization

Lexicalization was applied in all experiments, where one word was designated to represent a cluster. Implementing this was successful in labeling joined words related to nouns. Orthographic instances represented tenses and unstandardized variants such as seminars (seminar, seminarsdrill), floods (flooding), *basura* (*basurahan*) 'trash', and *kapaligiran* (*paligid*) 'surroundings'. Whereas semantic instances represented a set of words under an idea such as training (community, programs, assembly, government, technology, and more). In spite the attempt, there were still instances of unrelated and missed lexicalizations. Examples are evacuation-elevation (unrelated) and *kanal-imburnal* 'canal' (missed).

Primary learning in this experiment is its effectivity highly depends on the performance of clustering approaches. One thing to note of is its potential in displaying a single word that can represent an entire set of words given an adequate clustering approach. Selecting and retaining a single word however is another problem to tackle.

3.2.3 Normalized Version

Table 2. Normalization Samples

Original-Normalized		
andyan – nandiyan	kase – kasi	s – sa
anung – anong	konting – kaunting	san – saan
aq – ako	kpag – kapag	tas – tapos
cla – sila	meron – mayroon	tsaka – at saka
di – hindi	pano – paano	tv – television
dont – do not	pls – please	xa – siya
facebook – Facebook	pwd– puwede	yung – iyong

Shown on Table 2 are few samples of the corrections using [13] that covered shortcut texts, typographical errors, and Filipino colloquialisms. Even though the degree and coverage of the normalizer was sufficient to correct responses, not all were normalized correctly. Since it was built to be dependent on a statistical model, some instances resulted undesirable insertions and replacements. An instance of this inserted an extra *it* in between *as* and *possible* in the phrase “as early as possible”. Another is “*ayusin ang*” which has “metro rail” succeeding it.

There were also issues between colloquialisms and interlingual homographs[5]. In the English phrase “seminar for pre or post…”, the prefix *pre-* was found to be a shortcut for the colloquialism *pare* ‘buddy’, instead of leaving it as is. Other mistakes are *kits* in “medical kits” were considered as a typographical error for *kita* ‘you’, *to* into *ito* ‘this’, *non* into *noon* ‘previously’, and *my* into *may* ‘there’. Unfortunately, some repercussions of these mistakes removed words from clusters.

With incorrect normalizations, there were also unnormalized instances. Merged words were not covered by the normalizer such as *atpagbigay*, *sumunodkapag*, and *dahilbansa*. Another is a typographical error variant that substituted letter ‘q’ as ‘g’ such as *paqdating* and *paqaabiso*. Moreover, there were also shortcut variants, specifically the omission of vowels, that were not present in the statistical model such as *ngbbgay*, *dhlan*. *magki2ta*, *mlman*, *gwen*, and *magsgwa*.

Observing its effect, frequency counts fluctuated, shifting the order of clusters, but did not affect the highly frequent ones. It has been evident that there were shortcuts previously included in the clusters which was then removed, and new or corrected words appeared. Given this, there have been several instances that shortcuts as such were tagged properly.

In a similar way, joining Filipino/Tagalog prefixes with their separated root words as part of the normalization task caused a set of words to appear as insights. One instance, *mag-* prefix was joined with *karoon* which resulted into *magkaroon* ‘have’, a valid member of the proposed actions. More examples like this found on the experiment are: *nagpeperform* ‘performing’, *pagpapaalala* ‘reminder’, *pagbigay* ‘giving’, and many more.

3.2.4 Threshold Adjustment

Decreasing and increasing the similarity threshold from 50% (default) to 20% and 80%, respectively, affected the members of clusters. Reducing it signifies a loose acceptance in determining the similarity distance between two words. Since the condition is more lenient, there were more related words (variants) captured in clusters. In fact, it enabled the clustering process to capture variants with far distances that were not clustered in the base experiment. Examples for this are pairs: *dagdagan*-add, *maayos-ayusin* ‘fix’, and *mabigyan*-provide, all of which are synonyms of each other. Although in this adjustment, there were some clusters that mixed up ideas under a vaguely large topic such as medicine-cleanliness, barangay-typhoon, and cause-food.

Increasing the threshold on the other hand, tightens it, thus producing harder but more closely similar clusters. Since the condition is stricter, unrelated words were filtered, producing more accurate and clear manifestation of relationship between them. Positive samples separated evacuation-elevation, and clustered words such as drill-drills and training-community. Moreover, there were others that even produced better clusters; before under one cluster are *pagbaha* ‘flood’, *pagbabaha*, and *pagbara* ‘clog’, then after threshold increase created two clusters with pairs *pagbaha-pagbabaha* and *pagbara-pagbabara*.

However, this restriction is not perfect, as instances with slight variation (just a letter in some) can lead to either good, bad, or unaffected clustering. Negative samples were unable to cluster *kit-kits*, *training-taraining*, and *basura-basurahan*. In addition, the increase dispersed cluster members outside the condition of being under a single verb – meaning verbs must be the same to cluster specific nouns together.

Comprehensively for this experiment, it proved that increasing or decreasing the threshold affects the placement of words in their appropriate clusters. It controls the scope and composition of a cluster's contents and excessive amounts of these adjustments could decline the quality of clusters. Consequently, the threshold value was set into its default and balanced threshold of 50%.

3.2.5 Comparison of Clustering Approaches

Three clustering approaches were applied in the experiments, namely Dice’s Coefficient, Word2Vec, and FastText. Dice’s coefficient cluster words with orthographic similarity using n-grams, while Word2Vec uses a vector space to cluster semantically. FastText incorporates the idea of the two to determine whether words are related to each other.

Dice’s coefficient was able to group variants of a word, within the constraints of changes in affixes. Characteristic as such enabled it to cover shortcuts and typographical errors not far from the original’s form or structure, even without utilizing a normalizer. Provided, clusters have clearer and interpretable relationships.

Since Dice’s coefficient investigate features of words through a set of characters, one difference between it and Word2Vec is vectors are positioned based on usage – so words that operate the same way are closer together in the space. However, usage does not always mean it produces ideal groupings. There were clusters that produced antonyms like *tao-bagay* ‘human-object’ and *sakuna-sanhi* ‘disaster-cause’, and members with relationship indistinguishable from each other.

In word embeddings approaches, both were able to cluster based on usages, thematic similarities, and even those with orthographic similarities. Examples for these covered synonyms, antonyms, tense variants, code-switching, shortcut or typos, and other topic/themes. Coverage of orthographic similarity was more evident on FastText as it uses sub-word information (character n-grams) as vector representation. With this, verb tenses such as *improve*, *improved*, and *improving* were captured; as well as intra-word code-switching such as *ma-improve*. Moreover, it also inherits Dice’s capability in clustering shortcuts (e.g., *nagbbigay-magbigay* ‘give’, *magtulong2-magtulongan* ‘help’), one that is not covered by Word2Vec.

Synthesizing the results, characteristics in positive and negative aspects were similar, but quality of clustering has changed. FastText specifically has more characteristics taken from the other two, hence there were more combinations as to the resulting clusters.

Comprehensively, these three still have room for improvement, especially in capturing the right balance of relationships between words. Downside in Dice’s coefficient is its limitation on covering only in the confines of string similarity, where literal character distance matters. In word embeddings approaches, there were still

[5] words that exist on both English and Filipino language

fragments of word variants scattered across different clusters that could have been captured. Particularly obvious instances such as prepare-preparing, *linisin-maglinis* ‘clean’, giving-*pagbibigay*, announce-inform, and *gamit-bagay* ‘object/thing’. Undoubtedly, all approaches were able to fulfill their purposes in capturing conceptually similar ideas and in some instances formed one, interpretable idea in clusters.

3.3 Processing and Analyzing News Dataset

In terms of data experimentation, there was a test on another domain, News. The assumption is since the solution processes Part-of-Speech and Filipino texts, it should be able to handle inputs regardless of domain. This test is a proof of concept that it can handle any text given the current implementation.

The News dataset was taken from a Philippine language resource collection [16]. Sources used for this experiment are from Pang-Masa and Pilipino Star Ngayon 2015, both in Filipino with 2,000 sentences. There are four main categories, namely Entertainment (Movies and Showbiz), Sports, World, and Opinion. World category has 800 sentences and the rest with 200 sentences each.

In this dataset, sentences were processed through the API’s main tasks. Organization was implemented to process by category with prioritization set alphabetically. Like previous experiments, all three approaches were used in clustering.

Extracting information produced 3,503 insights from 1,689 sentences, where 311 sentences without insights. Based on this, the extraction algorithm exhibited decent results. It showed that as long as the Part-of-Speech tagger is reliable and the language is covered, it would be able to process texts that can point out actions found on a sentence. Sample sentences are as follows: for Entertainment, “Nag-tweet si Vice” ‘Vice tweeted’; for Opinion, “NAGULAT ang mga kidnaper” ‘Kidnappers were shocked’; for Sports, “sinimulan na ang fitness test” ‘fitness test has begun’; and for World, “Ituturo sa lahat ang tamang paghuhugas” ‘proper washing will be taught to everyone’.

In addition, as the tool covers English, there are positive instances of extracting English phrases. Few of these are “controlled the game”, “received proper medical attention”, “plays the title role”, and “worked hard for the movie”.

However, in some cases, the tagger failed to label words correctly, producing incorrect entries such as “asul na van” ‘blue van’, “liblib na lugar” ‘secluded place’, “bandang alas-7” ‘around 7’, and “nakaraang week” ‘past week’. These examples do not point to an action, instead describes a noun. Moreover, there are some that does not exhibit enough idea such as “am Maldives”, “sa semis”, “ani Chulani”, “taus puso”, and more.

On the condition that extracting information in news domain has been successful to capture actions and targets, potential use in news/media setting has been considered. One of which is insight phrases can provide a summary of the contents in an article. By taking parts of an article, specifically actions or events in it, readers would have an idea on what happened in the article. At the same time, word sets can be used as keywords or tags, displaying the article’s main occurrences (e.g., *pinatay* ‘killed’, *nadakip* ‘taken’, *nakaresponde* ‘responded’, etc.) and involved subjects (e.g., which celebrity, victim, team, area, etc.). Regardless, this experiment showed potential use to other domains.

Regarding results for clustering, Dice’s coefficient was able to cluster verbs mainly but was ineffective to targets. It is due to a lot of clusters containing distinct set of nouns that made string distances farther from each other. However, given the chance, it could cluster nouns that has been generally used such as *biktima-biktimang* ‘victim’, player-players, alas-9-alas-2, and the likes.

Differentiating with Dice’s coefficient, Word2Vec produced more but imperfect set of members for both verb and noun fields. It joined semantically related ideas that talks about movement (e.g., *lumisan* and *umalis* ‘leave’), franchise (e.g., actor and character), notable names (e.g., Aquino and Sharon), sports (e.g., players and championship), crime (e.g., *suspek* ‘suspect’ and *pulis* ‘police’), places (e.g., Quezon and mall), family (e.g., *ina* ‘mother’ and *anak* ‘child’), transportation (e.g., *kotse* ‘car’ and *pasahero* ‘passenger’), and other contextual similarities (e.g., kuwento-buhay ‘story-life’ and *kaibigan-kapatid* ‘friend-sibling’).

Applying FastText, results showed more ideas clustered together, significantly lowering the number of clusters – where most of the responses belonged to a cluster with members more than one. The most compelling quality of FastText is its ability to be able to capture both orthographic and semantic similarities. Having said, positive instances of the other two approaches were present, with more varieties in terminologies as compared to the original dataset. Even though this is the case, their negative qualities were also passed down, where it was still unable to combine closely related terminologies and instances with uncertain relationship were still clustered. Consequently, the large quantity of clustered sentences made it harder to interpret for its vague relationships.

3.4 Survey on API and Report

For this research, two outputs have been created, namely an Application Programming Interface (API) of the tool and a Report. In evaluating the outputs of this study, a survey was performed on both with questions pertaining to the quality of outputs and discussions on user feedback were provided.

3.4.1 API Functionalities

API is defined as a collection of functions, intended for researchers and developers alike. It has nine modules, namely Data Utilities, Normalization, Language Identification, Filipino Part-of-Speech Tagger, Information Extraction, Information Organization, Information Clustering, Information Ranking, and Report Generation. As a whole, the functions under each module are intended to be useful in future research or application, not necessarily within *Malasakit*, that involves the tasks related to the modules above.

3.4.2 Report Formatting

The generated report to be used by decision makers was initially a two-column Microsoft word document, with extracted insights/suggestions organized in two ways based on Information Organization. Each report contains three parts: introduction, insight list, and list of *Malasakit* responses.

In the insight list, each entry contains fields such as Sentence ID (of cluster members), Frequency Counts (number of responses), Proposed Action (verbs extracted and clustered), and Target (nouns under verb clusters). Additionally, lexicalizations are formatted by enclosing them in parentheses.

3.4.3 Survey Procedure and Results

Prior to answering the questionnaires, an informed consent form was provided. Upon signing, two forms were presented. One intended for API assessment, a form designed from USE Questionnaire [10] and NormAPI’s [12] version of the same. The other is intended for Report assessment designed from TAM

Model [2]. Both questionnaires have quantitative and qualitative part, measuring the agreement in certain aspects using numbers and acquiring comments and opinions, respectively.

The survey was conducted on five (5) *Malasakit* team members. In API assessment, the USE Questionnaire has four themes, namely Usefulness, Ease of Use, Ease of Learning, and Satisfaction. Its overall rating in each of the criterion produced a value of 4.3, 3.76, 3.8, and 4.4 out of 5, respectively. As a whole, it was rated with 4.07 out of 5 score.

Initially, using the API has been evaluated to be easy to learn and working as intended. It has been described to excel in providing increased productivity rate when partnered with other tools and it reduces time in accomplishing tasks. Moreover, it works as intended and applying it to other tools was deemed effective, recognizing its flexibility due to its modular design.

In qualitative part of the assessment, participants have pointed out a series of improvements. Recommendations mentioned better documentation and potentially extending the API by adding Graphical User Interface, packaging using pip install, and commercializing using REST API.

Better supporting documents were addressed by writing manuals. Combining the contents, these documents contain detailed list of functions and instructions on how to use the API properly. To ensure proper running of modules, *requirements.txt* were created, listing packages required by the API. These files were packaged and reflected on Github Repository and Google Drive.

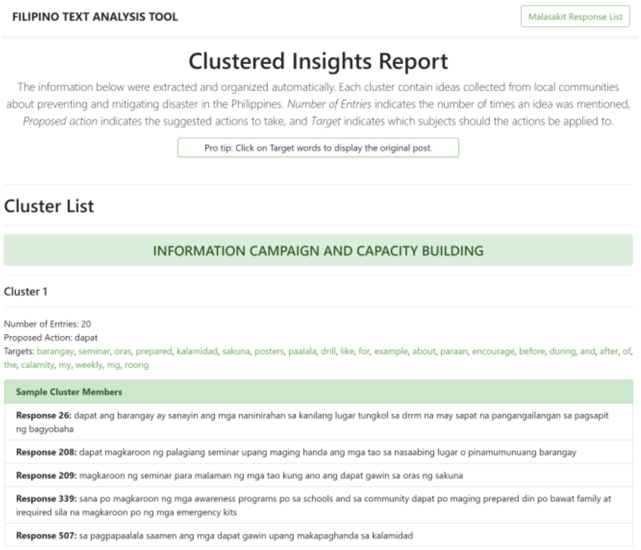
In assessing the generated report, it has been rated overall with 4.25 out of 5, a value that resides with agreeable to strongly agreeable elements. Among the series of questions, the highest average score garnered 4.6, where majority of the participants agreed to being satisfied with the report. Characteristics regarding the presentation and organization style of the information, as well as elements found in the report were deemed acceptable by participants. It has also been acknowledged to contain useful information and have an impact to the participants' job and tasks.

Although contents were found to be readable and sufficient to make strategic decisions, some pointed out the need to provide a background on what to expect in the report and instructions on how to interpret the information. It was also overwhelming for some, considering the length of pages and volume of the contents. To amend this, a localized HTML version (see Figure 2) designed through Bootstrap was supplied which can automatically be generated by reading an excel file with cluster values. Main intention is to make the report readable by adjusting the contents included, while maintaining its simplicity.

In this version, the introduction was modified to provide an idea on what to expect in the report. Similar to the word document, it contains the list of clusters and *Malasakit* responses. Additional features focused on enhancing navigation and readability by adding a scroll-to-top functionality, assigning hyperlinks to target words (clicking redirects users to the original response), and limiting the number of clusters (default shows top 25 clusters) with each at most five sample cluster members for display (all values are adjustable prior to generation).

Given this implementation, there are still other improvements that could be done for better user experience. The most straight-forward way is to extend the functionalities and design of this web report. It could be done by making hyperlinks and original posts more accessible such as pop-ups or collapsing text components. Although, it is highly recommended to present the information through interactive graphs or word networks.

Taking opinions regarding organizational and clustering preference, most preferred it by response categories and through Dice's Coefficient. Main takeaway is Dice's displays a clearer relationship on actions and targets through its orthographic similarities, as compared to semantic similarities which contains vaguely related and diverse clusters. Complemented with categories, the layout looked simpler and easier to interpret.



Clustered Insights Report

The information below were extracted and organized automatically. Each cluster contain ideas collected from local communities about preventing and mitigating disaster in the Philippines. *Number of Entries* indicates the number of times an idea was mentioned, *Proposed action* indicates the suggested actions to take, and *Target* indicates which subjects should the actions be applied to.

Pro tip: Click on Target words to display the original post.

Cluster List

INFORMATION CAMPAIGN AND CAPACITY BUILDING

Cluster 1

Number of Entries: 20
Proposed Action: dapat
Targets: barangay, seminar, oras, prepared, kalamidad, sakuna, posters, paalala, drill, like, for, example, about, paraan, encourage, before, during, and, after, of, the, calamity, my, weekly, mg, roong

Sample Cluster Members

Response 26: dapat ang barangay ay sanayin ang mga naninirahan sa kanilang lugar tungkol sa drrm na may sapat na pangangailangan sa pagsapit ng bagyobaha

Response 208: dapat magkaroon ng palagiang seminar upang maging handa ang mga tao sa nasaabing lugar o pinamumunuang barangay

Response 209: magkaroon ng seminar para malaman ng mga tao kung ano ang dapat gawin sa oras ng sakuna

Response 339: sana po magkaroon ng mga awareness programs po sa schools and sa community dapat po maging prepared din po bawat family at irequired sila na magkaroon po ng mga emergency kits

Response 507: sa pagpapaalala saamen ang mga dapat gawin upang makapaghanda sa kalamidad

Figure 2. Generated Report

4. Conclusion and Further Works

In this paper, the development of a Filipino text analysis tool with automated insight extraction and organization was discussed. Results showed that the use of Part-of-Speech Information Extraction achieved satisfactory marks on standard metrics such as Precision (P), Recall (R), Accuracy (A), and F-Measure (F). Specifically, on insight phrases and word sets, scores as such were recorded as P = 70.09, R = 80.87, A = 72.57, F = 75.09, and P = 62.14, R = 74.06, A = 51.03, F = 67.58, respectively.

Other modules undergone experimentations and through analyses found their advantages and disadvantages. Results showed the effectivity in organizing and clustering, regardless of approach. Assessing the API and report, both achieved satisfactory remarks with 4.07 and 4.25 out of 5 ratings, respectively. The tool excelled in usefulness and satisfaction, while the report on appropriateness, impact, presentation, usefulness, and satisfaction.

Specifying contributions of this research, the reports could serve as a data source for future disaster-related research and assist decision makers in disaster planning and response. The API could then be used to develop and supplement other Data Processing or Natural Language Processing tasks. Furthermore, it can be extended further for optimizations and addition of features.

Future directions of this study involve improving the heuristics of the extractor, exploring more novel clustering approaches, adding Graphical User Interface in the API, including other report formats such as infographics or visualizations, commercializing the tool, integrating the tool on other software applications, and studying the effects of applying the report in disaster planning and response.

5. Acknowledgments

Our thanks to Commission on Higher Education, Department of Science and Technology – Science Education Institute, Philippine-California Advanced Research Institutes, and De La Salle University for providing us support in this venture.

6. References

[1] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135-146.

[2] Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. 1989. User acceptance of computer technology: A comparison of two theoretical models.

[3] De La Cruz, A., Oco, N. and Roxas, R. E. 2017. A classifier module for analyzing community responses on disaster preparedness. In *Proceedings of the 2017 IEEE 9th international conference on humanoid, nanotechnology, information technology, communication and control, environment and management (HNICEM)*, 1-5. IEEE.

[4] Go, M. P. and Nocon, N. 2017. Using stanford part-of-speech tagger for the morphologically-rich filipino language. In *Proceedings of the 31st pacific asia conference on language, information and computation*, 81-88.

[5] Gorro, K., Ancheta, J. R., Capao, K., Oco, N., Roxas, R. E., Sabellano, M. J. … Goldberg, K. 2017. Qualitative data analysis of disaster risk reduction suggestions assisted by topic modeling and word2vec. In *Proceedings of the 2017 international conference on asian language processing (IALP),* 293-297. IEEE.

[6] Guha-Sapir, D., Hoyois, P., Wallemacq, P. and Below, R. 2017. Annual disaster statistical review 2016: The numbers and trends. *Brussels: Centre for Research on the Epidemiology of Disasters (CRED)*.

[7] Jurafsky, D. and Martin, J. H. 2018. *Speech and language processing.* Stanford.

[8] Loper, E. and Bird, S. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

[9] Lui, M. and Baldwin, T. 2012. Langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, 25-30.

[10] Lund, A. M. 2001. Measuring usability with the use questionnaire. *Usability interface, 8*(2), 3-6.

[11] Mikolov, T., Chen, K., Corrado, G. and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[12] Nocon, N., Cuevas, G., Magat, D., Suministrado, P., and Cheng, C. 2014. Normapi: An api for normalizing filipino shortcut texts. In *Proceedings of the 2014 international conference on asian language processing (IALP),* 207-210.

[13] Nocon, N., Kho, N. M. and Arroyo, J. 2018. Building a filipino colloquialism translator using sequence-to-sequence model. In *TENCON 2018-2018 IEEE region 10 conference*, 2199-2204.

[14] Nonnecke, B. M., Mohanty, S., Lee, A., Lee, J., Beckman, S., Mi, J. … Goldberg, K. 2017. Malasakit 1.0: a participatory online platform for crowdsourcing disaster risk reduction strategies in the philippines. In *Proceedings of the 2017 IEEE global humanitarian technology conference (GHTC),* 1-6. IEEE.

[15] Oco, N. and Borra, A. 2011. A grammar checker for tagalog using languagetool. In *Proceedings of the 9th workshop on asian language resources,* 2-9.

[16] Oco, N., Syliongka, L. R., Allman, T., and Roxas, R. E. 2016. Philippine language resources: Applications, issues, and directions. In *Proceedings of the 30th pacific asia conference on language, information and computation*, 433-438.

[17] Regalado, R. V., Kalaw, K. M. D., Lu, V., Dela Cruz, K. M. H. and Garcia, J. P. 2015. Filiet: an information extraction system for filipino disaster-related tweets. In *Proceedings of the DLSU research congress vol. 3*.

[18] Rehurek, R. and Sojka, P. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* 45-50. Valletta, Malta: ELRA. (http://is.muni.cz/publication/884893/en)