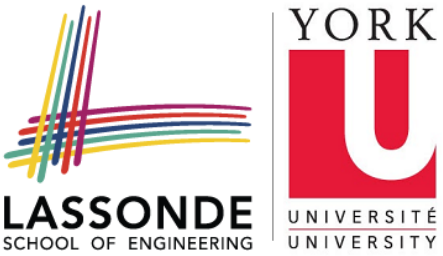

ENG 4000: Agile Roadmap

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

Date: 30/10/2020

Team P

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

Table of Contents

List of Figures

1. Project Overview

1.1. Product Vision

Purpose

As time is evolving people are relying more on social media. More specifically, to get updated on what is going on around the world. In times of emergency, people are turning to Twitter to provide them with reliable information. The purpose of this project is to develop an efficient Machine Learning model using Natural Language Processing (NLP) to classify a tweet as real or fake. Our team chose this project because it encompassed machine learning, which was an interesting topic for all of us. We all have the desire to learn machine learning. The original Kaggle project is a beginner level project, so we should be capable of learning how to do the Kaggle project and expanding from there.

Target Group

The product is targeting general users, news agencies, and first responders. For more details, refer to section 1.2 Stakeholders.

Problem Solved

As sources of news information become more plentiful on the Internet, differentiating between fake and real news is becoming increasingly difficult to the everyday user. The complexity of fake news sources has dramatically risen as creating authentic-appearing sites and social media accounts has become more accessible. Moreover, fake news is written stylistically to purposefully deceive Internet users, and contains fake content that imitates the presentation of real news. Vosoughi et al. found that in a number of categories (including but not limited to politics, business, terrorism and war, and natural disasters), falsehood diffused farther, faster, deeper and more broadly when spread through online mediums[1].

To address the problem of the spread of fake news on the Internet, this project roadmap seeks to outline a product that has the ability to gather mass amounts of readily available, user generated data (via Tweets), extract the truthful information and streamline it such that useful and reliable news information is provided. The scope of the project is limited to natural disaster news information due to the time constraints of the project, but beyond the scope of the ENG4000 course, it can be extended into multifaceted news dimensions.

Product Description, Uniqueness, Feasibility

Natural Language Processing classification models have already been significantly explored, producing models that generate more than 90% accuracy on a binary true or false classification on multiple datasets (e.g. FakeNewsTweet, Liar, Fever)[2]. However, the data produced by these models are highly technical and immensely large in size, and a more widely usable product based off of these machine learning models requires further development. This product aims to compile the binary true or false classification of a large dataset of Tweets, and generalize the truthful information into a format that can be accessible and readable by the everyday, casual Internet user. As such, the proposed product strives to reduce the technical sophistication of machine learning model results by categorizing the classified datasets and presenting them to users in a format that is useful to them.

The feasibility of the project is supported by the existing work on natural disaster classification models. Improvements to an initial classification model as expected to be produced in the first Sprint of the project roadmap can be continuously developed by the testing of various classification models (for example Naive Bayes, Random Forest, Logistic Regression), and the addition of handcrafted features to the model (such as a Bag of Words feature extraction).

---

[1] *Vosoughi, S., D. Roy, and S. Aral. "The Spread of True and False News Online." Science, vol. 359, no. 6380, 2018, pp. 1146-1151. SCOPUS, www.scopus.com, doi:10.1126/science.aap9559.*

[2] *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.*

Furthermore, the major challenge of the project will be to categorize the results of the classification data and present relevant and reliable information to the users. To ensure the feasibility of the project within the ENG 4000 timeline, the scope has been limited to natural disaster focused Tweets. Further refinement on which aspects of natural disasters (e.g. geographic, temporal, disaster event type based) may be defined as the project progresses and the accuracy and usefulness of the classification model to be developed can be analyzed.

Product Societal Benefits

In times of emergency, people rely on Twitter to provide accurate news. False news generated by an individual is misleading a large group of people trusting these tweets. The impact of fake tweets is not only affecting Twitter users but also people in need affected by a disaster. Thus, developing a successful project, using Machine Learning frameworks and models, which can detect whether or not a tweet is real or not, will provide accurate information to first aid responders and news agencies.

1.2. Stakeholders

General Public/Internet Users

The primary object of the proposed project is to produce a viable product in the form of a web application or website that produces reliable information on natural disasters. The product is planned to do so by using a classification model to process the mass amounts of user-provided data in Tweets. The portion of the general public that uses the Internet as a source of information regarding natural disasters is therefore a direct stakeholder in the project, as there currently exists a hugely problematic existence of fake news on the Internet that misleads the public in a far-reaching and lasting manner. This phenomenon can be greatly reduced by the planned product, and is important in maintaining the truthfulness of news content that is available to and consumed by Internet users.

News Sources

Current sources of news information are a direct stakeholder in the proposed product due to their content similarities. Traditional news (newspapers, television), as well as Internet news sources (social media platforms such as Twitter, Facebook, BuzzFeed News) share a role with the proposed product as their objective is to provide the public with news information. Fake news propagates more prolifically on Internet mediums, due to its inherent accessibility and low cost. Everyday users are capable of posting “news information” on social media platforms without the requirement of fact checking or any expertise. In fact, in the event of false news spreading, attributes of the source such as number of followers, or likes/retweets

on a Tweet was found to have little effect[3]. As such, the impact of the product may have more of an effect on Internet news sources. The issue of fake news dissemination is already widely discussed, therefore a framework to autonomously or semi-autonomously detect truthfulness and generate news content with the assurance of truthfulness will have an impact on current news disseminators and their services.

First Responders

First responders are a stakeholder that are less directly affected by the initial conception of the project, but could be a primary user of an expansion of the product. First responders enter disaster situations with little to no situational information. If equipped with more context and awareness of the disaster event, first responders could more effectively react to the event. In disaster situations, even a marginal advantage could result in one or multiple lives being saved, or the prevention of catastrophic damages.

The multitude of freely available data on Internet platforms provides a unique opportunity for data extraction that could produce such an advantage for first responders. The difficulty lies in the fact that first responders cannot manually filter through the vast amount of information available and accurately extract just the truthful information. Furthermore, there is the question of whether people trust information retrieved by classification models enough to deploy crisis response actions based on it.

2. Project Timeline and Management

2.1. Feature Roadmap

The team took the Now - Next - Later approach to break down the project's feature roadmap. The roadmap was divided into three sections to show which features we are currently working on, what we have planned for the next few Sprints, and what we wish to achieve much later in the project. The *Now* stage includes features that the team is currently working on and features that are expected for the upcoming product release. The *Next* stage includes features planned after the release of the product. Plans on the next stage are flexible since changes can be made after we get feedback from the release. Even more flexible is the *Later* stage depending on how sprint reviews go. This stage includes high level features that we plan for the project. The diagram below shows how the team planned the features to be added in the product:

[3] *Vosoughi, S., D. Roy, and S. Aral.*

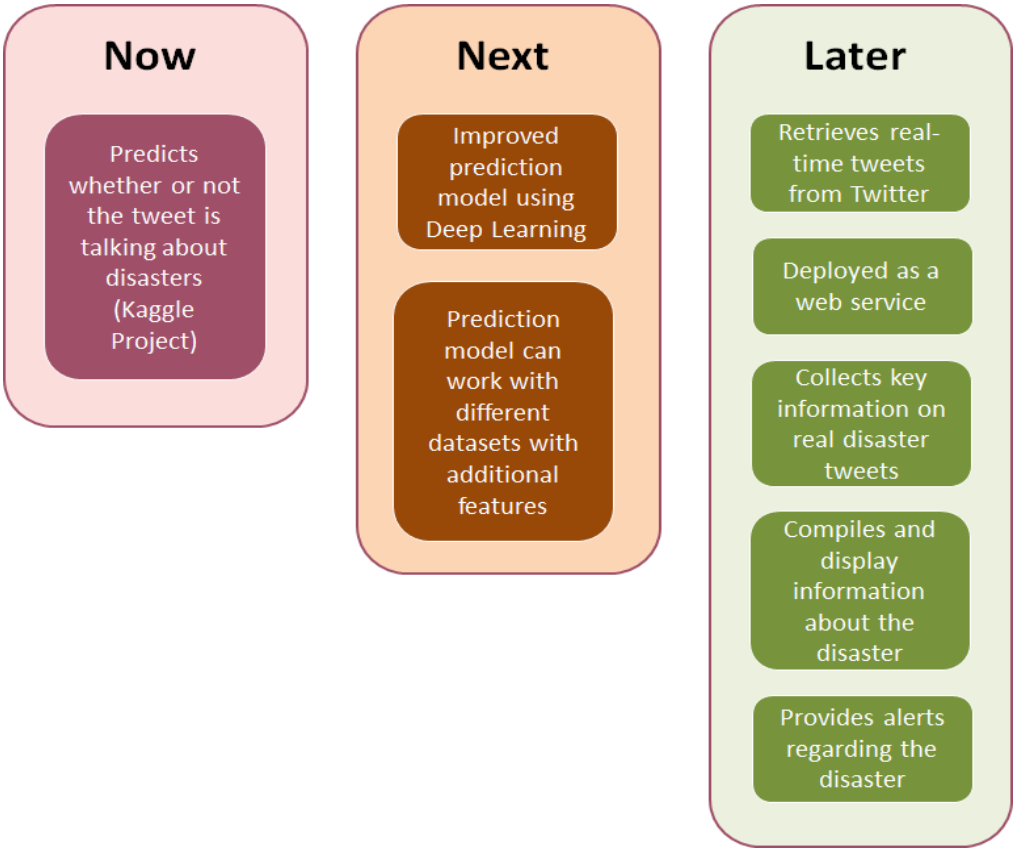

Figure 1: *Now-Next-Later Framework*

2.2. Release Planning

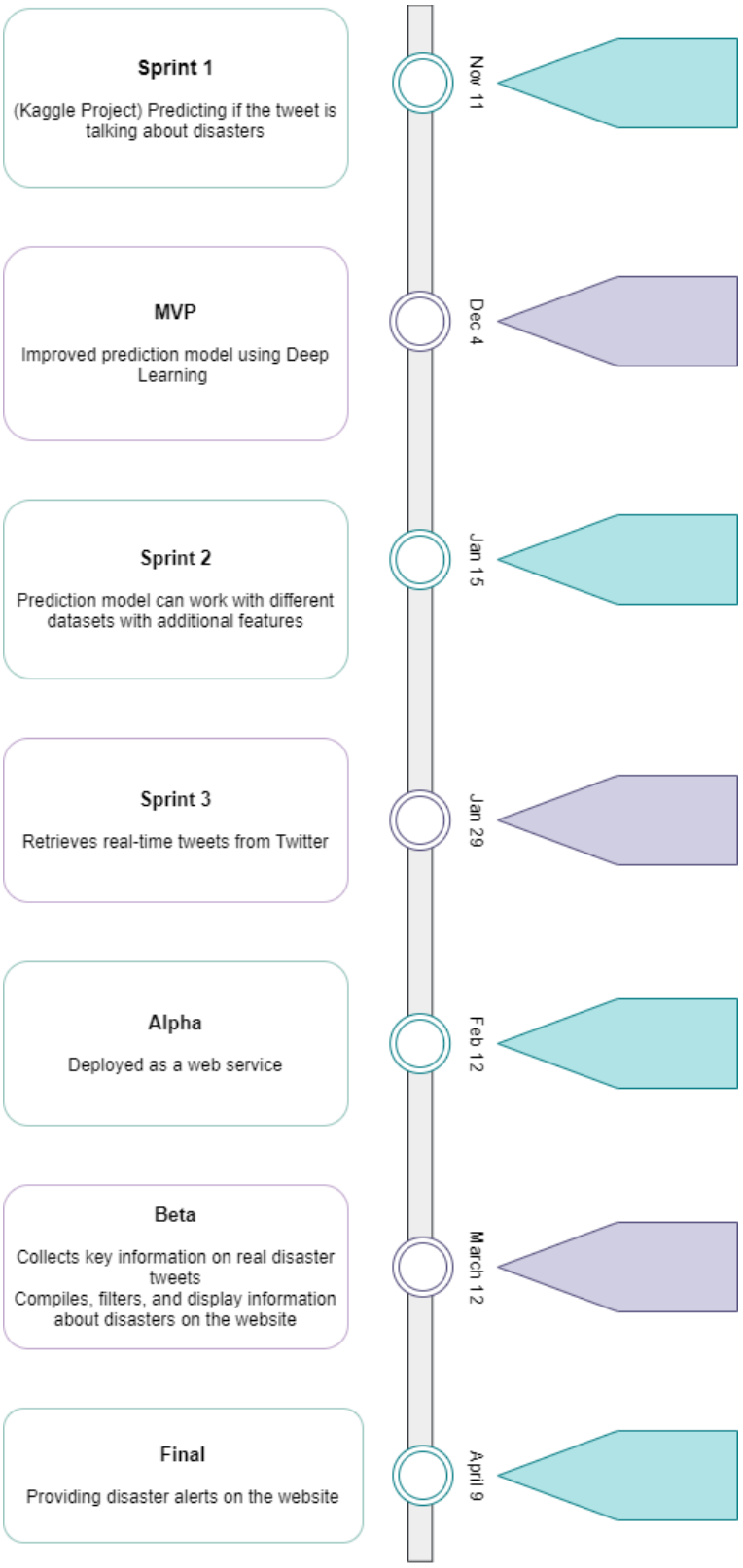

Figure 2: *Project Timeline*

2.3. Sprint Planning

Sprint Planning is the first major step a team takes to start a single Sprint in an Agile workflow. In a Sprint Planning session, the team reflects on the Product Backlog, making necessary adjustments to it, and commits Product Backlog Items (PBIs) to the Sprint backlog. PBIs are any work that provides business value to the project and consumes time from the team. Committed PBIs, which ideally should meet the current Sprint goals, are implemented, tested, and deployed during

Sprint execution.

Scheduling

The team will hold Sprint Planning sessions on a Friday, following a Sprint Retrospective session if the team happens to hold one on the same day—in other times, the Sprint Planning happens two days after a Sprint Retrospective. The Scrum Master, which will be a rotating, predetermined role, will lead and mediate the Sprint Planning sessions via Zoom. An estimated two to four-hour timebox will be allotted for this meeting.

Backlog Refinement

The Product Owner (PO), which is a role assigned to one team member throughout the project, confirms the order of priority of the PBIs listed on the Product Backlog. PBIs that add the most product value are often given higher priority. The PO also clarifies any vague PBIs to the team and decomposes them into smaller, more manageable PBIs if needed. The team then deliberates on the amount of work required in each of the prioritized PBIs. A simple Small-Medium-Large metric or a Story Point system will be used to describe these amounts. As a guideline, PBIs should have a Who, What, and Why in describing its user stories. Doing so serves as further validation of a PBI's value and atomicity.

Commiting to the Sprint Backlog

Before committing any PBIs to the Sprint Backlog, the PO and the rest of the team first agrees on a Sprint goal. A Sprint goal is often a high-level goal that is descriptive of a set of PBIs that contributes to one major aspect or functionality of the product. The process of committing will be a team effort to ensure that everyone is comfortable with the amount of work being committed given the Sprint timeline. Once a PBI is in the Sprint Backlog, the team agrees upon a set of acceptance criteria that the PBI has to meet to be declared as "done" during Sprint Review. If the PBI happens to be a software feature, a sample acceptance criterion can either be one or all of the following: properly tested, runs in X platform, compatible with Y browser, extensively documented, etc.

Listing Tasks

With a filled Sprint Backlog, the team will then create a comprehensive list of tasks involved in completing each of the committed PBI. Potential tasks for PBIs in this project can range from, but not limited to, data cleaning, feature extraction, web page design, coding, testing, refactoring, integration, and deployment. Upon the group's preference, no tasks will be specifically assigned to members. Any member can volunteer for any task as they see fit.

SCRUM Board

The SCRUM board is where the team can track the status and progress of each committed PBI in the Sprint Backlog. During Sprint execution, members can volunteer to work on any task in the "Not started" column. Members then move their chosen task to the "In progress" column" as they work on it, and then to the "Done" column when they have completed the task. Ideally, the number of tasks in the "In progress" column at any given time does not exceed the number of members in the group. The team aims to work diligently on a single task at a time to maximize focus and efficiency.

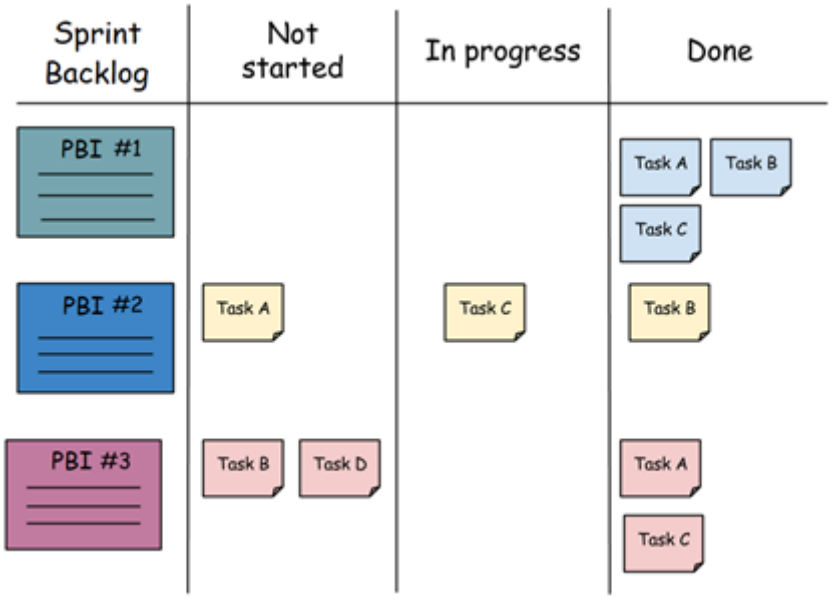

Figure 3: *SCRUM board*

2.4. Daily Scrum Plan

Our team will perform Scrum standups twice a week: Monday and Wednesday. The duration of the standups will be no longer than 15-20min.

The platform that will be used for the scrum meetings is Zoom. During the scrum meeting, each team member will talk about their progress, what they plan to do next and any obstacles they are facing. The Scrum Master will be responsible for overseeing that the

tasks done and the tasks to be done will keep the team on track to meet the sprint goal. The ScrumMaster will also be responsible for coming up with a solution or finding resources for solving any problems that occur. The team will answer these questions individually during every scrum meeting on Stormboard to reflect what they have done so far, what needs to be worked on next. The ScrumMaster will make updates to the Scrum Board on Trello after the meeting. This will give the team a chance to see their progress for the upcoming sprint.

The scrum planning will be organized by the ScrumMaster on Sunday and Tuesday, the days before the actual scrum. The role of a ScrumMaster will rotate through the group members, the order will be predetermined by the scrum master. Every group member will get to be the ScrumMaster twice a week(each sprint).

2.5. Sprint Review Plan

In the Sprint review, the team presents and demonstrates completed Product Backlog items (PBIs) to prospective users, stakeholders, and other interested parties. This is to solicit constructive feedback of the completed work so the team, especially the Product Owner, can assess customer satisfaction, maximize the product value, and find opportunities for further improvement on completed features.

Scheduling

The team plans to hold Sprint reviews on Wednesdays or Fridays, depending on what day the last Sprint finished. For example, if an ongoing Sprint has a finish date set on a Thursday, a Friday, or the weekends, then the Sprint review for the said Sprint will be on the Wednesday of the following week. If a Sprint is set to finish earlier in the week, *i.e.* from Monday to Wednesday, then the Sprint review will be held on the Friday of the same week. Wednesdays and Fridays are the most flexible days for the team considering each of the members' personal schedules, and are therefore, the more ideal days to hold meetings that usually last more than an hour. The few days in between the Sprint finish date and the Sprint review date is to allow for peers and stakeholders to respond to their invitation to join the review. These emails are sent out by the Scrum Master on the same day as the Sprint finish date.

Roles & Responsibilities

There are three roles that team members will assume in a Sprint Review. These are the Product Owner, the Scrum Master, and the Developers.

The Product Owner (PO) is a role that will be assumed by the same team member in all of the Sprints. It will be the PO's responsibility to maximize the product's value, or, at the very least, ensure that the product's value does not decline throughout the development process. In the Sprint Review, it is the PO's responsibility to make decisions on the status of each

committed Sprint Backlog item. The PO will also note and take into consideration all stakeholder feedback given during the session for the next Sprint.

As mentioned, the Scrum Master (SM) role is a rotating role and will be assumed by each team member not elected as PO. This role is assigned to a team member during each Sprint Planning session. Aside from the SM's duties of leading daily Scrum standups, the SM will also host and run the Sprint Review sessions via Zoom. The SM ensures that all parties of interest are invited to the session as well.

The other three members who are not in the PO or SM role will act as Developers in the Sprint Review. They will be leading the demonstrations of completed PBIs to the stakeholders. The Developers must also be prepared to answer stakeholder questions.

Procedure

The team will adhere to three simple steps to execute the Sprint review:

1. Demonstration
2. Evaluation
3. Adaptation

In the demonstration step, the developers will demo all the completed PBIs to the stakeholders. If the PBI is something tangible such as a new product feature or an improved functionality, then the developers will demonstrate how to access and use the said feature. If the PBI is a research task, then the developers will present the results of the study and explain to the stakeholders how the said study is of primary interest or how it adds value to the product. Any incomplete PBIs on the current Sprint Backlog will also be discussed. The developers will state the challenges the team faced and the reasons why some PBIs remain uncompleted. The developers can also raise feasibility concerns for the incomplete PBIs should the need arise.

In the evaluation step, the stakeholders get the chance to play with the product and try out the new functionality or features themselves. Having a personal experience with the product allows them to offer a more informed assessment of its performance. The stakeholders can direct questions to the developers regarding the features should they have any. Although not ideal, the stakeholders can also make requests for new features during this step to the PO. Any other feedback given by the stakeholders should be noted by the PO at this time.

Lastly in the adaptation step, the PO declares which PBIs are "done" and moves them to the completed column of the Scrum Board. A PBI is considered "done" if it meets all its acceptance criteria that was agreed upon during the Sprint Planning session. For uncompleted PBIs, the PO moves them back to the Product Backlog and evaluates each one whether it is a priority item for the next Sprint, taking into consideration any work that has already been made. PBIs can also be dropped at this point should feasibility become an issue. If any new requests have been made by the stakeholders, it will be upon the PO's discretion whether or not to entertain them. If the PO heeds the request, a new PBI will be

created and added to the Product Backlog. The PO then consolidates all of the solicited feedback from stakeholders and team members alike and reflects on the new Product Backlog with the team.

Metrics to measure Sprint Completeness and Efficiency

The team will track Sprint completion using a Burndown chart. In this chart, we are plotting the amount of work left—either by the number of tasks or by a story point system—against the number of days that have elapsed in the current Sprint. The chart paints a clear picture of the team's ability to do a consistent amount of work over time. In the example figure below, if the team does not wander too far away from the blue line then the team should be in good shape as the Sprint progresses.

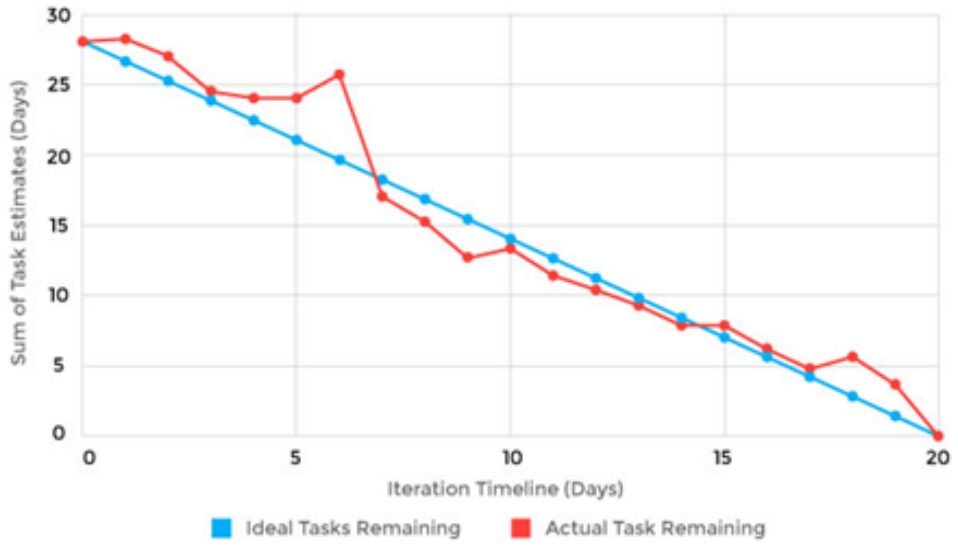

Figure 4: *Burndown Chart*

A Velocity graph can also be used to outline the team's work output over the number of iterations or Sprints that have passed. Assuming Sprints are of the same length, the team's velocity for a given Sprint is measured by the amount of Story points the team has completed during the Sprint. An increasing velocity over several Sprints conveys that the team is becoming more efficient in completing tasks.

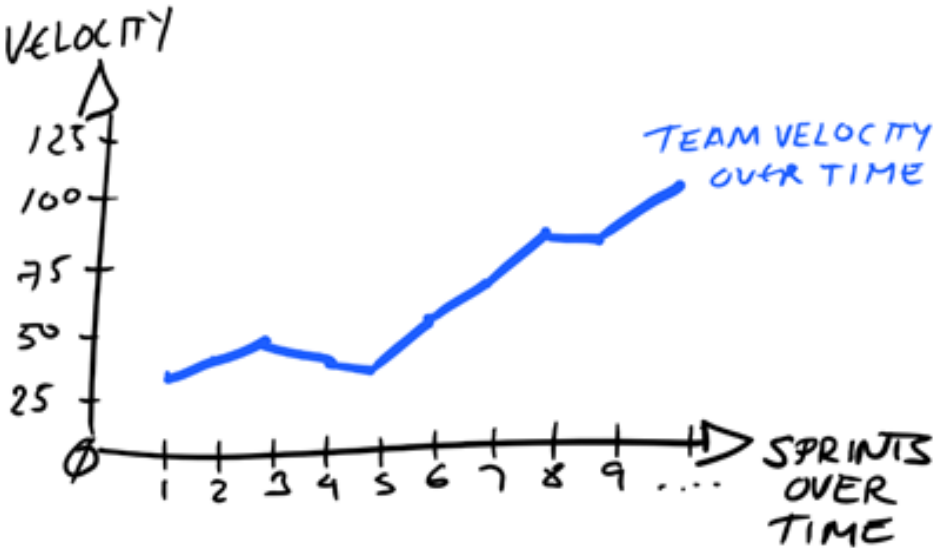

Figure 5: *Velocity Graph*

2.6. Sprint Retrospective Plan

The purpose of the Sprint Retrospective is for the team to self-inspect its performance during the Sprint, identify what worked well in the current Sprint and what can be improved for future Sprints. The Scrum Retrospective will be led by the Scrum Master, which is a role that will be rotated through each team member, such that the individual skills and strengths of each team member can be made use of and bring different perspectives to each Sprint.

Sprint Retrospective Discussion Topics

- What went well in the Sprint and why?
- What can be improved for the following Sprints?
- What are actionable steps we can commit to improve for the next Sprint?

Metrics to measure Sprint Effectiveness

	Team Cohesion	Use of Time	Ability to meet Sprint deadlines/goals	Team contribution	Value of Sprint deliverable	Confidence
Team Member #1						
Team Member #2						
Team Member #3						
Team Member #4						
Team Member #5						

Figure 6: *Pre-Sprint Progression Table*

Before the Sprint Retrospective, team members can fill out a table on how they feel about the progression of the Sprint that can be discussed during the Retrospective. Criteria that can be taken into consideration include:

- Team cohesion
- Efficient use of time
- Ability to meet Sprint deadlines/goals
- Team contribution
- Value of Sprint deliverable
- Confidence

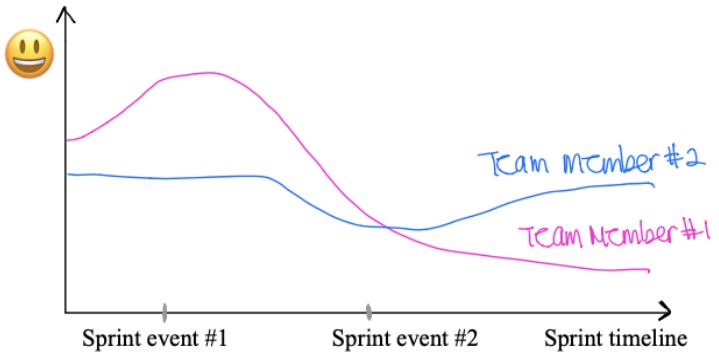

Figure 7: *"Happiness" Graph*

A "Happiness" graph can be used by the team to visually identify Sprint events that team members thought worked effectively towards the product goal. The Scrum Master can lead a conversation on why certain events have a positive or negative impact on the progression of the project.

Implementing Process Changes

The Retrospective review should result in actionable process improvements that can be implemented in future Sprints. Process changes should be measurable and actionable (e.g. "Team checks in with each other daily about what was accomplished during that daily Scrum" vs "Team should communicate more"). Improvements should be kept track of in a checklist, and be actively part of adjusting the next immediate Sprint. The checklist should be composed of actions to implement and incorporate process improvements, and the checklist should be continuously reviewed by the team to ensure they are being followed.

2.7. Project Management Tools

Microsoft Teams: Provides a space for team to work collaboratively on project deliverables, editing/working on a document simultaneously

Trello: To keep track of daily progress in the project; SCRUM boards, due dates for project deliverables

Stormboard: 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?”)

3. Project Team

3.1. Roles and Responsibilities

Roles

For each sprint, we plan to rotate (ScrumMaster) through the different Agile scrum roles (down the list).

The first sprint will be as follows:

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

Responsibilities for Fall Term

Our team decided to divide the responsibilities by various Machine Learning aspects:

Jessie is responsible for preprocessing work.

- Cleaning the data
- Removing unwanted tokens
- Filtering data

Binte and Neena are responsible for sentiment analysis.

- Data analysis
- Looking for specific keywords
- Interpreting data in regards to categorizing fake/real tweets

Paul and Jonas are responsible for other Machine Learning tasks.

- Feature extraction
- Model evaluation

Responsibilities for Winter Term - TBD

3.2. Motivation

Binte Zehra

The reason I picked this project is because it involves an aspect related to Machine Learning. The recent growth in the industry in regards to Machine Learning, made me interested in learning more about the topic. The project itself can be taken in many ways, societal aspects when it comes to designing an efficient system to detect fake/real tweets, which can help save people's lives affected by these disasters. Developing a system which can help people in need and not wasting resources being invested, to provide reliefs for the survivors. Personal aspect for individual growth, by learning about a tool which will be beneficial in the future. Hence, designing a successful project will not only prove to be a personal achievement but also serve the community.

Neena Govindhan

Learning Machine Learning (ML) was my main motivation to select this project. ML was first introduced to me in my previous degree in Genetics, where I learned that ML was being used in genomic sequencing prediction and found it a fascinating topic. And throughout the years I have been Engineering, I have been hearing more and more about it. I also had planned to take an Introduction to Machine Learning course in the winter term. Designing a model that would predict whether or not disaster tweets were real or not, seemed like a good starting point and motivation to learn ML. With expansion to deciphering whether or not a disaster tweet is true or not, would be most beneficial to society as fake news is wide and rampant these days.

Jessie Leung

As with most of my teammates, I am motivated to learn about how machine learning algorithms work and how to implement them. Also, as more and more information is generated via the Internet leading to the rise of fake news dissemination as a major public concern, I believe that it is important for countermeasures to be put in place to contain the spread of false information. Beyond that, I think that classification models will be a multilaterally useful tool in categorizing the immense volume of data available on the Internet in a variety of societal dimensions.

Jonas Laya

My motivation for this project is my personal interest in machine learning. My personal hobby is playing games. I love playing tactical games such as chess, trading card games, and video games such as turn based RPGs. I became interested in machine learning when I saw AIs being used in such games. I was amazed at how well they perform and it made me wonder how it is possible that machines can be so intelligent that they can beat professional players' minds. Since then, machine learning quickly became one of my personal research interests and this project is the first opportunity for me to learn and dig deep into machine

learning. As for me who grew up in a place where first responders are very slow and almost unreliable, I greatly appreciate first responders in Canada. Developing an AI system where we can help them further to react more quickly with more accurate information is something I can offer them as gratitude.

Paul Sison

Personal growth and the irresistible allure of cutting-edge technology are my main motivators for picking this project. Having been taught all about traditional programming in my earlier years at University, Machine Learning approaches to problems appear nothing short of refreshing. Being able to have a chance to teach myself Machine Learning as well as its applications to Natural Language Processing while completing my Engineering Capstone course then becomes a no-brainer. With the project's aim of leveraging Social Media data to help in disseminating emergency information and to aid in disaster response also comes as a huge bonus. Having come from a typhoon-ridden country in Southeast Asia, I find unlimited appreciation for the kind of services that can arise from the successful completion of this project.

3.3. Team Strengths and Weaknesses

As a team at first we did not know each other very well, but we were able to quickly come together as a group. We are all fairly new to machine learning as a topic and that is our motivation to work on the project. This did initially pose a problem as to where to start since we all were fairly new to the topic. Under the guidance of our supervisor, feedback from peers, our own research and helping each other, we had a better understanding of the problem at hand and have figured out a direction we want to go with the project.

As per our ITP Metrics, Team Dynamic report, as a whole we work well as a team. In terms of Communicate, we have been messaging each other on Whatsapp and having at least one Zoom meeting a week to go over what needs to be done for that week. We keep all of the different files on Microsoft Teams, so that everyone has access to it. To improve our communication we will be using Trello boards more to keep track of the tasks at hand and who will be responsible for it in our scrums.

In terms of Adapt, we are doing well to coordinate between each other to figure out what to do and monitoring what has been done. The thing we need to improve on as a team is to work on our time management. As discussed earlier tools such as Trello, Stormboard, etc. and starting the "daily" scrum meetings will help to effectively plan out our time and pace our goal progression more efficiently.

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. To improve on Relate, we should be more prepared in our meeting to be able to have more healthy, fact-driven conflicts, or bring about different perspectives. The objectives of our meeting should be

stated ahead of time, so discussion will also be efficiently done. This will be the ScrumMasters role to make sure that this is stated before the scrum meetings.

In Educate, since this is a new topic for all of us, this has been a very important part process so far. We have each done some literature research on the problem to understand how we can expand on the Kaggle project given, and have come together to teach each other the important aspects of the literature research done. We have also done some online tutorials to understand machine learning and natural language processing. We also feel comfortable sharing that information with each other.

3.4. Communication Channels

The team's main communication methods are as follows:

Zoom: Video conferencing platform to hold our scrums and meetings with the supervisor, team meetings will be held on Zoom to have discussions about the project

Whatsapp: Group chat to provide weekly updates and discuss any minor/major issues faced by team members, as well as, communicating daily to bond better and developing a good team understanding

Google Docs/Slides and Microsoft Teams files: To share useful links and information, simultaneously work on shared documents, have targeted discussions as documentation for deliverables are produced by using comment/resolved threads, creating presentations for supervisor and peer reviews

Email: Professional platform to communicate with course directors and supervisors, asking questions, clarifying ideas, or getting help