

ENG 4000: Minimum Viable Product Review

Project: Disaster Tweets - Real or Not: Natural Language Processing

Date: December 10th, 2020

Team P

Binte Zehra	215624141
Neena Govindhan	212137600
Jessie Leung	215985948
Jonas Laya	214095715
Paul Sison	214447510

Table of Contents

Executive Summary	6
Introduction	7
Budget Proposal	8
Key Stakeholders	9
Social Impact and SDG Goals	10
Product Backlog	12
Sprint 1	14
Sprint Planning	14
Timeline	15
Roles	15
Sprint Goal(s)	15
Sprint Backlog	16
Tasks	17
Sprint 1 Review	21
Completeness	21
Burndown Chart	21
Velocity	21
Challenges	22
Sprint 1 Retrospective	23
Happiness Evaluation Metric	23
Process Related Issues	24
Sprint 2	25
Sprint Planning	25
Timeline	25

Roles	25
Sprint Goal(s)	26
Sprint Backlog	26
Tasks	27
Sprint 2 Review	36
Completeness	36
Burndown Chart	36
Velocity	36
Challenges	37
Sprint 3 Retrospective	38
Happiness Evaluation Metric	38
Process Related Issues	39
Additional Project Management Content	40
Project Schedule	40
Project Risk Register	41
Project Management Tools	43
Team Content	44
Individual statements on Contributions to Project	44
Individual statements on ITPMetrics survey (Peer feedback)	46
Team statement on ITPMetrics survey	48
Self-Evaluation	49
Appendix A: Research Papers	50
Appendix B: Glossary	54
References	55

List of Figures

Figure 1: Formulae to calculate Fscore	17
Figure 2: Supervised classification process from NLTK	19
Figure 3: Burndown Chart	21
Figure 4: Sprint 1 Happiness Graph	23
Figure 5: Non-neural Network Models Results	29
Figure 6: Confusion Matrix for Best and Worst Performing Classifier	29
Figure 7: LSTM Confusion Matrix	30
Figure 8: News and Social Media Survey	31
Figure 9: Age Pie Graph	32
Figure 10: News Source Survey	32
Figure 11: News Content Survey	33
Figure 12: Familiarity with Twitter Survey	33
Figure 13: Reliability Survey	33
Figure 14: News Content on Social Media During Emergency Survey	34
Figure 15: Social Media to Notify People Survey	34
Figure 16: Sprint 2 Burndown Chart	36
Figure 17: Sprint 2 Happiness Graph	38
Figure 18: Now-Next-Later Framework	40

List of Tables

Table 1: Product Backlog	13
Table 2: Self Perceived Expertise	14
Table 3: Sprint 1 Backlog	16
Table 4: Results of classification techniques for cross-validated training set using 10 stratified k-folds	20
Table 5: Happiness Evaluation Metric	23
Table 6: Sprint 1 Process Related Issues	24
Table 7: Sprint 2 Backlog	27
Table 8: LSTM Models Score	28
Table 9: Classification Model Results	30
Table 10: Happiness Evaluation Metric	38
Table 11: Process Related Issues	39
Table 12: Project Risk Register	43
Table 13: Self-Evaluation	49

Executive Summary

In times of emergency, Twitter and social media have become an important communication channel for people to stay updated on what is going on in the world. However, extracting useful information from Twitter can be difficult due to the noisiness of social media platforms. The crowd-sourced nature of Twitter means that it provides instant, first-hand eyewitness accounts of live events, but also that it can be cluttered by spam, misinformation, and fake news.

The purpose of this project is to explore the usefulness of Twitter as a source of information. Ultimately, the project objective is to combine openly accessible crowd-intelligence with machine learning such that useful and actionable information can be extracted from Tweets. To limit the scope of the project such that it can be accomplished within the eight month timeframe of the ENG 4000 capstone, the project is focused on extracting reliable information about natural disasters.

Key deliverables of the project include:

- A machine learning model using a Natural Language Processing pipeline to predict whether a Tweet is about a real disaster or not (Kaggle challenge)
- A website to present reliable natural disaster information extracted from Tweets in a manner that is consumable by the general public
- A targeted website or application that provides actionable natural disaster information to first responders

The dissemination of fake news on the Internet is currently hugely problematic in that it misleads the public in a far-reaching and lasting manner. The proposed project seeks to automate truthful news extraction from social media posts in a timely manner, such that in times of emergency actionable information can be accessed by first responders. The importance of this task is that in occurrences of natural disasters, immediate situational information can aid in the highly time-sensitive rescue operations of first responders. In disaster situations, even a marginal advantage could result in one or multiple lives being saved, or the prevention of catastrophic damages.

The technical risks of the project are associated with the hugely noisy nature of the Twitter platform and the technical sophistication of machine learning models. The sheer amount of information available on Twitter makes extracting useful information a highly complicated task. To fulfill stakeholder needs, a classification model must be developed that performs to an acceptable level in determining whether information contained in a Tweet is real and useful or not. Another technical risk is the potential mismatch between the technical sophistication of the model and stakeholder trust. The topic of machine learning is complicated, and consumers of the proposed project (both general public and first responders) are non-experts of the subject. As technical sophistication rises, stakeholder trust in the end product may decrease. As such, a key challenge of the project will be to complement technical implementation with a presentation of the product and its results in a manner that is easily consumable by stakeholders.

1. Introduction

The purpose of this document is to present the Minimum Viable Product for the ENG 4000 Mid-term Capstone submission. The MVP project presented achieves the first project objective outlined in the executive summary, in which the Kaggle challenge (“Real or Not? NLP with Disaster Tweets”) is completed by implementing a machine learning model using a Natural Language Processing pipeline to predict whether a Tweet is about a real disaster or not.

This document first outlines the Budget Proposal, Key Stakeholders and SDG Goals of the entire project including post-MVP extensions. The need and scope of the project is presented in these sections. The document will then outline the process of the first two sprints of the project, including planning, sprint, review and retrospectives following an iterative agile approach. The sprint tasks include development of deliverables for the project MVP.

Each sprint will track progress towards the project objectives, as well as analyze the progression and project management process of the sprint. Actionable project risk and management tasks from each sprint will be used to evaluate the overall progress rate of the project, and identify processes that work well for the development team. This content is summarized in the Project Schedule section following, which will be continuously updated in future stages of the project.

Lastly, the Team Content section includes both individual and team statements on contributions and progress of the project, as well as on feedback of the Peer Review ITP Metrics reports.

2. Budget Proposal

The proposed project aims to deploy a client-server style website such that users can view live natural disaster related news content. The user interface of the website must be able to direct and manage user requests loads from the server end of the website, which will be responsible for scanning for new disaster-related Tweets, running them through the classification model, and returning the outputted news information to the user upon request. As such, user load management and database hosting will be key components of the project

Given the \$200 budget allocation for the ENG 4000 capstone, this project proposes that the budget will be allocated primarily for deploying the website using services such as Amazon Web Services (including services such as MongoDB, EC2 instance hosting, Docker Hub container management).

A detailed outline of the proposed budget allocation is as follows:

- \$100 for the initial website deployment costs (e.g. AWS credits) targeted at the general public (requires larger amount of load management resources)
- \$50 for project extension website deployment targeted at first responders (e.g. AWS)
- \$50 reserved for miscellaneous or additional dataset acquisition costs (e.g. outsourcing of Tweet dataset gathering and labelling)

3. Key Stakeholders

MVP Stakeholder: General Public

As stated previously, a key deliverable of the project is a website to present reliable natural disaster information extracted from Tweets in a manner that is consumable by the general public. The best performing prediction model produced in the current MVP stage is the Logistic Regression with Bag of Words and lemmatization model, producing a cross-validated F1-score of 0.765. While this score can still be improved by tuning the model, it produces a classification label that the team believes fulfills the general needs of the current stakeholder. Next steps will be to deploy a website that scans live Tweets for up-to-date natural disaster information, runs it through the classification model, and aggregates these labels to output reliable natural disaster information in a manner that the general public can consume. However, there is a chance that a negative unintended consequence can occur because any incorrect information can lead to unnecessary panic.

Stakeholder Refinement for Future Stages: First Responders

In the extension of the initial website, the project seeks to target first responders as a key stakeholder. Because the nature of first responder activities is such that they are highly time-sensitive and important, a higher F1-score will be necessary to justify actionable information (as this stakeholder will have a lower acceptable threshold for false positive or false negative disaster information). For first responders, even marginal situational information can aid their rescue operations such that one or many lives can be saved, or catastrophic damages prevented. However, a negative unintended consequence can be that incorrect information can lead to wastage of precious resources and rescue operation time. As such, the current classification models need to be further improved in future stages of the project.

A challenge will also be documenting and conveying the process of building the classification model such that first responders can trust the results of the model. To convey the criticality of the proposed project, the methods and results of the model in gathering important natural disaster information will need to be presented to the stakeholder, who is a non-expert in machine learning, such that they will be encouraged to trust the technology and adopt it as actionable information.

4. Social Impact and SDG Goals

The dissemination of fake news on the Internet is currently hugely problematic in that it misleads the public in a far-reaching and lasting manner. The proposed project seeks to automate truthful news extraction from social media posts in a timely manner, such that in times of emergency, actionable information can be accessed by first responders. The importance of this task is that in occurrences of natural disasters, immediate situational information can aid in the highly time-sensitive rescue operations of first responders. In disaster situations, even a marginal advantage could result in one or multiple lives being saved, or the prevention of catastrophic damages.

As an engineering capstone, one of the key project goals is to ensure the maintenance of the SDG goals. The following sections detail some of the ways that the proposed project will benefit sustainable development.

Goal #3: Good Health and Well-Being. *Ensure healthy lives and promote well-being for all at all ages.*

By extracting reliable and useful natural disaster information from the mass amounts of information available on Twitter, the project aims to prevent the loss of life and increase the standard of living. The project mainly targets countries in which most of the population has Internet access readily available to them, but future extensions of the project could expand beyond the scope of the project defined in this document. The goals of this project and any project built off this one are such that healthy lives and the well-being of all people can be ensured through the dissemination of reliable, useful, and actionable crowd-sourced information.

Goal #8: Decent Work and Economic Growth. *Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all.*

The target stakeholders, first responders, are a fundamental component in ensuring protection and service of society and the economy. As such, the project goal to provide first responders with timely and actionable emergency response information is such that in times of natural disaster or crisis, disruption to human life and society can be minimized.

Goal #14 & #15: Life Below Water & Life on Land. *Conserve and sustainably use the oceans, seas and marine resources for sustainable development. Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss.*

By extracting truthful natural disaster information in a timely manner, a key objective of the project is that in situations of natural disaster, destruction to natural resources and natural life both below water and on land can be prevented and minimized. The dissemination of truthful information can aid in efforts to protect and restore both human and natural ecosystems.

Goal #16: Peace, Justice, and Strong Institutions. *Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels.*

The dissemination of truthful information from extractions of social media posts is an attempt to provide a service that is accountable and inclusive with the natural disaster information that it provides. A current weakness of the project scope is that its inclusivity is limited to where the general public and first responders have access to available Internet service. However, the project benefits attempt to overcome the malicious effects of fake news dissemination, and therefore aims to ensure accountable institutions and sustainable development goals.

5. Product Backlog

The product backlog consists of requirements for the project that are needed to be accomplished throughout the project. For each requirement, there is a priority associated with it.

The priority used for this backlog is as followed:

High = crucial task

Medium = important task

Low = desirable task

Legend



Completed PBIs

PBI#	Product Backlog	Priority
1	Introductory tutorials and familiarization with the Kaggle challenge	High
2	Begin research about natural disasters in relation to tweets	Medium
3	Research about existing projects related to our project	Low
4	Familiarize with common ML classifiers and metrics	High
5	Familiarize with NLP approaches and techniques	High
6	Familiarize with Tweet format and metadata	Medium
7	Familiarize with Kaggle test dataset	Medium
8	Implement first model to generate predictions on Kaggle dataset	High
9	Review performance of first model	High
10	Implement and combine hand-crafted preprocessing features with classification models	High
11	Implement an LSTM prediction model	High
12	Compare model performance	High
13	Survey current stakeholders to extract social media and news content sentiment	High
14	Discuss social need and scope of project	High
15	Find additional datasets on natural disaster tweets	High

15	Find additional datasets on natural disaster tweets	High
16	Clean and label new datasets	Medium
17	Test the new dataset on model	High
18	Combine classification models (e.g. combine non-neural models with LSTM models) to improve the model	High
19	Learn about web development and familiarize with deployment techniques	Medium
20	Implement data extraction feature	High
21	Build the website	High
22	Deploy the ML model on the website	High
23	Retrieve live Tweets related to natural disaster from Twitter and add to the model	High
24	Survey current stakeholders to get feedback on the website	High
25	Improve the model/website based on stakeholders feedback	High
26	Analyze the social impact of the project	High
27	Expand stakeholders and learn about their needs (i.e. First Responders)	High
28	Design methods to further improve the classification models based on new data	High
29	Develop tools to allow semi-autonomous maintenance of the website	High
30	Analyze SDG and the social impact of this project	High

Table 1: Product Backlog

6. Sprint 1

6.1. Sprint Planning

Sprint Planning Zoom Meeting - November 1st

Since the team is relatively new to working together and is looking to plan the first agile sprint, it is difficult to forecast sprint velocity. Therefore, before planning the initial sprint, the team each filled out the following table and reviewed the first ITPmetrics report before the sprint planning meeting. Doing so helps the team identify both individual and team strengths and weaknesses, as well as analyzes the familiarity of each team member with the project topics, such that a sprint plan can be more effectively created based on these metrics.

The confidence levels used to fill out the pre-sprint plan meeting are as follows:

- 1 = No experience
- 2 = Limited experience
- 3 = Fair experience
- 4 = Intermediate
- 5 = Expert

Team member	Machine Learning	NLP	Python	Twitter
Binte	2	2	3	2
Neena	2	2	3	2
Jessie	2	2	3	3
Jonas	2	1	2	2
Paul	2	1	2	1

Table 2: Self Perceived Expertise

During the sprint planning meeting, the team discussed project management tools that will be used during the sprint. Reviewing the ITPmetrics report, it was clear that managing the varying schedules of the team would be a challenging task since each team member would be balancing busy work, academic and social responsibilities. It will therefore be fundamental during sprint planning processes to set clear product features to be developed during each sprint, and to accurately review the performance of the team at the end of each sprint.

Furthermore, review of Table 1 finds that as a whole, the team required more experience in many of the fundamental project topics. To overcome this, the following tasks are outlined for Sprint 1:

- Review the general practices for machine learning regarding natural language processing (e.g. Classifiers = Naive Bayes, Decision Trees, SVC, and classification techniques = Bag of Words, Lemmatization, Stop Words) (**Week 1 of Sprint 1**)
- Review Tweet format and metadata, as well as the Kaggle dataset (**Week 1 of Sprint 1**)
- Implement a first prediction model on the Kaggle dataset and review performance metrics of the classification results (**Week 2 of Sprint 1**)

Given the two week timeline allocated for Sprint 1, the feasibility of accomplishing these tasks was discussed during the Sprint Planning meeting. The team is confident that given clear roles and tasks during Sprint 1, the above mentioned Sprint goals can be completed.

The roles and timeline for Sprint 1 as defined by the team are discussed in following sections, a summary of the sprint progress is discussed, and finally the performance and productivity of Sprint 1 is reviewed and a refinement of the process and tasks for future sprints is examined.

6.2. Timeline

Sprint 1 Duration: November 2nd - November 16th

Week 1 of Sprint 1: November 2nd - November 9th

Week 2 of Sprint 1: November 9th - November 16th

6.3. Roles

The first sprint roles are defined as follows:

Jessie Leung: Product Owner
Neena Govindhan: Scrum Master
Binte Zehra: Development Team
Jonas Laya: Development Team
Paul Sison: Development Team

6.4. Sprint Goal(s)

The sprint goals for the first sprint were as follows:

- Review research papers similar to the project to get familiar with Machine Learning tools and techniques
- Review Tweet format and metadata, as well as, the Kaggle dataset.
- Implement the model using different Machine Learning Classifiers: Naive Bayes, Decision Tree, SVC to find the best one, giving reasonable performance
- Furthermore, implement the model using classification techniques (Bag of Words, Lemmatization, Stop Words) to find better results
- Implement a first prediction model on the Kaggle dataset.

6.5. Sprint Backlog

Sprint Backlog	To-Do (For Sprint 2)	In Progress (For Sprint 2)	Done	Work Estimation
PBI #1: Familiarize with common ML classifiers and metrics	-	Learn about commonly used deep learning models (LSTM) and study implementations using libraries such as keras.	Learn about common classifiers (Naive Bayes, Decision Tree, Random Forest, SVC) and performance metrics (Recall, precision, F-scores). Study sample code of classifier implementations using libraries such as scikit-learn.	Medium
PBI #2: Familiarize with NLP approaches and techniques	-	-	Learn about NLP techniques (Bag of words, lemmatization, stop words). Study sample code of implementations of these techniques in existing NLP classification models.	Large
PBI #3: Familiarize with Tweet format and metadata	-	-	Familiarize with Tweet format and metadata such as followers, retweets, likes, hashtags	Small
PBI #4: Familiarize with Kaggle test dataset	-	-	Look through Kaggle dataset and familiarize with data columns	Small
PBI #5: Implement first model to generate predictions on Kaggle dataset	Implement a model using deep learning model (e.g. LSTM).	-	Implement an initial prediction model to generate predictions for the 'test' dataset using common classifiers. Implement a model with cross-validation technique to run on a 'train' dataset.	Large
PBI #6: Review performance of first model	-	-	First model predictions submitted to Kaggle test to review accuracy of predictions against actual labels. Review performance of cross-validated recall, precision and f-scores on 'train' dataset.	Medium

Table 3: Sprint 1 Backlog

6.6. Tasks

PBI #1:

A study of research papers (as summarized in Appendix A) concerning commonly used classifiers shows some classification models that are effectively used for text classification¹:

- Non-neural network models: SVM, Naive Bayes, Logistic Regression, Decision Trees, Random Forest (many included in scikit-learn library)
- Neural network models: Long Short-Term Memory, Convolution Neural Networks (many included in keras library)

Furthermore, the metrics that are most commonly used to evaluate the performance of the models use methods of cross-validation (e.g. Stratified K-folds as supplied by the scikit-learn library) on test datasets reviewed by the following performance measures, where tp is a true positive (Tweet is classified as a valid Tweet where it is truly valid), fp is a false positive (Tweet is classified as valid when it is actually invalid), fn is false negative (where Tweet is classified as invalid when it is actually valid)²:

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$Fscore = \frac{2 * precision * recall}{precision + recall}$$

Figure 1: Formulae to calculate Fscore

The team came up with research questions, which helped with coming up with product features. It was useful in terms of structuring the features in an efficient manner. The research questions were divided into four categories, descriptive, diagnostic, predictive, and prescriptive.

Example research questions from each category:

Descriptive (observative)

Q1: What is the relation between the type of disaster (earthquake, wildfire, etc.) and the number of fake tweets?

- NULL HYPOTHESIS (1_1): there is no significant difference between the distribution of the number of tweets for different emergencies.
-- test: ANOVA, Kruskal-wallis, Chi-Square

Diagnostic

Q8: What factors prevent first responders to trust machine generated solutions?

¹ Oshikawa, R., J. Qian, and W. Y. Wang.

² Shu, K., A. Sliva, S. Wang, J. Tang and H. Liu.

- How? --- Qualitative research – interview/survey

Predictive

Q11: What is the Likelihood of significantly different location of fake tweets if the disaster happens over night?

Prescriptive

Q12: How do we leverage a machine learning model that can identify authentic disaster tweets into a product/service for use of emergency and disaster response organizations?

PBI #2:

To address the team's general unfamiliarity with machine learning models, the process of building a model for natural language processing was studied. Upon review of multiple research papers on the topic of NLP and building predictive models, a summary of the NLP process first consists of preprocessing the given dataset into extracted features³. Data preprocessing involves techniques such as tokenization, lemmatization, and the use of a stop-words list. Exploring the scikit-learn API finds that built-in methods for these techniques are included (e.g. `sklearn.feature_extraction.text.CountVectorizer` includes parameters for tokenization and stop-words). These are NLP techniques that will likely be useful in the project's implementation of an NLP-based classification model.

PBI #3:

Familiarizing with Twitter and structure of tweets. Analyzing tweets in terms of its text, vocabulary, keywords used in it, number of times it has been retweeted, followers/followees relationship. After reviewing various research papers, there were some important takeaways that were useful when analyzing Twitter data:

- Methodologies
 - Analyzing Twitter within hours and days after events
 - Examine how rumors and news are propagated in relation to followers/followees
 - Correlation of key terms in various tweets
- Techniques
 - Feature Extraction
 - Model Construction
 - Performance Evaluation
 - Coding scheme- for categorizing disasters
- Tools
 - Latent Dirichlet Allocation (LDA) topic modelling
 - Tweet2Vec - Tweet semantic similarity tool
 - Non-neural network models: SVM, Naive Bayes, Logistic Regression, Decision Trees, Random Forest
 - Neural network models: Long Short-Term Memory, Convolution Neural Networks

³ Oshikawa, R., J. Qian, and W. Y. Wang.

PBI #4:

The team analyzed the datasets provided by the Kaggle challenge. The test dataset consists of four columns: id, keyword, location, and text. Similarly, the training dataset consists of the same data column, with the addition of a target data column. The classification task prescribed by the Kaggle challenge is a binary classification problem, wherein the target label indicates the validity of a Tweet given the following labels:

- 1:** about a real natural disaster
- 0:** not about a natural disaster

By observing the structure of the Kaggle dataset, the primary focus of the initial model was decided to use data from the 'text' column to predict the value of the corresponding 'target' column.

PBI #5:

For the first sprint, building a simple classification model implemented SVC, Decision Tree, Random Forest, Logistic Regression, and Multinomial NB classifiers using the following approach for building a machine learning model. The training dataset is split using a 10 stratified k-fold method, such that nine folds are used for training and one fold used for testing.

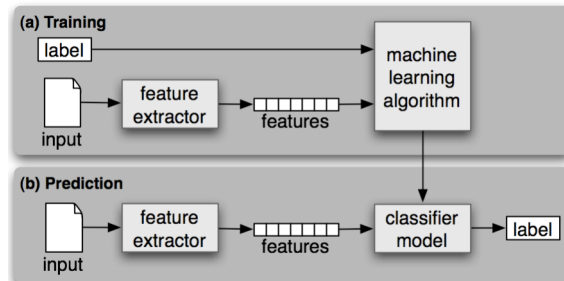


Figure 2: Supervised classification process from NLTK⁴

In the feature extraction process, the 'text' data is preprocessed by using techniques such as tokenization, lemmatization and the removal of stop words to compose a Bag of Words in the form of a vector. This feature is then trained using a classifier from the scikit-learn library, and outputs prediction labels that are then verified using the 'target' column.

⁴ "Learn to classify text." NLTK. <https://www.nltk.org/book/ch06.html>

PBI #6:

Classifier	Precision	Recall	F1
SVC			
BOW + Stopwords	0.70	0.70	0.68
BOW + Lemmatization + Stopwords	0.71	0.70	0.69
Decision Tree			
BOW + Stopwords	0.57	0.58	0.57
BOW + Lemmatization + Stopwords	0.61	0.61	0.61
Random Forest			
BOW + Stopwords	0.63	0.64	0.61
BOW + Lemmatization + Stopwords	0.67	0.66	0.64
Logistic Regression			
BOW + Stopwords	0.65	0.66	0.65
BOW + Lemmatization + Stopwords	0.66	0.66	0.66
Multinomial NB			
BOW + Stopwords	0.67	0.67	0.67
BOW + Lemmatization + Stopwords	0.69	0.69	0.69

Table 4: Results of classification techniques for cross-validated training set using 10 stratified k-folds

7. Sprint 1 Review

Date: November 17th

7.1. Completeness

The team was able to successfully complete Sprint 1 on time. During week 1 of sprint 1, the team familiarize themselves with different Machine Learning tools and techniques, classifiers (Naive Bayes, Decision Tree, SVC, etc.), and classification techniques (Bag of Words, Lemmatization, Stop words, etc.). Though there is plenty of material available, the team decided to review the basics and begin week 2 of sprint 1. In week 2, the team completed sprint 1 with implementing a first prediction model on the Kaggle dataset and compared results of performance metrics of different classifiers.

7.1.1. Burndown Chart

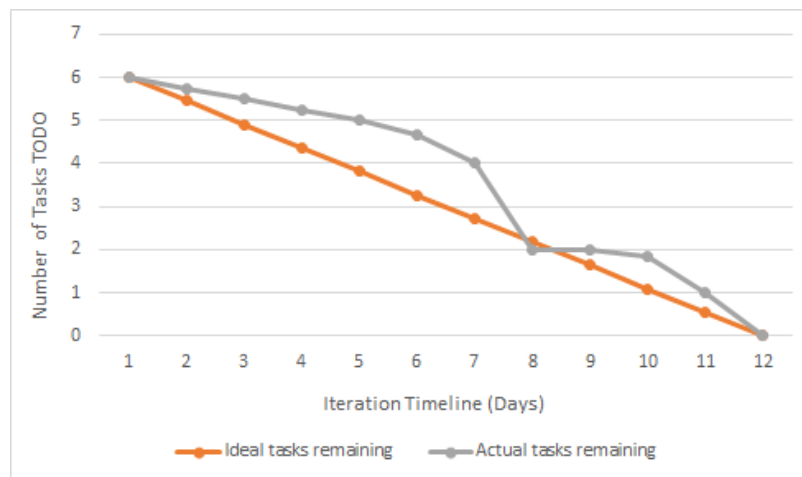


Figure 3: Burndown Chart

The goal of this Sprint was mainly to learn about Machine Learning (ML) and Natural Language Processing (NLP) well enough so the team can implement an initial model. Since all the team members are eager to learn about the project topic/domain and everyone wishes to be heavily involved in the development process, the tasks for this Sprint cannot be delegated to only a few members. PBIs 1 and 2 took more time than expected and were the first tasks to be completed. PBIs 3 and 4 were trivial tasks and only took a day to complete. And due to other course commitments and difficulties in setting up our coding environments, the team was delayed by a few days in implementing the model.

7.1.2. Velocity

Having completed 6 PBI tasks in a span of 2 weeks, the team's velocity for this Sprint is 3 PBI tasks per week. A velocity graph will be provided in further reports when the team have completed a few more Sprints.

7.2. Challenges

The challenges in the first sprint occurred when the team started to code the model. The coding had not been attempted previously, and the team was more focused on planning and research about the problem. However, the knowledge acquired through the research and course work had helped in creating the initial model. With Python, there were some initial hesitations since the team as a whole was a little inexperienced with it, but it was actually much more simple to implement than expected. There was some trouble with managing other courses, so some tasks took longer than expected, but it was good that the sprint backlog was not overloaded so the team managed to get through in time.

8. Sprint 1 Retrospective

Date: November 18th

8.1. Happiness Evaluation Metric

To measure the overall team satisfaction levels with Sprint 1, a Happiness graph is used to visually identify the sprint events that team members thought worked effectively towards the product goals. The following scale was used as a metric for the team to measure their 'happiness' levels with each sprint event.

Happiness level	1 😞	2 😞	3 😞	4 😊	5 😊
Criteria	Unhappy with effectiveness of sprint activity, very demotivated	Slightly demotivated, feel that more could have been done with sprint activity	Sprint activity was fair, neutral motivation levels	Generally happy with sprint event, motivation and momentum levels fair	Happy with sprint event, feeling motivated and happy with sprint momentum

Table 5: Happiness Evaluation Metric

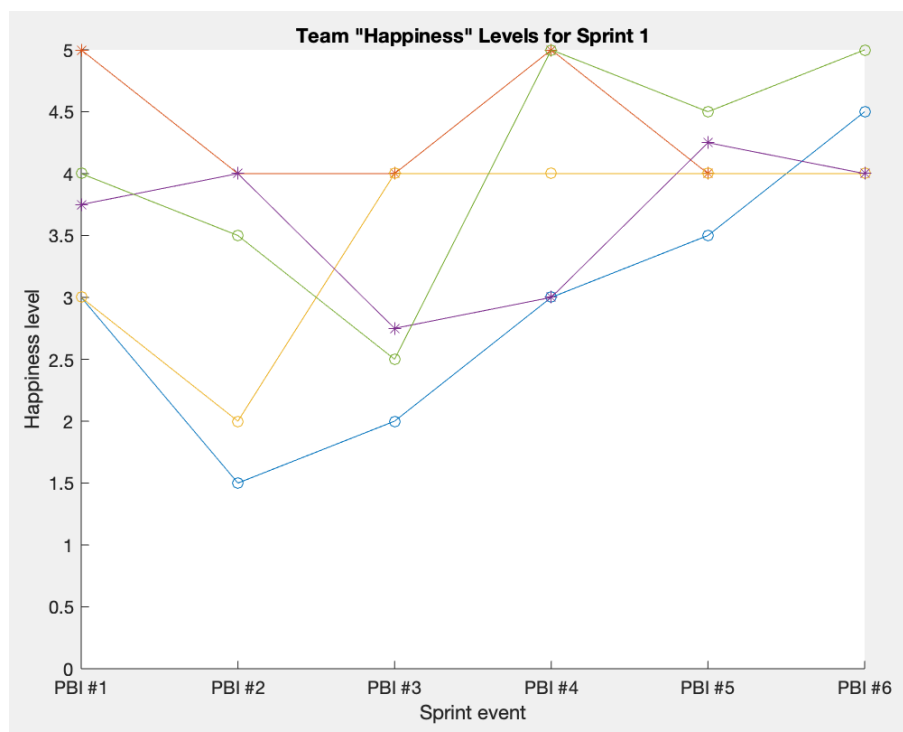


Figure 4: Sprint 1 Happiness Graph

During the Scrum Master-led Sprint Retrospective, the effectiveness of Sprint 1 was discussed with members of the team. Sprint activities that had generally lower happiness ratings were discussed to address what aspects of the event could have been improved. Similarly, generally positively rated events were discussed to identify why team members felt that that activity significantly aided the progression of the project. The findings of the retrospective discussions are detailed in the following sections.

8.2. Process Related Issues

A discussion of the collaboration process during Sprint 1 resulted in the following actionable procedure changes such that following sprints can be streamlined and more effective.

Sprint Event	What Went Wrong	What Went Well	Actionable Procedure Objectives
PBI #2	Access to mass amounts of academic papers felt overwhelming at times, many topics to learn since the team is relatively new to ML. Difficulty understanding how to bridge ML/NLP learning with code implementation.	Reading through many academic papers, found techniques and classifiers that were commonly used throughout most papers. Clear that these were topics that the team should focus on. Helped clarify NLP techniques conceptually. Research papers split up amongst team members and summarized in PowerPoint documents presented amongst team members.	Continue to divide learning workload amongst team members, generate documentation and make accessible amongst members of the team to facilitate overall team learning (documentation available on MS teams, presentation of academic learning such that the entire team benefits).
PBI #5		Team began learning once the process of implementation began. Concepts that were less clear in previous sprint events (NLP & ML concepts) became easier to understand once in the process of implementing.	Pair academic learning with implementation activities to validate conceptual learning.
PBI #6		Overall sentiment of progress since results (cross validation scores, testing with Kaggle prediction submissions) helped validate that learning done in previous sprint events was implemented to a satisfactory level.	Continue to generate testable implementations (via cross-validation, Kaggle submissions) and improve on scores from previous models.

Table 6: Sprint 1 Process Related Issues

9. Sprint 2

9.1. Sprint Planning

Sprint Planning Zoom Meeting - November 22rd

Following the completion of the first sprint process, as well as its associated review and retrospectives, the team initiated the sprint planning phase for Sprint 2. From the initial sprint, it was determined that the team worked well when learning individually, then coming together as a team to implement collaboratively. Furthermore, the team overall felt more productive when their work produced a working product (for example, the Kaggle submission and cross-validation scores from Sprint 1).

The overall goal of Sprint 2 is to create a Minimum Viable Product for the ENG 4000 mid-term gate submission. As such, the activities of the sprint were two-fold. The first activity of the sprint focused on improving the previously implemented prediction model to a threshold of satisfaction of the team. Sprint planning discussion with the members of the team found that team members were interested in implementing an LSTM Neural Network model for MVP, since academic research from Sprint 1 found this classification model to produce high accuracy scores for other datasets. Furthermore, another goal was to gather the cross-validation and other testing scores for the generated models and present the information using figures and visualization tools in a manner that would be consumable to a wider audience, including those less familiar with the ML process. This also lends to the second goal of Sprint 2, wherein the societal aspect of the project would be further explored. Here, the goal is to elicit the social need and impacts of the project, as well as to discuss the future roadmap for the project beyond the mid-term submission.

Given the two week time-frame of Sprint 2 and the progress rate from Sprint 1, the team was confident that these tasks could be accomplished by setting out a timeframe and planning out tasks and collaborative sessions.

9.2. Timeline

Sprint 2 Duration: November 23rd - December 7th

Week 1 of Sprint 2: November 23rd - November 30th

Week 2 of Sprint 2: November 30th - December 7th

9.3. Roles

The second sprint roles are defined as follows:

Binte Zehra: Product Owner

Jessie Leung: Scrum Master

Neena Govindhan: Development Team

Jonas Laya: Development Team

Paul Sison: Development Team

9.4. Sprint Goal(s)

The sprint goals for the second sprint are defined as follows:

- Improve performance of Machine Learning models by hand-crafting data preprocessing features
- Implement additional classification models by utilizing Neural Network models (LSTM)
- Compare performance of all models using metrics such as recall, precision, f1-score and accuracy of Kaggle submission of the testing dataset
- Explore social aspect of the project by surveying current stakeholders (general public) and discussing stakeholder refinement for future stages of the project
- Better define social need and scope of the project at both MVP stage and in future stages

9.5. Sprint Backlog

Sprint Backlog	To-Do (For Sprint 3)	In Progress (For Sprint 3)	Done	Work Estimation
PBI #1: Implement and combine hand-crafted preprocessing features with classification models	Continue to improve hand-crafted text preprocessing features for social media posts data (e.g. normalization of emojis).	-	Replaced built-in preprocessing methods with manual text preprocessing to better suit Kaggle dataset.	Medium
PBI #2: Implement an LSTM prediction model	Combine classification models (e.g. combine non-neural model with LSTM model).	-	Implemented an LSTM classification model.	Medium
PBI #3: Compare model performance	-	-	Create tables comparing performance of different classification model. Use visualization tools (e.g. Confusion Matrix) to present classification results in a consumable form.	Medium
PBI #4: Survey current	-	-	Created and distributed an online survey to elicit	Small

stakeholders to extract social media and news content sentiment			general public familiarity with Twitter and sentiment towards news content retrieved from social media.	
PBI #5: Discuss social need and scope of project (MVP and future stages)	Continue to research and discuss the societal project scope in the context of future releases to elicit requirements.	Analyze survey results and discuss direction for stakeholder refinement beyond MVP stage.	Discuss social impact and scope of project at its current stage of development.	Large

Table 7: Sprint 2 Backlog

9.6. Tasks

PBI #1:

To improve the performance of the classification models, hand-crafted text preprocessing features replaced the standard built-in models used in the first model. An in-depth view of the pipeline used for the model built during this task can be found at:

https://colab.research.google.com/drive/1TSDbvVCh15fkBthKsLcqFOW0Qkayka_v?usp=sharing

While the previous models utilized only a Bag of Words feature extraction technique, the team decided to experiment with two additional feature extraction techniques during the second sprint:

- Bag of Words
- Term Frequency-Inverse Document Frequency
- Word Embeddings

Some of the manual text cleaning features implemented include:

- Expanding contractions
- Emoji removal
- HTML tag/URL removal

Beyond the standard tokenization, stemming and stopword removal implemented in the first models, the extension during the second sprint also included testing multiple custom tokenizers, as well as a word lemmatizer. As with models implemented in Sprint 1, evaluation metrics on the training dataset are determined using a 10 stratified k-fold cross-validation technique.

In the weekly discussions with the team supervisor, the model extensions implemented during the first week of Sprint 2 were presented. Potential future steps were discussed, such as the possibility of incorporating features such as emojis in the model analysis, due to their

relation to conveying sentiment. These possible additions to the classification model were noted as steps to be implemented in future sprints.

The results of the implemented models are presented in PBI #3.

PBI #2:

In the academic reading completed in Sprint 1, team members noted that across multiple papers and datasets, text classification models using a neural network LSTM approach appeared to produce high accuracy scores across the metrics introduced in Figure 1 (Recall, Precision, and F1-score). As such, the team implemented an LSTM model using the TensorFlow Keras library to run on the Kaggle dataset. A full view of the LSTM pipeline used can be found at:

<https://colab.research.google.com/drive/1guDExEyF51KPSQGxKM0V1UrQWU8ZEsP?usp=sharing>

The team also implemented another LSTM model that utilizes word embeddings pre-trained on Twitter posts. The full implementation of the model is also be viewed on Google Colab at:

<https://colab.research.google.com/drive/1auQrS7tKjo9m1F8mY0dmNnjPHeaQvw99?usp=sharing>

A similar preprocessing pipeline as used for the non-neural network models was included in the LSTM model, and the results of the LSTM models on the cross-validated Kaggle training dataset are as follows.

	Precision	Recall	Accuracy	F1-score
Without word embeddings	0.765	0.713	0.765	0.717
With word embeddings	0.822	0.720	0.814	0.767

Table 8: LSTM Models Score

PBI #3:

The cross-validated results of the non-neural network models implemented in PBI #1 are summarized as follows:

	Bernoulli Naïve Bayes		Multinomial Naïve Bayes		Stochastic Gradient Descent		Support Vector Machines		Logistic Regression		Decision Trees		Random Forest	
	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.
Bag of words														
Simple preprocessing	0.751	0.805	0.759	0.803	0.739	0.787	0.739	0.783	0.755	0.802	0.692	0.746	0.726	0.795
- stopwords	0.749	0.805	0.758	0.799	0.734	0.785	0.736	0.784	0.754	0.805	0.703	0.756	0.732	0.795
+ lemmatization	0.754	0.806	0.757	0.802	0.740	0.786	0.741	0.784	0.765	0.810	0.701	0.751	0.733	0.799
- stopwords + lemm	0.754	0.809	0.756	0.798	0.739	0.787	0.738	0.785	0.755	0.803	0.709	0.760	0.728	0.790
- stopwords + stem	0.755	0.809	0.758	0.799	0.732	0.778	0.731	0.778	0.750	0.800	0.703	0.756	0.725	0.787
TF-IDF														
Simple preprocessing	0.751	0.805	0.733	0.800	0.757	0.807	0.752	0.801	0.752	0.810	0.674	0.725	0.710	0.788
- stopwords	0.749	0.805	0.744	0.802	0.750	0.802	0.743	0.796	0.740	0.805	0.713	0.756	0.726	0.790
+ lemmatization	0.754	0.806	0.736	0.802	0.759	0.808	0.754	0.802	0.753	0.810	0.682	0.731	0.718	0.790
- stopwords + lemm	0.754	0.809	0.745	0.803	0.749	0.799	0.743	0.795	0.742	0.805	0.707	0.751	0.724	0.788
- stopwords + stem	0.755	0.809	0.746	0.804	0.745	0.796	0.741	0.792	0.740	0.802	0.702	0.748	0.728	0.791
Word embeddings														
Simple preprocessing	0.678	0.711			0.749	0.798	0.755	0.800	0.754	0.799	0.648	0.699	0.742	0.798

Figure 5: Non-neural Network Models Results

Comparing the results of the non-neural network models (Figure 5) wherein the best performing classifier results are highlighted in green and the worst performing classifier is highlighted in red) and the LSTM model (Figure 5), the Logistic Regression using Bag of Words with lemmatization produces the highest F1-score. Overall, compared to the results of the initial models implemented in Sprint 1 (summarized in Table 3), the hand-crafted feature extraction and text preprocessing techniques increased the F1-score of all the classifiers previously implemented by a significant margin.

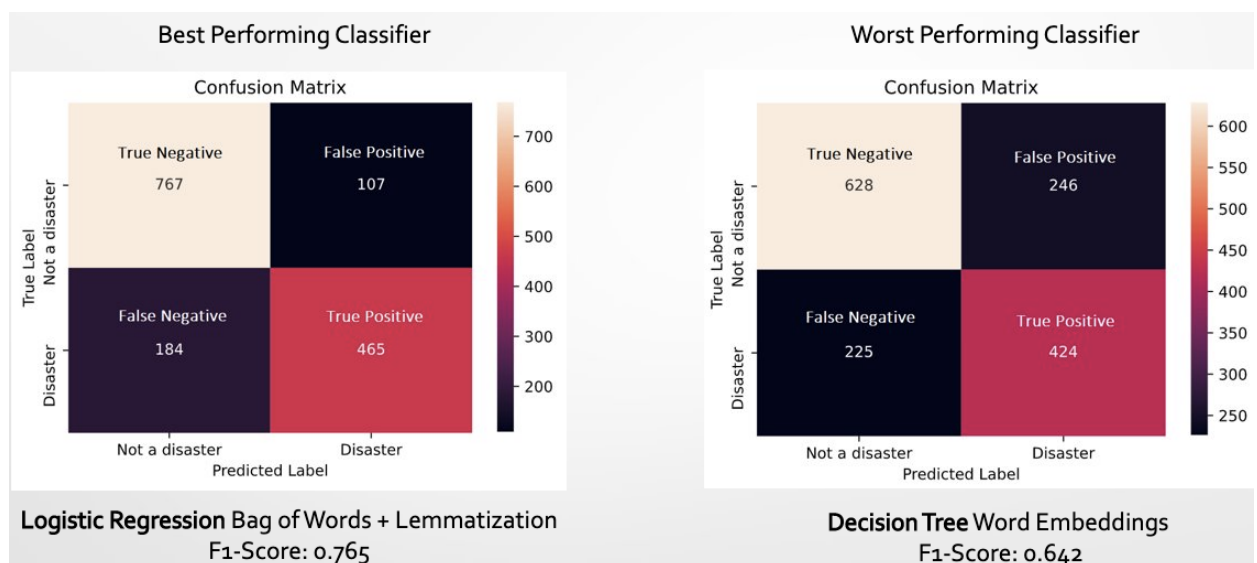


Figure 6: Confusion Matrix for Best and Worst Performing Classifier

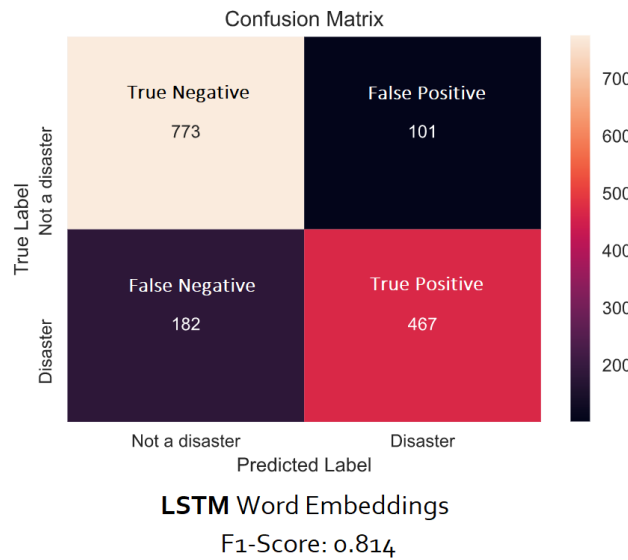


Figure 7: LSTM Confusion Matrix

In the figures above, the corresponding confusion matrices are presented for the Logistic Regression, Decision Tree, and LSTM classification models. These figures present the performance of the individual models using a visualization tool such that the information can be more easily consumed by both machine learning experts and non-experts. These visualization tools will be helpful for Alpha and Beta releases of the project as the prediction models will be further developed in upcoming Sprints.

For the ENG 4000 Mid-term Capstone MVP, the project objective was to complete the Kaggle challenge “Real or Not? NLP with Disaster Tweets” (found at <https://www.kaggle.com/c/nlp-getting-started>). As such, taking all knowledge gained by the team from academic research completed in Sprint 1, and implementing completed in both sprints, two classification models were chosen to output prediction labels on the Kaggle test dataset (unlabelled target column or whether the Tweet is about a real disaster or not), and submitted to the Kaggle challenge for feedback. The accuracy of the target labels generated by the implemented Logistic Regression and LSTM classification models are as follows:

Model used	Accuracy of model predictions on the Kaggle test dataset
Logistic Regression + BoW + lemmatization	0.798
LSTM + word embeddings	0.813

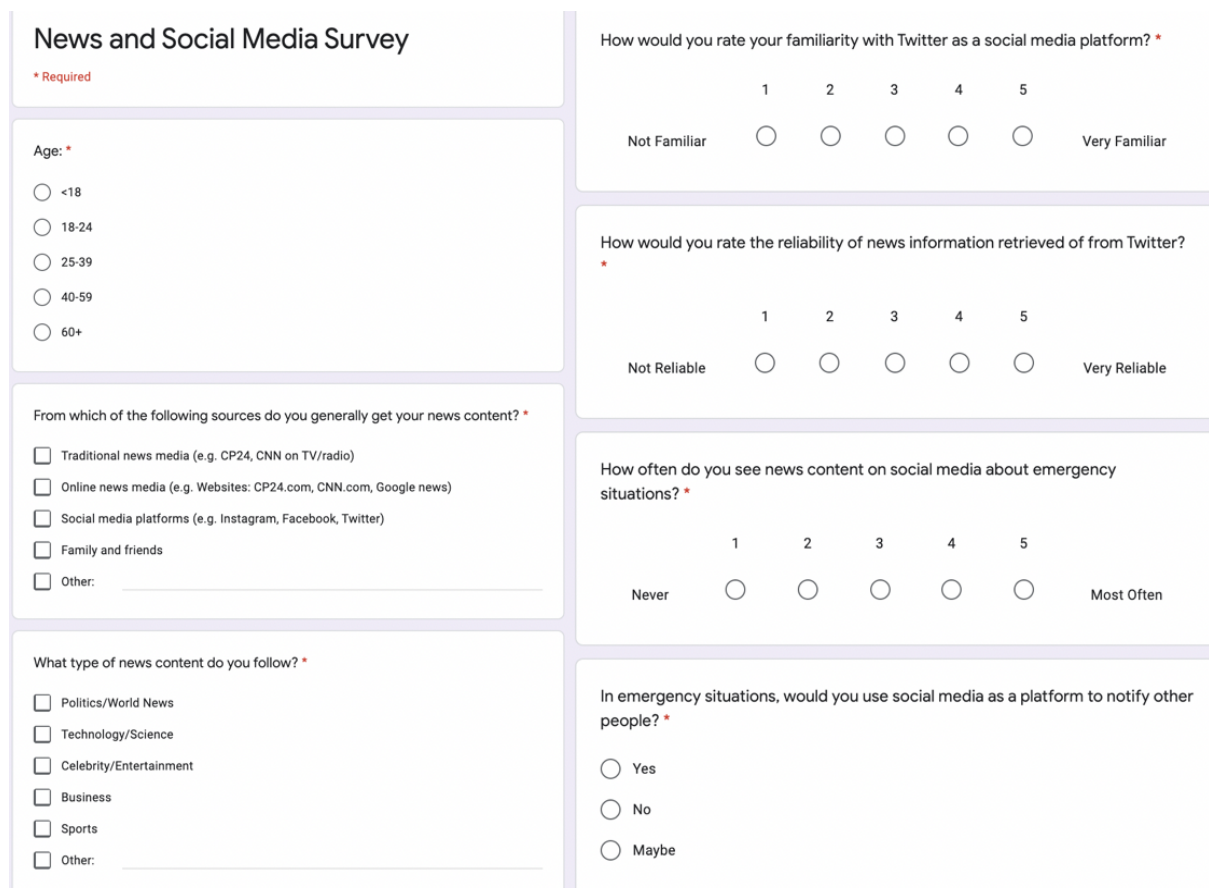
Table 9: Classification Model Results

As expected, the cross-validation scores closely reflected what the Kaggle test results will be. With an accuracy score of 81.3%, the team fulfills the MVP requirement of implementing a model that can score at least 80% on the Kaggle challenge. Moving forward to the winter term, the team will look into utilizing more sophisticated models such as CNNs, hybrid neural networks, transformers, etc. and train them on larger Twitter datasets made available by CrisisLex and CrisisNLP.

PBI #4:

To gain a better understanding of general public interaction with Twitter and sentiment towards news content retrieved from social media, the following survey was created and distributed to members of the general public to anonymously complete:

https://docs.google.com/forms/d/e/1FAIpQLSeM08I-TjoMXw7SkbcZLwgQJU-Im08acSoxhAf_v_wQnryfW5w/viewform



News and Social Media Survey

* Required

Age: *

☐ <18

☐ 18-24

☐ 25-39

☐ 40-59

☐ 60+

From which of the following sources do you generally get your news content? *

☐ Traditional news media (e.g. CP24, CNN on TV/radio)

☐ Online news media (e.g. Websites: CP24.com, CNN.com, Google news)

☐ Social media platforms (e.g. Instagram, Facebook, Twitter)

☐ Family and friends

☐ Other: _____

What type of news content do you follow? *

☐ Politics/World News

☐ Technology/Science

☐ Celebrity/Entertainment

☐ Business

☐ Sports

☐ Other: _____

How would you rate your familiarity with Twitter as a social media platform? *

1 2 3 4 5

Not Familiar ☐ ☐ ☐ ☐ ☐ Very Familiar

How would you rate the reliability of news information retrieved of from Twitter? *

1 2 3 4 5

Not Reliable ☐ ☐ ☐ ☐ ☐ Very Reliable

How often do you see news content on social media about emergency situations? *

1 2 3 4 5

Never ☐ ☐ ☐ ☐ ☐ Most Often

In emergency situations, would you use social media as a platform to notify other people? *

☐ Yes

☐ No

☐ Maybe

Figure 8: News and Social Media Survey

The results of the survey are shown in the figures below.

Age:

32 responses

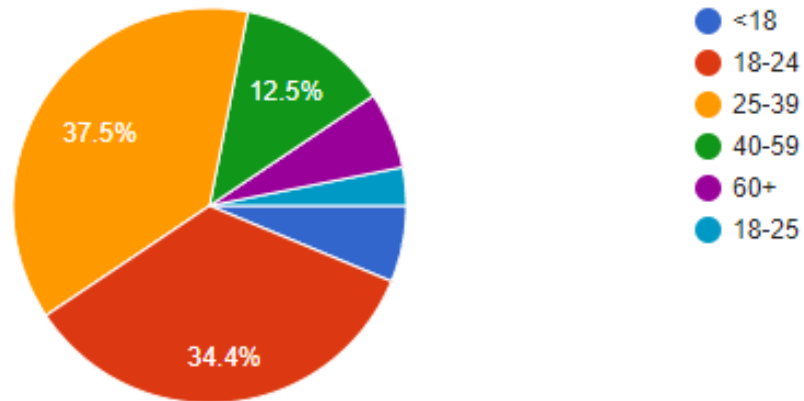


Figure 9: Age Pie Graph

From which of the following sources do you generally get your news content?

32 responses

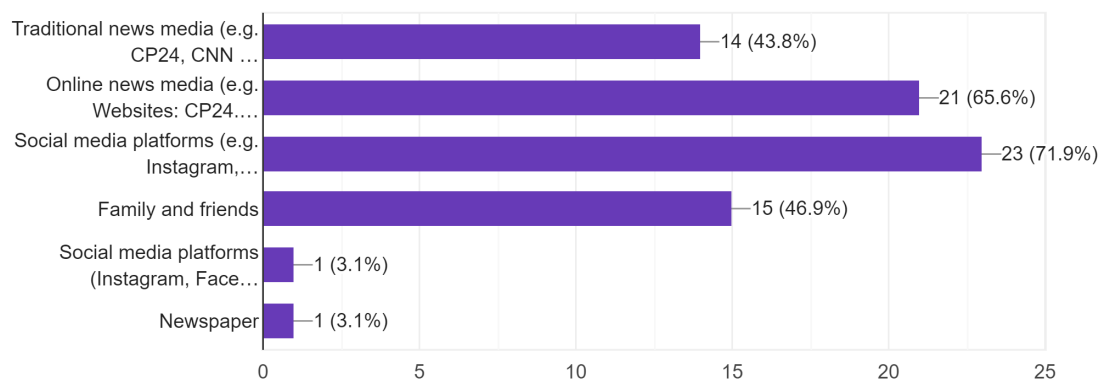


Figure 10: News Source Survey

What type of news content do you follow?

32 responses

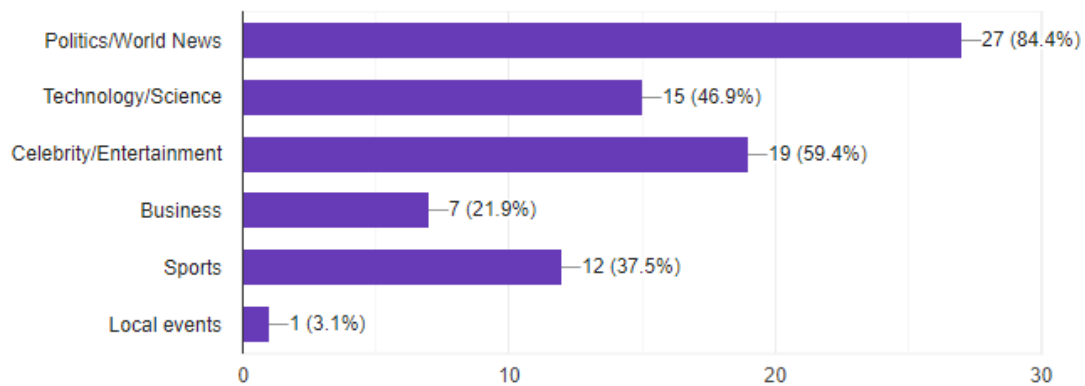


Figure 11: News Content Survey

How would you rate your familiarity with Twitter as a social media platform?

32 responses

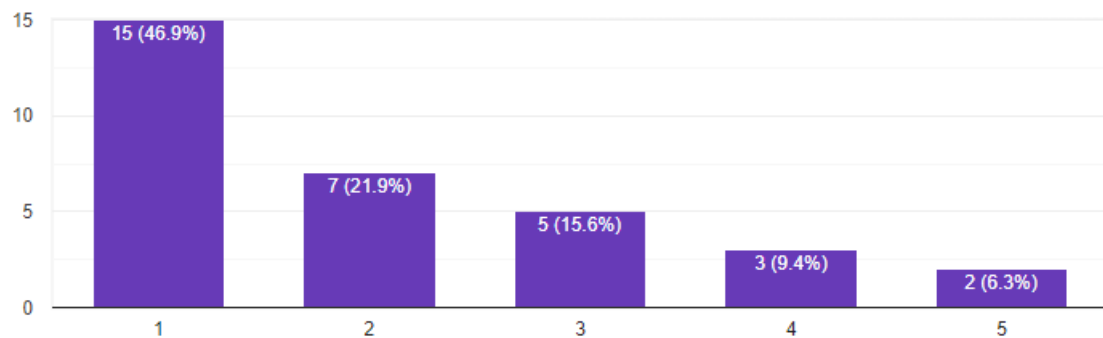


Figure 12: Familiarity with Twitter Survey

How would you rate the reliability of news information retrieved of from Twitter?

32 responses

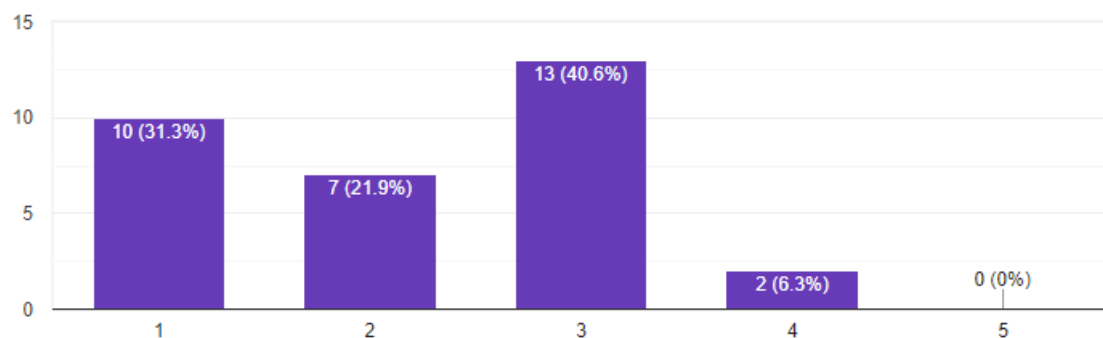


Figure 13: Reliability Survey

How often do you see news content on social media about emergency situations?

32 responses

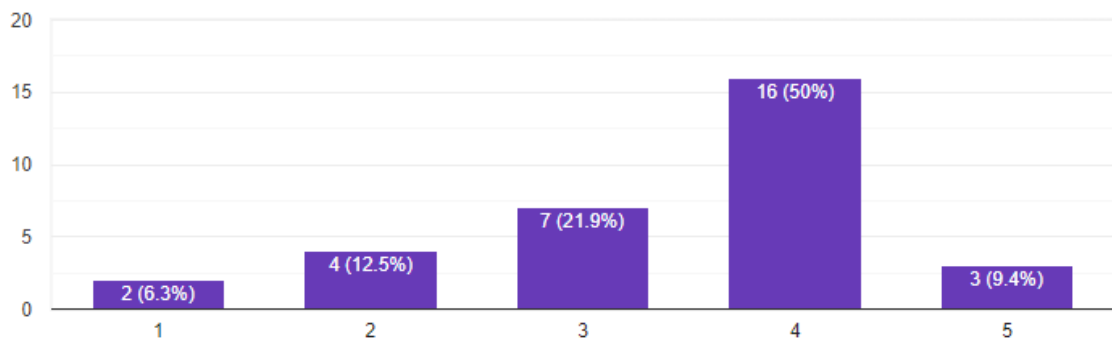


Figure 14: News Content on Social Media During Emergency Survey

In emergency situations, would you use social media as a platform to notify other people?

32 responses

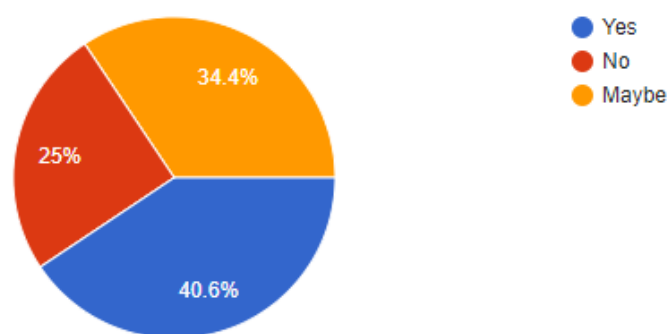


Figure 15: Social Media to Notify People Survey

Reviewing the results of the survey finds that amongst multiple age groups, typically users tend to see news content from social media platforms (Figure 11: News Source). However, the sentiment on the reliability of news content retrieved from Twitter generally ranges from 'Not Reliable' to 'Neutral' (Figure 14: Reliability). 40.6% of survey takers also noted that they would use social media platforms such as Twitter to notify other people of emergency situations (Figure 16: Social Media to Notify People).

This survey therefore reinforces the project's initial hypothesis that while the general public agrees that social media platforms contain important news content, it is difficult to extract useful information from social media posts due to the inherent noisiness (e.g. spam, misleading content, fake news) cluttering social media posts.

PBI #5:

The project thus far had a focus on the technical aspect, but more research into the societal aspect is needed. One step towards this was first done in the initial research papers that were analyzed (Appendix A), however, at that time more focus was given to learning about ML and NLP techniques. For this sprint, the focus has shifted to learning more about the stakeholders, as can be seen in the survey conducted, and determining the SDG(s) the project is focused on.

A summary of the deliverables from this PBI are outlined in the Key Stakeholders and Social Impact and SDG Goals in previous sections of this report.

10. Sprint 2 Review

10.1. Completeness

The PBI tasks planned for Sprint 2 were completed to the satisfaction of the Scrum Master and team. The Sprint Goals included implementing additional classification models to complete the Kaggle challenge, which was accomplished in Week 1 of Sprint 2. Having completed the Kaggle challenge, an in-depth exploration of the social aspect of the project was conducted. This included a breakdown of project impacts on SDGs and a stakeholder refinement for future stages of the project development. Overall, by discussing the progress of Sprint 2 during the Sprint 2 Review meeting, it was determined that the team was happy with the progression of the sprint.

10.1.1. Burndown Chart

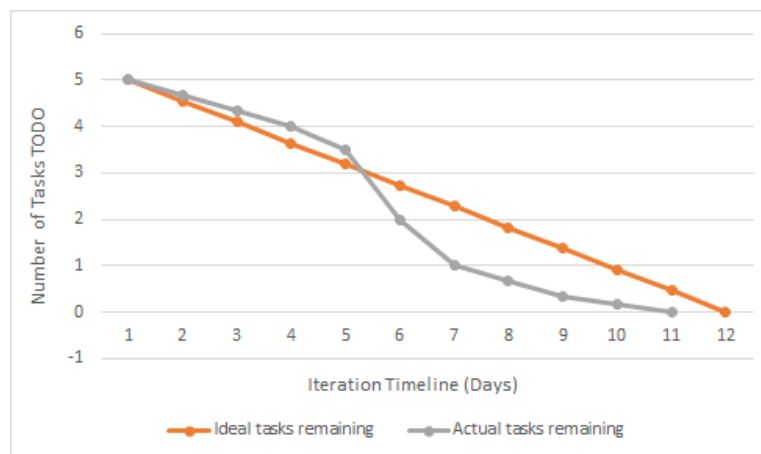


Figure 16: Sprint 2 Burndown Chart

The implementation of the LSTM model (PBIs 1 to 3) scheduled in the first week of the Sprint was completed promptly as evidenced in the Burndown chart above. The experience we gained from implementing our initial model during the first Sprint transitioned smoothly to our coding workflow in the second Sprint. Research on relevant Deep Learning libraries and the LSTM model were quickly followed by implementation and test runs. Exploring the project's societal impact (PBIs 4 and 5) were also completed ahead of schedule.

10.1.2. Velocity

Having completed 5 PBI tasks in 11 days (the team counts 6 working days in a week), the team's velocity for this Sprint is 2.73 PBI tasks per week, which is a slight reduction in efficiency from the previous Sprint. The team, however, finds this understandable as the Sprint was executed during the end of term when other course deliverables are piling and revisions for upcoming exams take some of the team's time. A velocity graph will be provided in further reports when the team have completed a few more Sprints.

10.2. Challenges

The challenges occurred during the second sprint were again to do with not having enough experience with the topic of deep learning. There are plenty of techniques available, but choosing the right one that would improve the prediction model results, was the main focus of Sprint 2. Therefore, after doing a lot of research, the team decided to learn more about the technique of LSTM, since it better suited the teams requirements. Another challenge faced during the week 2 of the second Sprint, was related to the SDGs. Deciding on goals that best relates to the project. As to the social aspect, what impact does the project have on society. Where to begin the and how to tie it the project resulted in challenges.

11. Sprint 3 Retrospective

11.1. Happiness Evaluation Metric

To measure the overall team satisfaction levels with Sprint 2, a Happiness graph is used to visually identify the sprint events that team members thought worked effectively towards the product goals. The following scale was used as a metric for the team to measure their 'happiness' levels with each sprint event.

Happiness level	1 😞	2 ☹️	3 😐	4 😊	5 😄
Criteria	Unhappy with effectiveness of sprint activity, very demotivated	Slightly demotivated, feel that more could have been done with sprint activity	Sprint activity was fair, neutral motivation levels	Generally happy with sprint event, motivation and momentum levels fair	Happy with sprint event, feeling motivated and happy with sprint momentum

Table 10: Happiness Evaluation Metric

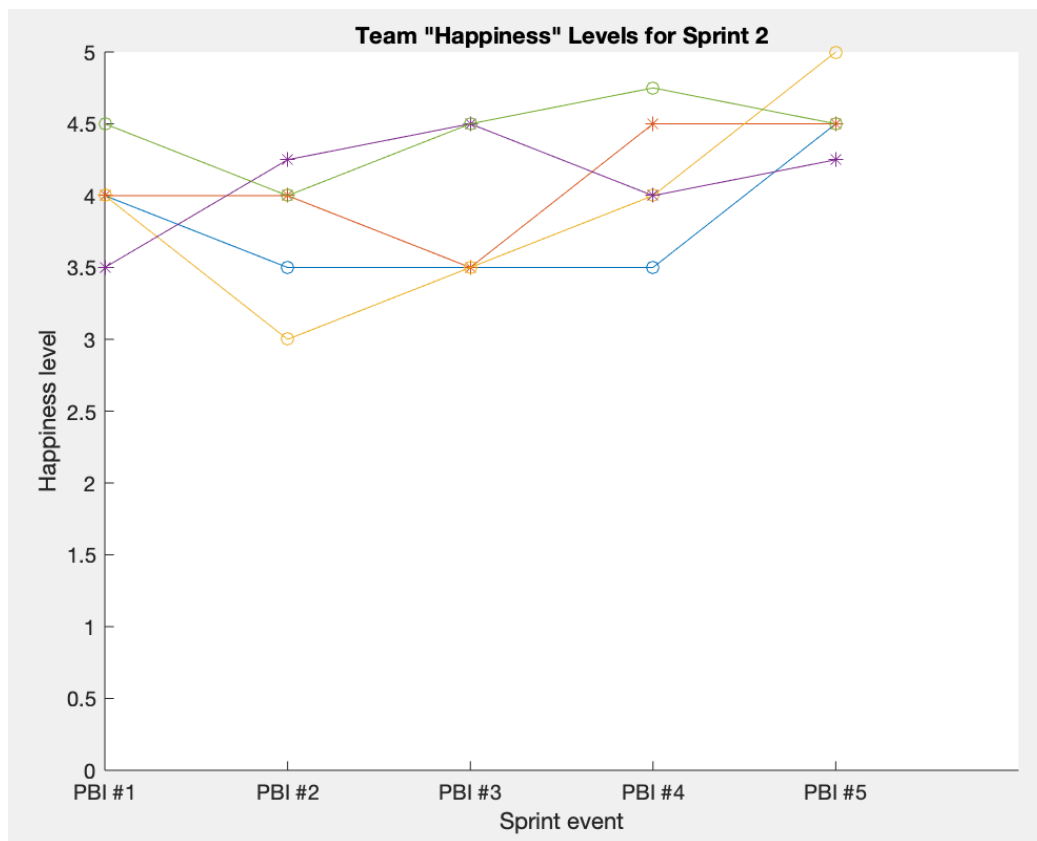


Figure 17: Sprint 2 Happiness Graph

During the Scrum Master-led Sprint Retrospective, the effectiveness of Sprint 2 was discussed with members of the team. Overall, the Sprint 2 Happiness Graph showed that the team was generally “happier” during the second sprint compared to Sprint 1. Discussions during the Sprint Retrospective meeting found that having completed two successful sprints together, the team felt more comfortable working together and had good team morale levels. Team members felt comfortable collaborating and discussing different opinions as a team.

Sprint activities that had generally lower happiness ratings were discussed to address what aspects of the event could have been improved. Similarly, generally positively rated events were discussed to identify why team members felt that that activity significantly aided the progression of the project. The findings of the retrospective discussions are detailed in the following sections.

11.2. Process Related Issues

A discussion of the collaboration process during Sprint resulted in the following actionable procedure changes such that following sprints can be streamlined and more effective.

Sprint Event	What Went Wrong	What Went Well	Actionable Procedure Objectives
PBI #1	-	Updating team members on progress and future steps led to a streamlined process, work done by one team member could be smoothly handed off to another team member to continue.	Continue updating the entire team on progress and next steps. Carefully organize code and comments such that work can easily be handed off with minimal time wasted to understand code.
PBI #6	Team members sometimes had different understandings of project objectives.	Discussion of different understandings of project objectives led to new perspectives on the project for the entire team.	Continue discussing project objectives and end goals with the entire team, leads to collaborative discussions and interesting perspectives for the project.

Table 11: Process Related Issues

12. Additional Project Management Content

12.1. Project Schedule



Figure 18: Now-Next-Later Framework

12.2. Project Risk Register

ID	Risk	Likelihood	Impact	Consequence	Mitigation
1	Inexperience with natural language processing	High	High	Difficulty in implementing and training the model	Allot time for learning natural language processing techniques
2	Inexperience with web development	Medium	Medium	Web page has many bugs	Allot time for learning web development
3	Failure to retrieve real time tweets	Medium	High	Unable to provide updates about real time disasters	Plan an alternative service for the prediction model
4	Failure to classify real and fake tweets accurately	Low	High	Prediction model is unreliable	Use wide variety of classifiers and preprocessing techniques
5	Failure to extract important information about classified real disasters	Low	High	Unable to provide the information needed by the users	Use different deep learning techniques and perform a lot of testing
6	Model does not improve after training	Medium	Medium	Model prediction accuracy does not satisfy the goal	Perform more preprocessing

7	Model does not work well with other datasets	Low	High	Provides inaccurate predictions	Ensure the model does not suffer from overfitting
8	UI displaying unintended information	Low	Medium	Information delivery system does not function as intended	Allot time for debugging
9	Delivering fake news to the users	Low	High	Unreliable service that takes away users' trust	Show prediction accuracy to the users and provide a disclaimer
10	Users does not trust our service	Medium	Medium	No one using our service and just wasting resources	Improve model's accuracy to gain trust from users
11	Run into bugs	High	Varies	Some functionalities might be restricted	Perform a lot of testing during implementation
12	Unsuccessful deployment to the web	Low	High	Service is not available for stakeholders	Plan an alternative platform for deployment
13	Unable to provide project's costs	Low	Low	Model may not be deployed as a web service	Carefully plan the budget for the project and request the university for more fund
14	Web server overload	Medium	Medium	Service becomes temporarily unavailable	Filter incoming traffic

15	Project does not go according to plan	Medium	High	Project may become unfeasible given the time restriction	Make contingency plans
16	Failure to meet high standards	Low	High	Unable to gain trust from the first responders	Keep training and building the model. Make a backup plan in case high standards become unfeasible.

Table 12: Project Risk Register

12.3. Project Management Tools

Microsoft Teams:

The team used MSTEams to work collaboratively on documenting the sprints, editing/working on documents/presentations simultaneously. Mainly used to store documents for the project.

Google Docs:

This was used for working on documents simultaneously.

Trello:

The team used Trello to manage tasks and keep track of product backlog, sprint backlog, tasks that are not started, in progress, and completed.

Jamboard:

For brainstorming ideas and during scrums for answering questions (“What did I do “yesterday” that helped meet the Sprint Goal?”, “What will I do “today” to help meet the Sprint Goal?”, and “Do I see any impediment that prevents me or the team from meeting the Sprint Goal?”)

13. Team Content

13.1. Individual statements on Contributions to Project

Binte Zehra

This team works amazing well with each other. Team members help each other overcome areas of weaknesses and identify room for improvements. All in all, I take initiative in every aspect of the project. I try to have a positive attitude towards my group and group members which encourages them to keep going. I motivate my team to work hard so that we can design a project that represents us. I tried my best to contribute towards the technical aspect, while focusing more on the documentation writing for the Gates (1, 2, and 3) and presentations for the peer reviews, done throughout this semester. Since we have to balance Capstone and other courses, time management was an evident issue. As a team member, I took the responsibility of keeping track of time so that everything can be managed smoothly, and we meet all the deadlines in a timely manner. Since, the focus for this term was more directed towards researching, I devoted my time learning about the Machine Learning tools and techniques, specifically the ones that can be applied on the metadata (i.e. looking at keywords, analyzing the text, etc.) and preprocessing of Tweets that are full of noise (i.e. spam, slang, ambiguous statements). As for the winter term, I am going to shift my focus more on the technical part (i.e. related to deployment) of the project. All in all, every team member has done their best and will continue to work hard in the future.

Neena Govindhan

The team as whole has been very good at staying on top of the tasks at hand and ensuring that the work is distributed to those who can contribute to the work more efficiently. As a team member I have had the opportunity to do various research, about ML, DL and NLP in general and about specific processes such as looking into specific classifiers and preprocessing techniques to help the team to construct the model. More of my efforts this semester went into preparing presentations and documentations for the gates, supervisor, and peer reviews, researching about the societal benefits and also assisting with the model. While in the group, I maintained a positive demeanor and assisted others when they needed help. For the next term, I would be working on more of the technical side as more effort will be needed as we will be deploying and improving the model.

Jessie Leung

Overall, I find that my team has been excellent in supporting the strengths and contributions of each team member in advancing the progress of the product. Each team member has contributed both to the technical implementation and report documentation, as well as in supporting the collaborative morale of the team. In regards to my personal contribution, beyond developing the deliverables alongside my team members, I believe that I encourage the progress of the project by encouraging the team to approach difficult tasks with a positive attitude.

Jonas Laya

Our team has been amazing getting the job done throughout the semester. Each team member contributed to all aspects of the project's development. In terms of individual contribution, I try to stay positive with my team members through tough times, taking away pressure and stress among the team members. Due to my lack of skill in documentation, I tend to contribute more on the technical aspects of the project, specifically in building and training the model.

Paul Sison

My trust and confidence in this group in completing tasks and meeting project objectives grow by the day, and that should speak volumes on how lucid this team performs. Each member of the team engages with the project professionally and produces work of good quality and ample standards. Suffice it to say, I look forward to the winter term and continue my exploration of Machine Learning with this group. As far as individual contributions go, a majority of my work went on the technical side of the project in researching and implementing the team's machine learning models using classical classification algorithms and some using neural networks. I also made minor contributions on documentation and in preparing presentation materials but a huge credit of those should go to my teammates. Moving forward, I plan to document everything I learn about our problem domain and present these findings to my teammates from time to time.

13.2. Individual statements on ITPMetrics survey (Peer feedback)

Binte Zehra

The ITP Metrics is a great tool to measure and reflect on my strengths and weaknesses as well as, how my team sees me versus how I judge myself. Specifically, the Peer Feedback helped me realize how my team appreciates my efforts and values my opinions. Their positive feedback regarding my performance, motivates me to work harder to achieve more and keep up the good work. The suggestions that my team made regarding how we could improve as a team and work together better to design a project that leaves an impact on the society. In the future, if my team has any suggestions on how I could improve myself, I will take their suggestions and work towards achieving that goal. I want to work collaboratively with my team as well as enjoy this whole experience and come up with a project that can help the society in one way or the other.

Neena Govindhan

Doing the ITP Metrics Peer Feedback not only gave my peers a chance to give me feedback, but also allowed me to reflect on my own performance. The feedback I received from my peers has helped me to improve on the different aspects of the teamwork competencies, some of which I had also realized I needed to improve on. The positive feedback I had received made me realize what I was able to achieve and needed to maintain in order to be a good team member. However, I do believe there are always improvements to be made, so I will strive to improve in all aspects and ask for my peers' feedback to keep improving.

Jessie Leung

The Peer Feedback given in the ITP Metric survey reinforced my sentiment that as a whole, my team was invested in working collaboratively to produce the best project that we could. My team constructively voiced both their appreciation and suggestions for how we could best work together. The feedback provided in the ITP Metric survey encouraged me to take note of my social and work responsibilities and how this might affect my contribution to the ENG 4000 capstone project. Resulting from this, I decided to make sure to set aside time to contribute for the project. My team members were also very supportive to this, and noted that the entire team would be accommodating to the varying schedules of each team member.

Jonas Laya

The ITP Metric survey played a key role to my success in working with our team project. The feedback I received from the survey highlighted my strengths and weaknesses which decided my role in the team. As for the peer feedback, I greatly appreciated getting honest comments and suggestions from my team. With their constructive feedback, I was able to adjust myself accordingly to develop myself better with my team members. I am now able to communicate more with the team which created a more positive environment for us. The ITP Metric survey also created a great opportunity for me to show my appreciation and to voice out my suggestions to my team members.

Paul Sison

Prior to completing ITP Metric surveys, my suspicion was that peer feedback would only reinforce what I already knew were my weaknesses. But reading my teammates comments and their limitless support motivated me to work in a manner that would always see great benefit to the team. It was also surprising to learn that what I see as weakness can be appreciated by others.

13.3. Team statement on ITPMetrics survey

As per our ITP Metrics, Team Dynamic report, for both Gate 1 and 3, as a whole we work well as a team. In terms of Communicate, compared to the Gate 1 report, now our Communicate score had increased. Because we are following the Agile project management approach we have been in more contact in terms of scrum meetings and we have more defined roles that are still flexible. We also have gotten to know one another much better, that we are more comfortable to be more open with one another.

In terms of Adapt, again our scores have increased in this gate. This is because prior to Gate 1, we were splitting up the tasks and putting it together and going over what needs to be changed at the end. But now there is more communication because of the scrum meeting and more in tune with what each other is doing. Still there are some days where time management is still an issue, because of other courses and personal lives, but with the planning we are more on track than before. Even if certain tasks don't go exactly as planned, we planned some buffer time so that there is time to catch up with the tasks.

In the Relate section, as a team we have been good at contributing to work equally and have had a positive environment where no conflicts have arisen as such. That hasn't changed much since. In terms of conflict, we hardly have major ones, we usually talk it out and get everyone's opinion and if someone is wrong, they would openly admit it. As a team we try to maintain a positive energy and environment for work. And we have built that trust in one another to do that work.

In the Educate section, since this is a new topic for all of us, this has been by far a very important part of the process. Since we all have different learning styles, we leveraged on team members that were able to grasp material much faster to get certain tasks done which we planned during the scrum meeting, but we all continue to learn and contribute to the progress of our project. Especially, when we started the sprint, we focused on learning specific topics such as learning the different classification models we want to use for the project. Since this was more specific this task was easier to accomplish, than before where we did more general research about ML and NLP. But we do understand that performing the general research has enabled us to understand the specific topics. We still continue to be open to learning and especially now will be focusing on the social aspect.

14. Self-Evaluation

Criterion	Self Evaluation Ranking	Justification
Contribute to a team in an appropriate and meaningful way	Exceeding	Team Content (Section: 13)
Apply an iterative process to refine or assign solutions for a given engineering design problem	Exceeding	Sprint 1 and Sprint 2 (Section: 6 and 9)
Achieve a system design breakdown for management and implementation	Exceeding	Product Backlog and Sprint 1&2 Backlogs (Section: 5 and 6.5&9.5)
Justify the strength and limitations of the solution and make recommendation for possible improvements	Exceeding	Sprint 1&2 Review (Section: 7&10)
Concise and coherent document that reflects critical analysis and synthesis	Exceeding	Team completed all the sections required for Gate 3. (Section: 1-13)
Design has been reviewed for sustainability impact and potential for negative unintended consequences	Exceeding	Social Impact and SDG Goals (Section: 4) Key Stakeholders (Section: 3)
Adjust project schedule based on project status	Exceeding	Project Schedule (Section: 12.1)
Monitor Risks during the lifecycle of the Project	Exceeding	Project Risk Register (Section: 12.2)
Applies all appropriate engineering concepts and fundamentals, theories, and practices to solve the engineering problems	Exceeding	Applied the Engineering Design Process: Requirements (Product Backlog, Section: 5) Design (Sprint 1&2 Planning, Section: 6.1&9.1) Implementation (Sprint 1&2 Tasks, Section: 6.6&9.6) Testing (Sprint 1&2 Tasks, Section: 6.6&9.6) Evolution (Sprint 2, Section: 9)

Table 13: Self-Evaluation

Appendix A: Research Papers

A Survey on Natural Language Processing for Fake News

By Ray Oshikawa, Jing Qian and William Yang Wang [i]

This research paper details the process of using Natural Language Processing techniques and machine learning classifiers for fake news detection. First, the paper provides an explanation of the task formulation, detailing the difference between a binary versus regression classification model. Next, the composition of commonly used social media post datasets (e.g. Liar, Fever, and Fakenewstweet) is explained. The paper then describes the methodology of building the prediction model by first preprocessing the dataset text using an NLP pipeline consisting of text tokenization, stemming, N-generalization or word weighting (TF-IDF), and other techniques. The preprocessed text is then used to train classifiers of two types: non-neural networks (e.g. SVM, Naive Bayes, Logistic Regression, Random Forest) and neural networks (e.g. Long Short Term Memory, Convolutional Neural Networks).

In the following section, the results of training and testing the previously mentioned datasets on selected classification models are presented using figures and discussion. The paper finds that certain models, for example LSTM, was able to achieve accuracy scores of greater than 90% on a binary classification of the test datasets. The paper then extends these findings by making recommendations for future models, indicating that more work can be done in non-binary classifications of truthfulness (e.g. A sophisticated measure of truthfulness being 3: Fully true, 2: Mostly true, 1: Mostly fake, 0: Fake). Furthermore, the temporal and content-transitional applicability of classification models can be explored by combining hand-crafted features with neural network models.

Twitter Under Crisis: Can we trust what we RT?

By Marcelo Mendoza, Barbara Poblete, Carlos Castillo [ii]

In this research paper they talked about how twitter users reacted in case of an emergency. To be particular, the 2010 chile earthquake. They analyzed Twitter and observed people's reaction within hours and days of the disaster.

Twitters reliability under extreme circumstances and its ability to distinguish between false rumors and confirmed news. Rumors are questioned more than the news by the Twitter community.

There were two types of studies done over the post-quake tweet data. First was characterizing the usage and social networks of the days immediately after the event. The goal of this task is to observe how rumors and news are propagated and the dynamics of the followers/followees relationship. As well as how most authoritative users influence topics discussed in the network. Second was investigating the ability of the social network to discriminate between false rumors and confirmed news. To do this they examined tweets

related to confirmed news and to rumors, classifying manually each tweet. The aim of this task was to measure if and how the network filters false information from accurate news.

One thing that gave us the idea that we also performed with our dataset, was to look for keywords. So, what was done in the article was, they focused on the community that surrounded the topic of the earthquake. To do this, they selected tweets which included a set of keywords which characterized this event. These keywords included hash-tags such as #terremotochile and the names of the affected geographic locations.

Fake News Detection on Social Media: A Data Mining Perspective ***By Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang and Huan Liu [iii]***

In this paper, they provided a detailed description on different techniques on detecting fake news based on the tweet's news contents and social contexts. They divided the fake news detection into phases, feature extraction and model construction. Feature extraction aims to represent news content and related auxiliary information into a formal mathematical representation. Model construction phase builds machine learning models to differentiate fake and real news based on the feature representation.

They talked about different approaches on extracting features from the news content and social context. On news content such as source, headline, body text and multimedia, they used linguistic based approach and visual based approach. Linguistic based approach uses lexical features including character-level and word-level features such as total words, characters per word, frequency of large words and unique words; and syntactic features including sentence-level features such as frequency of function words and phrases of punctuation and parts of speech tagging for natural language processing. Visual based approach uses visual features such as clarity score, coherence score, clustering score etc., and statistical features such as count, image ratio, hot image ratio etc. For social context features, they used 3 different approaches. User based approach which captures user's profiles and characteristics to identify malicious users. Post based approach finds potential fake news from different aspects of the social media post like the reaction of the public. Lastly, network based approach which extracts network based features from users to represent network patterns for fake news detection.

Similarly, the model construction phase has two different models, news content model and social context model. News content models rely on news content features and existing factual sources to classify fake news. They categorized existing approaches as knowledge based and style based. Knowledge based approach uses external sources to fact check proposed claims to assign its truth value. Style based approach captures the manipulators in the writing style of the news content. Social context model uses a stance based and propagation based approach. Stance based approach utilizes users' view-points from relevant post contents to infer the veracity of original news articles and propagation based approach reasons about the interrelations of relevant social media posts to predict news credibility.

The spread of true and false news online

By Soroush Vosoughi, Deb Roy, and Sinan Aral [iv]

This research paper looked at the diffusion of true, false and mixed new stories using a comprehensive data set of all the fact-checked rumor cascades that spread on Twitter from 2006 to 2017. A rumor cascade is defined as the original tweet plus the retweets. Cascade depth would be defined as the number of times a tweet is retweeted. For example, if a tweet was retweeted by User 1 and User 2 then the cascade depth is 1, but if the tweet was retweeted by User 1 and then User 2 retweets User 1's retweet, then it is a cascade depth of 2. The sampled rumor cascades were fact-checked by 6 different organizations and were also manually verified by undergraduate students. In the study, they had found that the largest rumor category was politics, which meant that it had the largest number of rumor cascades, and natural disasters had the smallest number of rumor cascades. They had also found that falsehood diffused farther, faster, deeper, and more broadly than the truth in all categories.

Some key points they found for the paper were that more people retweeted the falsehoods than they did the truth and the truth took about six times as long as falsehood to reach 1500 people. They realized that more work is needed to find the reason as to why false news spreads faster and reaches more people as they had found that the number of followers, being verified, tweeting more, had very little impact. One possible explanation could be that novel information is valuable in that it aids decision-making and in a social perspective. They used a LDA topic modelling and tweet semantic tool called Tweet2Vec and found that false rumors were significantly more novel than the truth. However further research is needed to understand the behavioral explanations of the results.

Microblogging during two natural hazard events: What Twitter may contribute to Situational Awareness

By Sarah Vieweg, Amanda Hughes, Kate Starbird, Leysia Palen [v]

This paper is an analysis of the utility of microblogged information via Twitter in improving situational awareness. Per Sarter and Woods, Situational Awareness (SA) is "all knowledge that is accessible and can be integrated into a coherent picture, when required, to assess and cope with a situation." SA is applicable in many domains, including weather and emergency response. Having more (and credible) information about people affected by a crisis and the condition of their environment can lead to a more optimal allocation of resources and aid coming from government agencies and emergency response units.

In this study, the authors analyzed twitter activity during two disaster events, which both occurred in the Spring of 2009 in the US. These are the North Dakota Red River floods (RR) and the Oklahoma grassfires (OK). The authors collected 13,153 tweets from 4,983 unique users in a span of 51 days for the RR dataset. On the other hand, 6,674 tweets from 3,852 unique users were collected in a span of only 5 days for the OK dataset. To ensure quality, the datasets were reduced to "on-topic" tweets that have content relating to the said disaster events, and to users that are local to the affected areas. Every tweet was then coded into

one or more categories, which the authors identified from emergent themes in both datasets. The authors also accounted for the time-sensitivity of the tweets and identified three phases of an emergency—warning, impact, and recovery.

The author's findings are as follows:

- Geo-location and location-reference information were more prevalent in the OK dataset making up 40% of OK tweets versus the 18% found in RR tweets. It is suggested that the nature of the disaster contributes to this result, i.e. wildfires are unpredictable and where it spreads has to be noted while in a predicted flooding of a river, affected areas are implicitly conveyed from flood level information.
- There are notable differences in the amount of geo-location information included in each tweet for either emergency event. The OK dataset contains an average of 1.5 different geo-location features per tweet, while it's 1.35 features per tweet in the RR dataset.
- The nature of the disaster events also contributes to the differences in the type of information included in each *Situational Update* tweet. To explain the differences, for example, there are higher instances of OK tweets containing *Wind* information because its direction and speed are potential indications of the grassfire's path. *Preparatory activity* and *Volunteer information*, on the other hand, are higher in the RR dataset considering that it's an anticipated event.

The immediate application of this paper's findings cannot be fully realized at the moment as the current project objective does not require classifications between identified disaster events. However, the coding scheme developed by Kendra and Wachtendorf can be useful in defining keywords when mining for actionable information from disaster-related tweets.

Appendix B: Glossary

- BoW: Bag of Words (feature extraction technique)
- CNN: Convolutional Neural Network
- DL: Deep Learning
- LDA: Latent Dirichlet Allocation (topic modelling)
- LSTM: Long Short Term Memory (Neural Network Classification Model)
- ML: Machine Learning
- MVP: Minimum Viable Product
- NLP: Natural Language Processing
- NLTK: Natural Language Toolkit
- PBI: Product Backlog Item
- RNN: Recurrent Neural Network
- SDG: Sustainable Development Goals
- SVM: Support Vector Machine (Non-Neural Classification Model)
- TF-IDF: Term Frequency-Inverse Document Frequency (feature extraction technique)

References

- [i] Oshikawa, R., J. Qian, and W. Y. Wang. "A Survey on Natural Language Processing for Fake News Detection." Cornell University. <https://arxiv.org/abs/1811.00770>.
- [ii] M. Mendoza, B. Poblete, and C. Castillo, "Twitter under crisis: can we trust what we RT?," *Twitter under crisis | Proceedings of the First Workshop on Social Media Analytics*, 01-Jul-2010. [Online]. Available: <https://dl.acm.org/doi/abs/10.1145/1964858.1964869>.
- [iii] Shu, K., A. Sliva, S. Wang, J. Tang and H. Liu. "Fake News Detection on Social Media: A Data Mining Perspective." Cornell University. <https://arxiv.org/pdf/1708.01967.pdf>.
- [iv] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
- [v] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. "Microblogging during two natural hazards events: What twitter may contribute to situational awareness." *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1079–1088. 2010. <https://doi.org/10.1145/1753326.1753486>