

**TEXT CLASSIFICATION OF CLIMATE CHANGE TWEETS USING ARTIFICIAL  
NEURAL NETWORKS, FASTTEXT WORD EMBEDDINGS, AND LATENT  
DIRICHLET ALLOCATION**

A Thesis Proposal  
Presented to the Faculty of the  
Department of Computer, Information Sciences and Mathematics  
University of San Carlos

In Partial Fulfillment  
of the Requirements for the Degree  
**BACHELOR OF SCIENCE IN COMPUTER SCIENCE**

By  
**JOHN DAVES S. BAGUIO**  
**BILLY A. LU**

**CHRISTINE F. PEÑA, MMath**  
Faculty Adviser

January 14, 2023

## **ABSTRACT**

Text classification is a Machine Learning (ML) technique in the domain of Natural Language Processing (NLP), that assigns a set of predefined categories to open-ended text. It is used to handle problems such as spam filtering in emails, and document classification. Text classification has also been used in a variety of other domains such as social media. Twitter, one of the most popular social media platforms, is a place for users to provide public comments called tweets. Through social media, people are able to convey information easily, especially with information concerning climate change. Climate change refers to long-term changes in temperature and weather patterns in a particular area. Also, global warming is also connected with climate change, which is the long-term heating of Earth due to industrial human activities. Both problems not only affect the Earth's atmosphere, but also all lifeforms. Even with scientific facts by scientists and environmentalists, there are still some who spread misinformation, and say that climate change overall is a hoax, or doesn't exist. This study aims to classify climate change tweets in the given dataset by creating a text classification model through an Artificial Neural Network (ANN) model architecture, incorporated with techniques such as word embedding generation through fastText, and topic modeling through Latent Dirichlet Allocation (LDA). A dataset that consists of 43,943 rows with 2 attributes, being the tweet and its given sentiment, will be split to training and testing. A 10-fold cross validation will be performed to verify the accuracy of the model, with performance metrics such as accuracy, precision, recall, and f1 score.

## TABLE OF CONTENTS

ABSTRACT .....	ii
TABLE OF CONTENTS .....	iii
LIST OF FIGURES .....	v
LIST OF TABLES .....	vi
CHAPTER 1 INTRODUCTION .....	1
1.1 Rationale of the Study .....	1
1.2 Statement of the Problem .....	2
1.2.1 General Objective .....	3
1.2.2 Specific Objectives .....	3
1.3 Significance of the Study .....	3
1.4 Scope and Limitation .....	3
CHAPTER 2 REVIEW OF RELATED LITERATURE .....	5
CHAPTER 3 TECHNICAL BACKGROUND .....	14
CHAPTER 4 DESIGN AND METHODOLOGY .....	21
4.1 Dataset Source and Interpretation .....	21
4.2 Dataset Treatment .....	21
4.3 Conceptual Framework .....	21
4.3.1 Dataset Preprocessing with LDA .....	23
4.3.1.1 Tokenization of Tweets .....	23
4.3.1.2 Tweet Preprocessing .....	23
4.3.1.3 Create a Dictionary .....	24
4.3.1.4 Create a Document-term Matrix .....	25
4.3.1.5 LDA .....	25
4.3.1.6 Add topics next to each word .....	26
4.3.1.7 Detokenize Tweets .....	26
4.3.2 Word Embedding Generation .....	28
4.3.2.1 Split into training & testing .....	28
4.3.2.2 Create FastText Word Embeddings .....	28
4.3.3 Developing a Text Classification Model .....	30
4.3.3.1 Text Classification Model .....	30

4.3.3.2 Classification Results	30
4.4 Development Model	30
4.5 Development Approaches	32
4.6 Software Development Tools	33
4.7 Project Management	35
4.7.1 Schedule and Timeline	35
4.7.2 Responsibilities	36
4.7.3 Budget and Cost Management	37
4.8 Verification, Validation and Testing	38
BIBLIOGRAPHY	40
CURRICULUM VITAE	

## LIST OF FIGURES

Figure 1: A simple ANN .....	15
Figure 2: Word Representations visualized in a vector space .....	15
Figure 3: FastText classifier model architecture.....	17
Figure 4: Text Data Hierarchy.....	19
Figure 5: A dictionary named numbers in Python .....	20
Figure 6: Conceptual Framework .....	22
Figure 7: Conceptual Framework of the study : .....	22
Dataset Preprocessing with LDA	
Figure 8: Conceptual Framework of the study : .....	28
Word Embedding Generation and	
Developing a Text Classification Model	
Figure 9: ANN for generating FastText Word .....	29
Embeddings, being matrix between the	
Input Layer and the Hidden Layer	
Figure 10 : Kanban Development Model .....	31
Figure 11: Application of Kanban to the study .....	32
Figure 12: Bottom-up Approach Model .....	32
Figure 13: 10-fold cross validation .....	39

## LIST OF TABLES

Table 1: The idea behind word embeddings -----	16
Table 2: Document-Term Matrix -----	20
Table 3: Tokenized Tweet -----	23
Table 4: Tweet Preprocessing -----	24
Table 5: Adding words from the ----- preprocessed tokenized tweet to the Dictionary	25
Table 6: Document-Term Matrix -----	25
Table 7: Input, Process, and Output of LDA -----	26
Table 8: Adding topics next to each word and Detokenize tweets process -----	27
Table 9: Document-Term matrix, and rows ----- are considered a one-hot vector	29
Table 10: Software Development Tools -----	34
Table 11: Gantt Chart of Activities, First Semester, A.Y. 2022 - 2023 -----	36
Table 12: Gantt Chart of Activities, Second Semester, A.Y. 2022 - 2023 -----	36
Table 13: Table of Roles and Responsibilities -----	37
Table 14: Table of Expenses -----	38

# **CHAPTER 1**

## **INTRODUCTION**

### **1.1 Rationale of the Study**

Text classification has been widely used to solve current real world problems, such as document classification, and spam filtering in emails. The process of classifying texts involves applying machine learning algorithms to group materials into predefined categories, and is considered as one of the most crucial ways to exploit the enormous volumes of information that are available in unstructured textual format (Altinel & Ganiz 2018). Aside from typical domains such as data mining, machine learning, and information retrieval, text classification has been involved in studies such as target marketing, and medical diagnosis (Aggarwal & Zhai, 2012). However, diverse domains for text classification have been emerging. Although many machine learning approaches maintained reliable performances on typical domains, diverse domains contain a number of complex documents and texts that require an even more deeper understanding of machine learning methods (Kowsari et al. 2019).

One diverse domain that has been searched upon is social media. Through social media, people can now freely and without any restrictions communicate with one another. People have the freedom to express themselves and engage with others through social media especially on Twitter. As a result, social media may now be regarded as a source of various types of information, more importantly textual data. Social media contains various discussions and topics, and climate change is no exception. Twitter is home to various public expressions, wherein opinions or messages called “tweets” are easily posted talking about different topics such as climate change, and the site also provides trends on how this topic is discussed at different times and places (Fownes et al., 2018).

Climate change has been a contentious topic over the past few decades despite being clearly observable from melting ice caps, depletion of the ozone layer, and rising temperatures. According to Tweets from 2013 to 2020, "climate change" mentions increased on average by 50% during the 7-year study period, indicating that many users on Twitter are concerned about it (Udelson and Nathalia, 2021). Despite the fact that its consequences are already apparent, some people do not think it exists. In a 2021 study, 6.8 million tweets were collected to examine twitter's fake news discourses around climate change and global warming. The study found that the majority (55.8%) of the top 500 most retweeted posts were categorized as representative of the belief that climate change either is not happening or that its causation is unrelated to human activities. Tweets that reflected behind belief in anthropogenic-caused climate change lagged far behind at 31.4% of the top 500, while 12.8% of the tweets were ambiguous or unclear (Al-Rawi et al., 2021). The hoax discourse pertaining to climate change shows 2 sides of the discussion, one being deniers of scientific facts, and the other being manufacturing false alarm (Brüggemann et al. 2020).

Thus, with these reasons in mind, this research aims to use machine learning in the domain of natural language processing in order to create a text classification model to quickly and reliably categorize tweets linked to climate change in the given dataset. Techniques such as the topic modeling technique LDA to enhance tweets for better text classification & the word embedding generation method through fastText will be used to improve the performance on the text classification model being created.

## **1.2 Statement of the Problem**

### **1.2.1 General Objective**

This research aims to develop a text classification model using an ANN feed forward architecture, with fastText word embeddings enriched with the topic modeling technique LDA.



### 1.2.2 Specific Objectives

1. Preprocess the dataset.
2. Perform LDA to the given preprocessed dataset and embed these topics to each word of every tweet.
3. Develop a text classification model using an ANN feed forward architecture, with fastText word embeddings enriched with the topic modeling technique LDA.
4. Test and evaluate the performance of the model using metrics such as accuracy, precision, recall, F1 score.

### 1.3 Significance of the Study

The outcome of this study may be relevant to the following parties:

**General Public.** The study will enable the general public to classify texts into those that support or oppose climate change.

**Researchers.** The researchers can apply what they have learned and be able to develop a work that can be benefited by the public. Moreover, through this research study, the researcher will be able to learn more and hone their skills in the field of computer science.

**Future Researchers.** The study provides a starting point for more advancement in the research field and may be used as a gateway for additional areas of study. The research study would also contribute in providing new knowledge in classifying text regarding climate change.

### 1.4 Scope and Limitation

The dataset used for this study only features climate change tweets ranging from April 27, 2015 to February 21, 2018. Text classification of climate change tweets will only be done with the tweets from the dataset being used.

Thus, recent climate change tweets in the present are not taken into account for. Dealing with a climate change dataset is only taken into account for this study and proposed text classification method, because this is the topic of interest of the researchers.

## **CHAPTER 2**

### **REVIEW OF RELATED LITERATURE**

#### **Text Classification optimized with word embeddings**

Word Embeddings are vector representations of words. These vector representations of words can also be visualized in a 2 dimensional vector space. A vectorized representation of words in a vector space helps machine learning algorithms to achieve better performance in NLP tasks by grouping similar words. Word Embeddings are manually created or pre-trained models that capture the context of a word in a given text. Word embedding techniques are widely used in sentiment analysis. In a study of Rezaeinia et al. (2017), they proposed a word embedding method called Improved Word Vectors (WIV), which increased the accuracy of pre-trained embeddings in sentiment analysis. They used Part-of-Speech (POS) tagging techniques, lexicon-based approaches, and Word2Vec/GloVe methods. The word embedding method improved Google's pre-trained Word2Vec to over 2%.

A study of Rudkowsky et al. (2018) used word embeddings for a supervised machine learning procedure for estimating negativity in parliamentary speeches. They concluded that this procedure is a potential word embedding approach in social sciences that has the potential to improve bag-of-words approaches in sentiment analysis.

Furthermore, a study of Yang et al. (2018) used word embeddings for Twitter election classification. They investigated the training of their word embedding and how it was influenced by the corpora they used. They compared the word embedding they created to other word embeddings of different corpora such as Wikipedia articles and Twitter microposts. They concluded that there is a significance in choosing the correct word embedding model given a classification task when compared with other word embedding models.

Merely reading a tweet's text can reveal whether it supports the topic or not, but categorizing a huge number of them according to their opinions requires time and work. Opportunely, this can be solved through text classification, a

natural language processing technique where text is categorized into groups. With NLP, text classifiers can automatically examine the text and then categorize it according to its content using a set of predetermined tags. To manually classify text, on the other hand, can be challenging and might lead to erroneous results, especially when working with large amounts of data.

Fortunately, word embedding, a method of natural language processing, can be used to tackle this issue. In this method, each word is represented as a real-valued vector in a smaller-dimensional space, capturing the semantics of the relationships between words. There are already lots of word embedding techniques that are present and can be used such as Word2Vec, GloVe, and FastText. But, out of those mentioned word embeddings, FastText gives a more advantage in terms of time and in handling vocabulary words.

FastText is a library developed by Facebook's AI Research (FAIR) team for word embedding and text classification learning ([fasttext.cc](https://fasttext.cc)). It is another word embedding method that expands on the word2vec paradigm. Instead of learning word vectors right away, FastText renders each word as an n-gram of characters. This enables the embeddings to comprehend suffixes and prefixes and helps to grasp the meaning of shorter words. Rare words fit the bill nicely. A big advantage over other embeddings is that even if a word was not observed during training, it may still be broken down into n-grams to obtain its embeddings. (Reddy, 2021).

### **Climate Change and its effects**

In 2018, the US National Climate Change Assessment concluded that mostly as a result of human activity, Earth's climate is changing quicker now than it has ever done throughout the history of contemporary civilization. The United States is already experiencing the effects of global climate change, and those effects are expected to worsen in the future. However, how severe those effects will be in the future will largely depend on the steps that are taken to reduce greenhouse gas emissions and adapt to the changes that will take place (Jay et al., 2018).

Moreover, in Intergovernmental Panel on Climate Change in 2022 stated the observed impacts from climate change, observed increases in the frequency and severity of climate and weather extremes, including hot extremes on land and in the ocean, heavy precipitation events, drought, and fire weather, have had widespread, pervasive effects on ecosystems, people, settlements, and infrastructure. These observed impacts have been attributed to human-induced climate change, particularly through increased frequency and severity of extreme events. Terrestrial, freshwater, coastal, and open ocean marine ecosystems have suffered significant harm from climate change, and these losses are becoming increasingly irreversible. The frequency and intensity of extreme weather events have increased due to climate change, which has decreased food and water security and made it more difficult to achieve the Sustainable Development Goals (Pörtner et al., 2022).

In a survey conducted by Kabir et al. (2016), the study group's understanding of climate change was ordinary, but there was a high level of perception and awareness of climate change-related events and their effects on health. Education was the main component that contributed to the understanding of climate change and its effects on health. To enhance awareness of climate change and the need for community-level health adaption, school-based interventions should be investigated.

In addition, an article made by Fagan and Huang (2019) stated some facts about how people globally see climate change based on the survey released in 2018 by Pew Research Center, it shows that the majority of nations surveyed believe that the threat posed by global climate change to their country is serious. However, a sizable portion of people consider climate change to be of negligible or no threat. The outcome also demonstrates that, since 2013, concerns about climate change have greatly increased in several countries. The research also demonstrates that those with higher levels of education are more likely to be concerned about climate change.

## **Data Collection Using Data Mining**

Text mining is the process of transforming unstructured text data into machine-processable structured form to discover hidden patterns, also known as a knowledge discovery database from the text (KDT), it deals with the machine learning supported analysis of the textual data. Textual data is extracted from semi-structured and unstructured datasets such as emails, full-text documents, HTML files (Yehia et al., 2019).

A study conducted by Öztürk and Ayvaz (2018) used Twitter and did a text mining approach in analyzing the sentiment about the Syrian refugees crisis. They collected a total of 2,381,297 relevant tweets in two languages including Turkish and English. They used keywords in gathering their dataset. For English Tweets, tweets were gathered with the keywords “Syrian” and “refugee”. For Turkish tweets, they used the corresponding Turkish keywords: “Suriyeli”, “mülteci” and “multeci” for the search.

On the other hand, there was also another method in getting data from twitter, which is through the Hydrator app. Hydrator is an Electron based desktop application for hydrating Twitter ID datasets. Twitter's Terms of Service do not allow the full JSON for datasets of tweets to be distributed to third parties. However they do allow datasets of tweet IDs to be shared (Summer, 2021).

In a 2021 study, Shofiya and Abidi conducted a sentiment analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data in which they used the twitter hydrator tool. They used the twitter hydrator tool to allow the hydration of tweets in JSON as well as CSV format. The hydrated tweets from 20 March 2020, to 20 April 2020, were downloaded into a CSV file. The tweets containing Canada in user location attribute and social distancing related keywords such as “social distancing”, “physical distancing”, “#social distancing”, “#physical distancing” in tweet texts were included in their final data.

In addition, a similar method was done in collecting dataset from twitter. Habbat et al. (2021) conducted a study through topic modeling and sentiment analysis with Latent Dirichlet Allocation and Non-negative Matrix Factorization on Moroccan tweets. He used twitter API in order to extract and analyze Moroccan

tweets. The resulting tweets collected are 25, 146 which are then stored in MongoDB database that later used in generating useful statistics, identifying different sentiments, and extracting then visualizing predominant topics.

### **Social Media as a Data Set**

A study conducted by Aimiwu (2017) about the efficacy of social media to promote green technology. The result of the study shows that social media can be used to reach millions of people to educate and keep them aware of the benefits of various green technologies that can be used to live a green-friendly lifestyle towards sustaining the environment, people, and firms. Also, he also mentioned that the results of this study may encourage humans to use social media to increase the use of green technology to combat the threat of global warming and climate change.

Moreover, a 2017 study conducted by Anderson about the effects of social media use on climate change opinion, knowledge, and behavior concluded that analyses of climate change-related online social content, such as Twitter debates, news comments, and online searches, offer proof of links between social media use and attitudes toward, and behaviors linked to, climate change. Early data points to a number of beneficial effects, including social media supporting increased awareness of climate change, mobilizing climate activists, providing a forum for discussion, and framing the issue in unfavorable terms for society online.

Anderson and Huntington (2017) studied about how twitter discussions of climate change use sarcasm and incivility. They analyzed the sarcasm and incivility in Twitter discussions of climate change during an extreme weather event and found out that instances of incivility and sarcasm were low overall. Incivility was used in association with political topics, and both incivility and sarcasm were used alongside skeptical perspectives of climate change and by those who mention right-leaning politics in their profiles.

Furthermore, a study conducted by Fownes and Margolin (2018) stated that twitter is a good resource for examining public discourse on climate change,

a global problem that sparks political and scientific debate. Twitter gives information on how attitudes on climate change change throughout different periods of time and locales by providing a forum for public expression. It was suggested that research evaluating people's perspectives on climate change may be made better by more closely relating them to statistics about the population that Twitter users represent. Tweets' open-ended content offers more details, such as what subjects are related to climate change and what terminology is used to describe it. They also propose that future research can expand on these findings to include a larger range of climate-related topics.

### **Various NLP Studies about Climate Change**

In a 2015 study, Cody et al. performed a sentiment analysis on tweets about climate change. They examine the "climate" related tweets that were gathered between September 2008 and July 2014. They examine how general sentiment changes in reaction to climate change news, events, and natural disasters using a previously established sentiment assessment tool called the Hedonometer. They discover that while climate rallies, book releases, and green ideas competitions can all lead to a rise in happiness, natural disasters, climate bills, and oil drilling can all contribute to a reduction. The uncovered words in their analysis suggests that, as opposed to climate change doubters, climate change activists are more likely to respond to news about climate change.

A study conducted by Laureiro and Allo (2020) assesses the sentiments and emotions related to climate change in the U.K. and Spain during the first six months of 2019, and how these relate to different preferences and concerns about energy policies. Using natural language processing (NLP) tools such as the NRC Emotion Lexicon that contains a list of English words and their associations with eight basic emotions, they examine tweets on climate change that are affecting both countries. The findings indicate that communications about climate change in the UK are less negative than in Spain, where fear is the feeling most frequently generated. The findings also indicate that these two Western European nations have rather comparable preferences for energy



policies. Particularly, attitudes toward nuclear energy are varied, whereas perceptions of renewable energy sources are associated with positive perceptions and coal is generally negative.

Similarly, there is also a study that analyzes sentiment and uses emotion as a classification in climate change related tweets. Fagbola et al. (2022) analyzed twitter micro-blogging streams to detect emotions and sentiments that surround the Global youth Climate Protest (GloClimePro) with respect to #ThisIsZeroHour, #ClimateJustice and #WeDontHaveTime hashtags. The analysis follows tweet scrapping, cleaning and preprocessing, extraction of GloClimePro-related items, sentiment analysis, emotion classification, and visualization. The results obtained reveal that most people expressed joy, anticipation and trust emotions in the #ThisIsZeroHour and #ClimateJustice action than the few who expressed disgust, sadness and surprise. #ClimateJustice conveys the most positive sentiments, followed by #ThisIsZeroHour and the #WeDontHaveTime. In all their evaluations, they found that there is a considerable number of people expressing fear in the climate action and consequences.

Furthermore, Taufek et al. (2021) conducted a sentiment analysis study about the public perception on climate change. The data utilized came from The Sun Daily's Malaysian Diachronic Climate Change Corpus, which was created. As part of the methodology, sentiment analysis using Azure Machine Learning software was used to explore the polarity of public sentiment. A corpus analysis approach was used to identify the sentiment lexicon, and discourse analysis was used to analyze public sentiment using the identified sentiment lexicon. The findings showed that the majority of public sentiments appeared to be negative, with terms like serious, long, and critical being used to describe them. Positive sentiment words also prevailed such as better, best and hope. Despite having negative attitudes, the population appears to be reasonably knowledgeable about climate change, according to the discourse analysis. However, the public's indignation over how policymakers approach the climate change issue had a significant role in the negative stance.

Moreover, Mucha (2018) conducted a study to analyze how people's perceptions have changed over the years for the past decade using sentiment analysis on Twitter data. He used Naïve Bayes classifier, Multinomial Naïve Bayes Classifier and Support Vector classification and Linear Support Vector classification algorithms to perform the classification using n-gram iterations, TF-IDF and additive smoothing and removing stop words. The testing shows that Linear SVC is considered more accurate among the other classifiers with an accuracy of 71% and was used to analyze Global warming tweets worth of past 10 years. The outcome demonstrates that the trend of tweets about global warming has been rising since 2008. In 2009 and 2010, the proportion of people who believed global warming was a hoax was almost equal to the proportion who believed it to be true. But the fashion has altered. In 2014, the proportion of positive tweets outweighs the proportion of negative tweets. Positive tweet percentages are greater than negative tweet percentages when compared to 2014, yet the positive tweet proportion has begun to decline.

On the other hand, Dahal et al. (2019) used a large dataset of geotagged tweets containing certain keywords relating to climate change and used volume analysis and text mining techniques such as topic modeling and sentiment analysis. To establish the various subjects of discussion, they utilized Latent Dirichlet allocation for topic modeling, and Valence Aware Dictionary and Sentiment Reasoner for sentiment analysis to ascertain the general attitudes and feelings observed in the dataset. Their examination of sentiment reveals that the general tone of the conversation is negative, particularly when individuals are responding to political or extreme weather occurrences. However, topic modeling reveals that while there are many different climate change conversation themes, some are more common than others. Compared to other nations, the USA places less emphasis on policy-related issues when discussing climate change.

Sham and Mohamed (2022) conducted a study that aims to find the most effective sentiment analysis approach for climate change tweets and related domains by performing a comparative evaluation of various sentiment analysis approaches. They employed the SentiWordNet, TextBlob, VADER, SentiStrength,

Hu and Liu, MPQA, and KWWSI lexicon-based techniques. While using two feature extraction approaches, Bag-of-Words and TF-IDF, three machine learning classifiers were used: Support Vector Machine, Naive Bayes, and Logistic Regression. Next, a hybridization of techniques based on machine learning and lexicons was carried out. The hybrid method fared better than the other two strategies, with hybrid TextBlob and Logistic Regression attaining an F1-score of 75.3 percent; as a result, this strategy was picked as the most efficient one. Lemmatization increased machine learning and hybrid approach accuracy by 1.6 percent, according to this study. By improving the accuracy of the Logistic Regression classifier by 0.6 percent, the TF-IDF feature extraction method was marginally superior to BoW.

A study by Effrosynidis et al. (2022) aimed to create a climate change dataset and make it publicly available. They made use of Harvard University's "The Climate Change Tweets IDs dataset" by Littman & Wrubel (2019), which contained almost 40 million tweets that pertain to climate change. They were able to create a dataset with about 15 million tweets with 7 features such as Gender, Stance, Sentiment, Aggressiveness, Temperature, Topics, Disasters, and an additional feature being Geolocation (longitude & latitude). This is done by employing machine learning algorithms and methods, both supervised and unsupervised, such as BERT, RNN, LSTM, CNN, SVM, Naive Bayes, VADER, Textblob, Flair, and LDA.

## **CHAPTER 3**

### **TECHNICAL BACKGROUND**

#### **Text Classification**

Text classification is a technique that assigns a set of predefined categories to open-ended text. It is used to handle problems such as spam filtering in emails, sentiment analysis in various domains, and classifying news articles (Pintas et al., 2021). The administration of content, contextual search, opinion mining, product review analysis, spam filtering, and text sentiment mining are all various uses of text classification (Dalal & Zaveri, 2011).

#### **Topic Modeling**

Topic modeling is a method for unsupervised classification of text documents, and is a statistical tool for extracting latent variables from large datasets. (Blei et al., 2012)

#### **Latent Dirichlet Allocation**

Latent Dirichlet Allocation (LDA) is a topic modeling technique and a probabilistic model for text. It assumes that each document is characterized by a particular set of topics. It's a mixed membership model, where based on the number of topics and document term matrix of the corpus and documents involved as an input, it returns a list of words associated with each topic with high probability, and the assignment of each document to topics. (Blei et al., 2012)

#### **Artificial Neural Network**

Artificial Neural Networks (ANN) are inspired by how the brain works. It is created by simulating a network of neurons in a computer, wherein a set of neurons are most of the time presented as a vector. By applying algorithms that follow how a neuron works, it makes a neural network learn to solve different machine learning tasks such as text classification and image processing. A

simple ANN has an input layer, may have a number of hidden layers, and an output layer (Krogh, 2008).

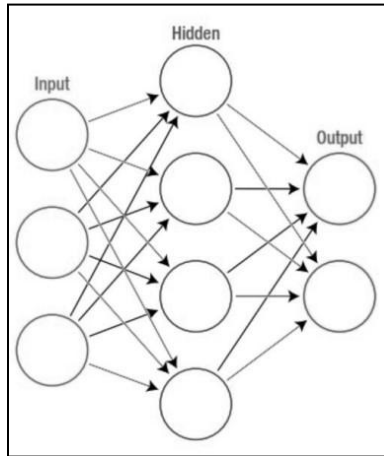


Figure 1: A simple ANN

## Word Embeddings

Word Embeddings are vector representations of words, and follows the Distributional Hypothesis of Harris in 1954, where “Words that occur in similar contexts tend to have similar meanings”. These vector representations of words can also be visualized in a 2 dimensional vector space. A vectorized representation of words in a vector space helps machine learning algorithms to achieve better performance in NLP tasks by grouping similar words (Mikolov et al., 2013). Generating word embeddings uses an unsupervised learning approach for ANNs. This results in the famous example of word embeddings, where: *king - man + woman = queen* is illustrated in a vector space.

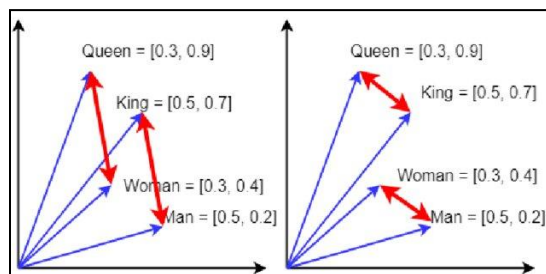


Figure 2: Word Representations visualized in a vector space (Sutor, 2019)

For values in a given vector representation of words, these can be seen as features on how the word is represented. For example, for the word cat, it can be described as: feline, furry, flexible, and so on. With these descriptions, it is then given a score, which is the overall idea of word embeddings, where it can have a high score for some columns if that word fits that feature, and a low score if it doesn't fit a particular feature.

*Table 1*

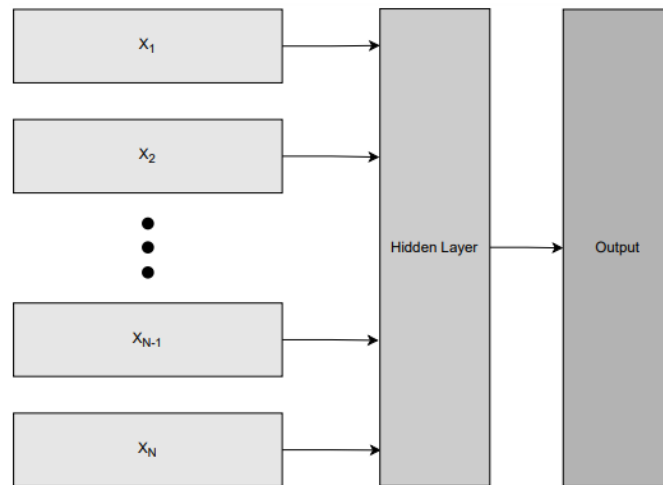
*The idea behind word embeddings*

Word	Feature description				
	animal	furry	flexible	verb	feline
cat	0.8	0.7	0.8	-0.2	0.9
run	-0.4	-0.6	-1.0	1.0	-0.8
human	0.6	0.2	0.3	-0.4	-0.6

## **FastText**

FastText is a library for efficient learning of word representations and sentence classification. It uses word embeddings for unsupervised or supervised learning tasks. For a classification task, it uses an ANN architecture, where the input is the word embeddings of all given words in a given document, and averaged to form the hidden layer, (fasttext.cc).

FastText word embeddings considers the character n-grams of words in order to create an enriched word vector. It emphasizes on the internal structure of a word through character n-grams, which improve vector representations. N-grams are subwords that create the word as a whole. These character n-grams also exist in the word embeddings (Mikolov et al., 2016). For example, the word "climate", with character 3-grams, comes in the form: <cl, cli, lim, ima, mat, ate, te>. These character n-grams solve out of vocabulary words (OOV) to a major extent especially for text classification and word representation.



*Figure 3: FastText classifier model architecture (Mikolov et al., 2016)*

## Climate Change

According to Matthews et al. (2021), the term climate refers to the statistical descriptions of temperature, wind speed, and precipitation, in a general sense “weather conditions”, for a particular location over a span of a time. The span of time of these weather conditions being averaged according to the World Meteorological Organization is about 30 years (WMO, 2021). According to Matthews et al. (2021), it is the change in the state of the climate. It may be caused due to natural internal processes or external forcings. However, the United Nations Framework Convention on Climate Change (UNFCCC) notes that climate change happens due to indirect human activity.

## Global Warming

Global warming refers to the effect on the climate of human activities, such as burning of fossil fuels like coal, oil, and natural gas, and large-scale deforestation, which causes emissions to the atmosphere of large amounts of greenhouse gasses, such as carbon dioxide (Haughton, 2005).

## **Twitter**

Twitter is a social networking and microblogging site. It is a service for communicating and staying connected through fast exchange of messages (“New user FAQ”, n.d.). People or users are able to see what’s happening and what people are currently talking about at the current moment (“About Twitter | Our company and priorities”, n.d.).

## **Tweet**

A Tweet is a message posted on Twitter that may contain photos, videos, GIFs, links, and text. A tweet can have up to 280 characters and up to 4 photos, a GIF, or a video (“How to Tweet – what is a Tweet, keyboard shortcuts, and sources”, n.d.).

## **Artificial Intelligence**

Artificial Intelligence (AI) is a way of making machines think and behave intelligently. The objective of Artificial Intelligence is to find theories and methodologies that can help machines understand the same way humans do, especially in doing tasks (Artificial Intelligence with Python, 2017).

## **Machine Learning**

Machine Learning (ML) is a part of AI. Machine Learning is programming and training computers to be optimized in performance through sample data or past experiences. (Introduction to Machine Learning, fourth edition, 2022)

## **Model**

A model is an output file. The model is developed and trained in order to be predictive to make predictions in the future, or descriptive to gain knowledge from data, or can be trained to perform both (Introduction to Machine Learning, fourth edition, 2022).



## Dataset

A dataset is a collection of data, and characterizes features such as how it's grouped, its overall content, how it's related, and its purpose given a particular field of study (Renear et al., 2011).

## Algorithm

An algorithm is a step-by-step procedure in performing a task (Data Structures and Algorithms in Python). An algorithm's main purpose is to help visualize and have a complete understanding of a task and its steps needed to carry it out and complete it.

## Natural Language Processing

Natural Language Processing (NLP) studies the processing and synthesizing of human languages. This field in AI aims to understand the significance of a group of text, and handles machine learning tasks such as text classification, sentiment analysis, topic modeling, and text generation. (Natural Language Processing in Artificial Intelligence, 2020)

## Corpus, Document, and Token

A corpus is a collection of text data in the form of numerous documents. A document is the building block of a corpus. Tokens are the words that represent a particular document (Mitrani, 2019). In this particular research, the corpus are the tweets that make up the dataset, documents are the individual tweets, and tokens are the words that make up a particular tweet.



Figure 4: Text Data Hierarchy (Mitrani, 2019)

## Document-Term Matrix

A Document-Term Matrix is a representation of a corpus in matrix format. Rows are the documents (tweets), and the columns are the words that appear in the entire corpus. For values in the matrix, it's the number of times each word appears in the corresponding document.

*Table 2*

*Document-Term Matrix*

Document	Term		
	Term #1	Term #2	Term n
Document #1	0	1	...
Document #2	1	1	...
Document #3	1	0	...
Document n	...	...	...

## Dictionary

A dictionary assigns an ID for each word, where each word is unique. The important operation of a dictionary is storing values with their corresponding keys and getting the value back with the key. A dictionary can be thought of as *Key-Value* pairs ([docs.python.org](https://docs.python.org)).

```
numbers = {1: "One", 2: "Two", 3: "Three"}  
Key: Value
```

*Figure 5: A dictionary named `numbers` in Python*

## **CHAPTER 4**

### **DESIGN AND METHODOLOGY**

#### **4.1 Dataset Source and interpretation**

The dataset is acquired online. The dataset contains 43943 tweets that pertain to climate change, and contains 3 attributes, namely: Sentiment, Tweet, and TweetID. This dataset is obtained from kaggle.com. The sentiment is classified into 4 features: Pro, Anti, Neutral, and News. The tweet is labeled Pro if it supports the belief that climate change is manmade, and labeled Anti if it supports the opposite. The tweet is labeled Neutral, if it is neither Pro nor Anti. Lastly the tweet is labeled News if it contains links to news items about climate change.

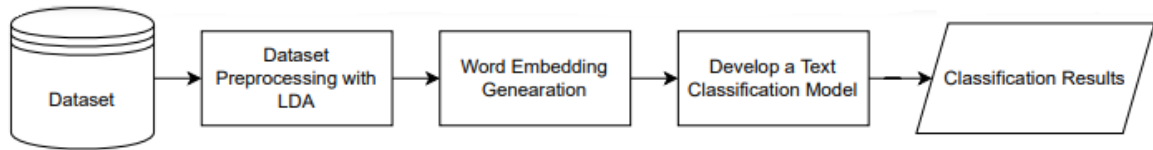
#### **4.2 Dataset Treatment**

The study will use tweets with sentiments labeled as: Pro, Anti, and Neutral. This study focuses on the classification of climate change tweets made by users and not from News items. Thus, tweets in the dataset that are labeled as News are discarded. Thus, before proper Dataset Preprocessing with LDA, tweets labeled as News are no longer present.

#### **4.3 Conceptual Framework**

This study focuses on classifying climate change tweets from the dataset acquired. This study attempts to determine the sentiment of a given tweet, based on the words that the tweet used. The conceptual framework is composed of 3 parts: *Data Preprocessing with LDA*, *Word Embedding Generation*, and *Developing a Text Classification Model*.

Figure 6 illustrates the entire Conceptual Framework of the study.



*Figure 6: Conceptual Framework*

Figure 6 shows the general process on how this study is done. The Dataset is first preprocessed, which also includes LDA. For word embedding generation, the dataset is partitioned into training data and testing data. The training set will be used in creating fastText word embeddings, and are used as an input for the ANN classification model. After training the model with the given training data, the model will then be tested with the testing data. Classification results will then be measured through performance metrics such as accuracy, precision, recall, and f1 score. Each part of the Conceptual Framework will be explained further.

Figure 7 illustrates the first part of the conceptual framework, which is *Dataset Preprocessing with LDA*.

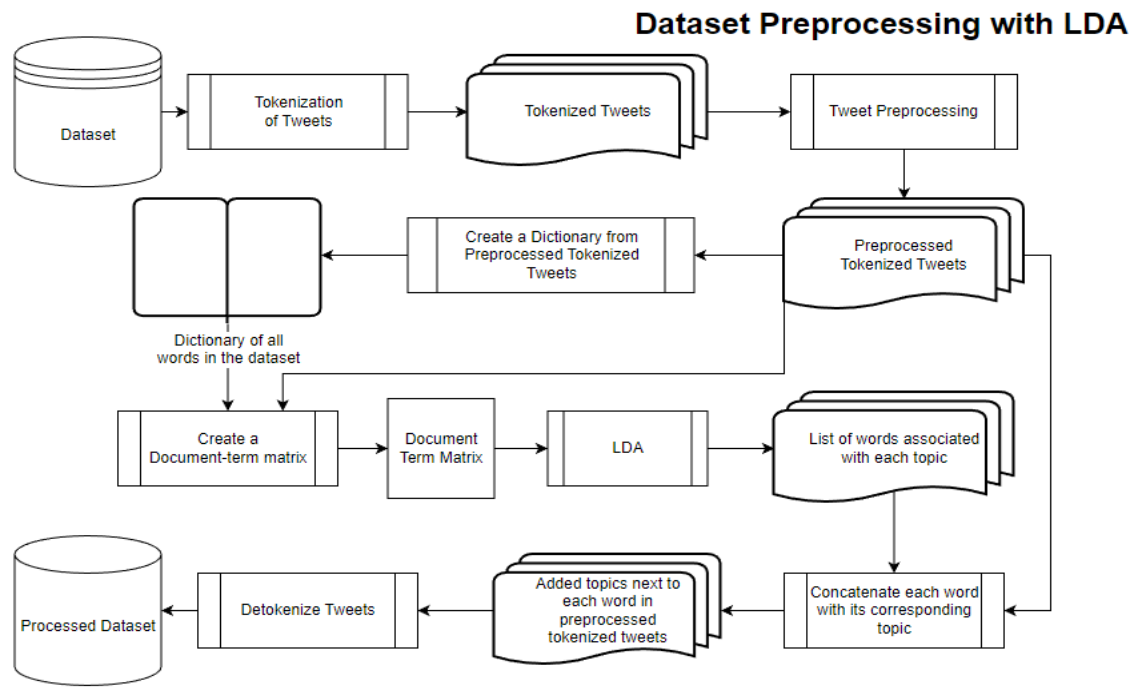


Figure 7: Conceptual Framework of the study : Dataset Preprocessing with LDA

#### 4.3.1 Dataset Preprocessing with LDA


The tweet column will be the feature to be preprocessed in this phase from the dataset being used.

##### 4.3.1.1 Tokenization of Tweets

Tokenization is done to every tweet, wherein each tweet is broken down into a list of words that make up that tweet. This is done to the tweets for easier preprocessing.

Table 3

Tokenized Tweet

Tweet	Tokenized Tweet
	["@ChrisHaines224", "@ZachHaller", "The", "wealthy", "+", "fossil", "fuel", "Industry", "know", "climate", "change", "is", "real.", "Their", "money", "is", "to", "make", "sure", "THEY", "are", "safe", "from", "it."]

##### 4.3.1.2 Tweet Preprocessing

In Tweet Preprocessing, this research makes use of 2 phases, first being the removal of the entire word, and second being removing parts of a word. For phase 1, this includes removing stop words, which majority of these kinds of words are prepositions. After, hashtags are removed, URLs, and lastly mentions For phase 2, special characters are removed, numbers, and all characters for every word are converted to lowercase. In the figure below, it's the example from the previous figure, which was the tokenized tweet. Words highlighted red during every phase are the words or characters to be removed for that phase. Preprocessing can be done

with the help of Python libraries such as *RE*, *SpaCy*, and *NLTK*.

Table 4

*Tweet Preprocessing*

Tokenized Tweet
[“@ChrisHaines224”, “@ZachHaller”, “The”, “wealthy”, “+”, “fossil”, “fuel”, “Industry”, “know”, “climate”, “change”, “is”, “real.”, “Their”, “money”, “is”, “to”, “make”, “sure”, “THEY”, “are”, “safe”, “from”, “it.”]
Phase 1
[“@ChrisHaines224”, “@ZachHaller”, “The”, “wealthy”, “+”, “fossil”, “fuel”, “Industry”, “know”, “climate”, “change”, “is”, “real.”, “Their”, “money”, “is”, “to”, “make”, “sure”, “THEY”, “are”, “safe”, “from”, “it.”]
Phase 2
[“wealthy”, “+”, “fossil”, “fuel”, “Industry”, “know”, “climate”, “change”, “real.”, “money”, “make”, “sure”, “safe”, “from”]
Preprocessed Tokenized Tweet
[“wealthy”, “fossil”, “fuel”, “industry”, “know”, “climate”, “change”, “real”, “money”, “make”, “sure”, “safe”, “from”]

#### 4.3.1.3 Create a Dictionary

A Dictionary is then created from all the preprocessed tokenized tweets, wherein each word is assigned with a unique ID, and all words in the dictionary are unique. This is done in order to create a Document-term Matrix, where the rows of matrix are the preprocessed tokenized tweets, and the columns are the words of the Dictionary. For example, from the preprocessed tokenized tweet, each word is paired with a key, and all other words in every tweet in the dataset are mapped to the dictionary. In trying to add a new word, the Dictionary first checks if that particular word exists or not. If it does exist, then the word will not be added.

**Table 5**

*Adding words from the preprocessed tokenized tweet to the Dictionary*

Preprocessed Tokenized Tweet	Dictionary
["wealthy", "fossil", "fuel", "industry", "know", "climate", "change", "real", "money", "make", "sure", "safe", "from"]	Dictionary = {1: "wealthy", 2: "fossil", 3: "fuel", 4: "industry", 5: "know", 6: "climate", 7: "change", 8: "real", 9: "money", 10: "make", 11: "sure", 12: "safe", 13: "from", 14: ..., nth key: nth word}

#### 4.3.1.4 Create a Document-term Matrix

The Document-term Matrix is created in order to convert text format into numeric data. It is the input for performing LDA.

**Table 6**

*Document-term Matrix*

Tweet	change	climate	energy	evolution	fossil	from	fuel	industry	know	make	money	necessary	ready	real	reduce	revolutionary	safe	steps	sure	take	through	wealthy	nth word
["wealthy", "fossil", "fuel", "industry", "know", "climate", "change", "real", "money", "make", "sure", "safe", "from"]	1	1	0	0	1	1	1	1	1	1	1	0	0	1	0	0	1	0	1	0	0	1	...
[ready, take, necessary, steps, reduce, climate, change, through, revolutionary, energy, evolution]	1	1	1	1	0	0	0	0	0	0	0	1	1	0	1	1	0	1	0	1	1	0	...
nth tweet	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

#### 4.3.1.5 LDA

LDA requires some hyperparameters, which are the number of topics, and the number of iterations. The number of topics is an estimate and an empirical choice for the researchers also with the number of iterations. The output of the LDA Model is a list of words associated for each topic with its given probabilities. The process of LDA can be done with the help of the Python library gensim. What LDA does is it first randomly assigns each word a topic in every tweet. After the random assignment of words in every tweet, LDA will go through every word and topic assignment for each tweet. LDA will look how often the topic occurs

in the tweet, and how often the word occurs in that particular topic overall. So, if the word in the tweet is assigned to a topic that doesn't occur much in the particular tweet that it's, then it will be reassigned to another topic.

*Table 7*

*Input, Process, and Output of LDA*

<b>Input:</b> Document-Term Matrix, <b>Hyperparameters:</b> # of topics, # of iterations
Randomly Assigned topics to each word for every tweet
[wealthy = Topic1, fossil = Topic3, fuel = Topic1, industry = Topic2, know = Topic1, climate = Topic3, change = Topic2, real = Topic1, money = Topic2, make = Topic2, sure = Topic1, safe = Topic3, from = Topic2]
Reassign topics to each word
[wealthy Topic2, fossil Topic2, fuel Topic2, industry Topic2, know Topic1, climate Topic1, change Topic1, real Topic1, money Topic2, make Topic3, sure Topic3, safe Topic3, from Topic3]
Repeat based on the # of iterations
<b>Output:</b> Word Distribution for each Topic <ul style="list-style-type: none"> <li>• Topic #1 - 40% climate, 40% change, nth word</li> <li>• Topic n - ...</li> </ul>

#### 4.3.1.6 Add topics next to each word

For the preprocessed tokenized tweet, each word is concatenated with its given topic to the right, which overall adds a lot of features to each given tweet. These Topics are then renamed, and how these topics are named is an empirical choice for the researchers.

#### 4.3.1.7 Detokenize Tweets

Detokenize Tweets is done so that all tweets are no longer a list of words, and are back to being a tweet as a whole.

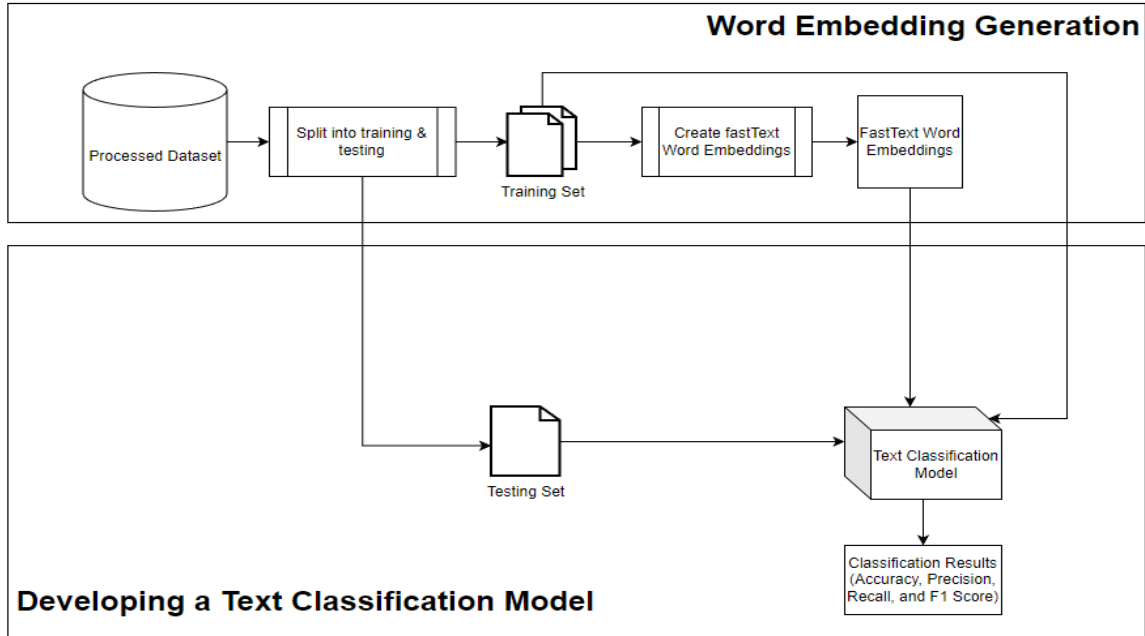


Table 8

*Adding topics next to each word and Detokenize tweets process*

Topics next to each word in the preprocessed tokenized tweet
[wealthy Topic2, fossil Topic2, fuel Topic2, industry Topic2, know Topic1, climate Topic1, change Topic1, real Topic1, money Topic2, make Topic3, sure Topic3, safe Topic3, from Topic3]
<b>Renaming each topic:</b> Topic1 = environment, Topic2 = energy, Topic3 = safety
[wealthy energy, fossil energy, fuel energy, industry energy, know environment, climate environment, change environment, real environment, money energy, make safety, sure safety, safe safety, from safety]
Detokenize Tweet
[wealthy energy fossil energy fuel energy industry energy know environment climate environment change environment real environment money energy make safety sure safety safe safety from safety]

Figure 8 illustrates the second and third part of the Conceptual Framework, which is *Word Embedding Generation* and *Developing a Text Classification Model* respectively.



*Figure 8: Conceptual Framework of the study : Word Embedding Generation and Developing a Text Classification Model*

### 4.3.2 Word Embedding Generation

With the preprocessed dataset from the Dataset Preprocessing with LDA step, this will be used to create word embeddings.

#### 4.3.2.1 Split into training & testing

The dataset is then split into Training and Testing. The Training Set is used in order to generate fastText Word Embeddings.

#### 4.3.2.2 Create FastText Word Embeddings

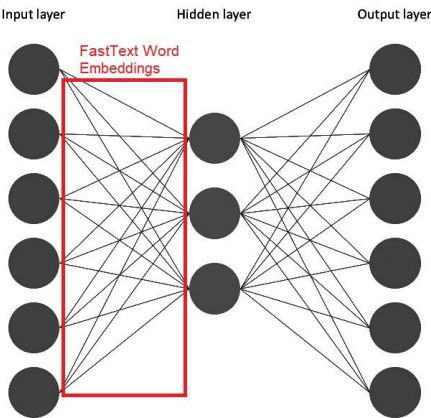
FastText word embeddings generation is done by an unsupervised ANN architecture, wherein the size of the input layer is the same as the size of the output layer. In this case, both the input and output layer of the network is the vocabulary of all words in the given preprocessed dataset. In this case, words are again represented as a Document-Term matrix, but this time, instead of tweets as the row, it's now the individual words, and the representation of words is considered as a One-hot vector. The table

below is an example of a Document-Term matrix, and each row is considered a one-hot vector, which basically means that a word represents itself.

*Table 9*  
*Document-Term matrix, and rows are considered a one-hot vector*

Word	Word			
	climate	change	real	nth word
climate	1	0	0	...
change	0	1	0	...
real	0	0	1	...
nth word	...	...	...	...

With the fastText word embeddings created, it is used as a dictionary when feeding training data to the text classification model, where every word in the training dataset already has its own word vector stored in the word embedding. Words are first searched in the dictionary and converted into its corresponding vectors.



*Figure 9: ANN for generating FastText Word Embeddings, being a matrix between the Input Layer and the Hidden Layer*

### **4.3.3 Developing a Text Classification Model**

With FastText word embeddings created, these are then used to convert words in a training tweet into numeric form, and then fed into an ANN for the text classification task.

#### **4.3.3.1 Text Classification Model**

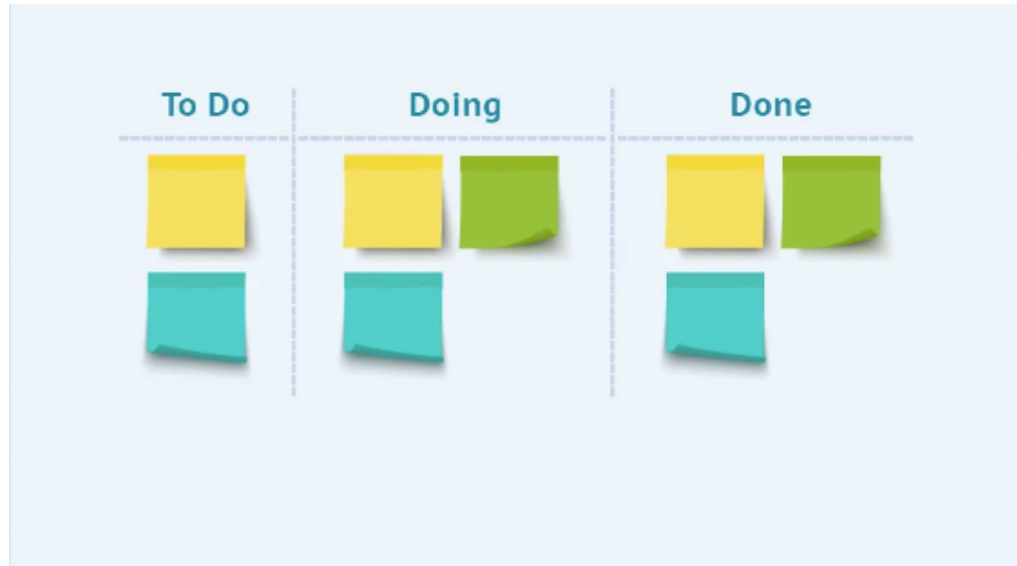
Training data is then used as an input to the created text classifier to train. The text classifier is created using an ANN feed forward architecture. Once training is finished, the text classification model has been created, and will be tested using the Testing Set.

#### **4.3.3.2 Classification Results**

After testing, performance metrics such as accuracy, precision, recall, and f1 score will be measured and used to see the overall performance of the model. K-fold Cross-Validation will also be performed along with the performance metrics mentioned to validate how well the model performs with the dataset overall and to check whether or not the model performs consistently.

### **4.4 Development Model**

The Kanban Method is a technique for creating, overseeing, and enhancing knowledge work flow systems (Agile Alliance, 2021). It is a method that helps teams facilitate the flow of value by visualizing workflow, establishing Work In Process (WIP) limits, measuring throughput, and continuously improving their process (Scaled Agile Framework, 2022).



*Figure 10 : Kanban Development Model*

The **To Do** column is the planning stage of the project development. This is the stage where researchers plan the steps on how to do the research study. Also, the stage where works are organized accordingly. This includes the creation of the model as well as the training and testing of the chosen model.

Furthermore, the **Doing** column are the tasks that the researchers will be working on. This is the stage where researchers will work on necessary changes to make the research study successful.

On the other hand, the **Done** column is the finished work that the researchers did in the study.

In applying Kanban to this study's development approach, each process demonstrated in the conceptual framework, such as Dataset Preprocessing with LDA, Word Embedding Generation, and Text Classification Model Development, will be broken down into smaller processes explained in the conceptual framework. Processes done for every phase should be done sequentially, in order to proceed to succeeding processes and phases of the study. All smaller processes will be in the To Do column, and should be at the Done column in order to proceed to the next process. For example, for Dataset Preprocessing

with LDA, the first step is to do the Tokenization of Tweets, which should first be Done in order to proceed to the next process, which is Tweets Preprocessing.

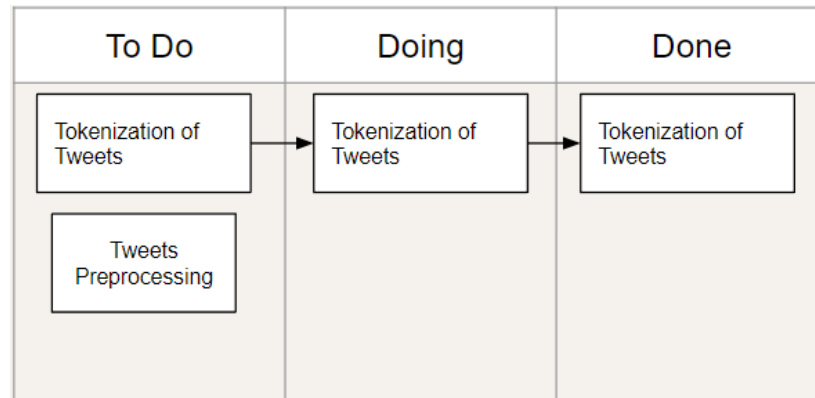


Figure 11: Application of Kanban to the study

#### 4.5 Development Approaches

The development of this research will utilize the bottom-up development approach. This means that the result is achieved by combining and performing different modules in the study.

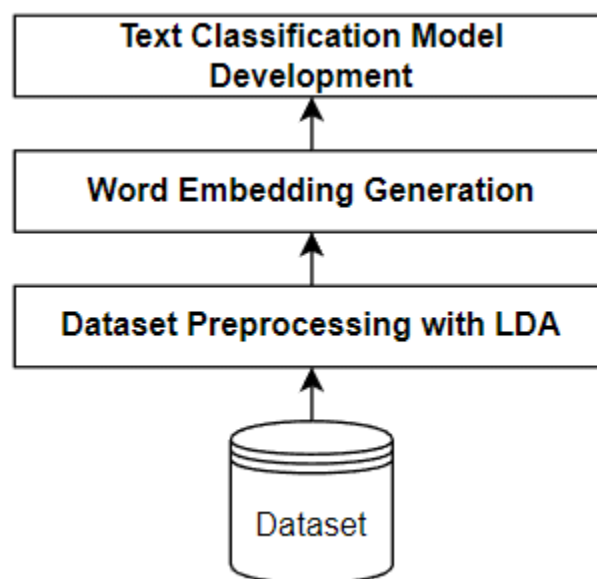


Figure 12: Bottom-up Approach Model

Figure 12 is the Bottom-up Development approach for this research. It shows the modules required in order to reach each succeeding module. These modules are: Dataset Preprocessing with LDA, Word Embedding Generation, and Text Classification Model Development. Each module also has different subprocesses in order to be completed. Following the Kanban Development Model, any kind of module to be worked with is first broken down to smaller processes, and are labeled as to-do. The Bottom-up Approach Model for this research follows a sequential procedure of tasks. Thus, each process is a prerequisite for a succeeding process. The current process being worked on is finished first before succeeding to the next, which will be on the doing column. Once the process is done, it will be on the done column and the researchers will then proceed to the succeeding process.

Modules are to be done sequentially in order to proceed to the next module. Thus, the sequential processes should be done with the first module: Dataset Preprocessing with LDA. This module will be put in the to-do column, wherein it's broken down to its sequential processes, and the first process of this module will be put in the doing column. Once done, it will proceed to the next process in the to do column. Once the Dataset Processing with LDA module is finished, it will then be followed by the Word Embedding Generation module, and lastly with the Text Classification Module, wherein both of these modules will follow how the first module was done: The module is broken into sequential processes placed in the to do column, the first process will then be placed on the doing column, and once placed on the done column, its succeeding process will be on the do column, and so on.

#### **4.6 Software Development Tools**

Table 10 shows various software tools, such as applications, programming languages, and Python packages that will be used in developing the research project, as well as the version and the use of the chosen software development tools.

Table 10  
Software Development Tools

Software	Version	Use
<b>Applications</b>		
Google Chrome	104.0.5112.81	Web browser to access sites such as research journals, articles, etc.
Google Collab	1.0.0	A workspace app for browsers that allows writing and executing python code
draw.io	20.2.3	A software used for making diagrams and charts
GitHub / GitHub Desktop	3.0.5	A software used for software and system development
Git	2.37.1	A source code management software that logs changes in local machine and used in pushing changes in Github
<b>Programming Languages</b>		
Python	3.9.13	The programming language used for data analysis and data visualization. It has many libraries that one can use for machine learning and research
<b>Python Packages</b>		
Pandas	1.4.2	Main library for data analysis in Python
Scikit-learn	1.0.2	Provides performance metrics
NumPy	1.21.5	Features a vast number of functions for arrays/vectors and matrices
Re (Regular Expression)	3.10.6	Data cleaning and text pattern matching
NLTK (Natural Language	3.7	Human language data for applying



Toolkit)		statistical natural language processing
SpaCy	3.3.1	Natural language processing library for dealing with word vectors and creating them
Gensim	4.1.2	Topic modeling, plain text processing, and natural language processing resources
fastText	0.9.2	Creates word embedding representations
googletrans	3.0.0	Translates words to English from any different language and vice versa

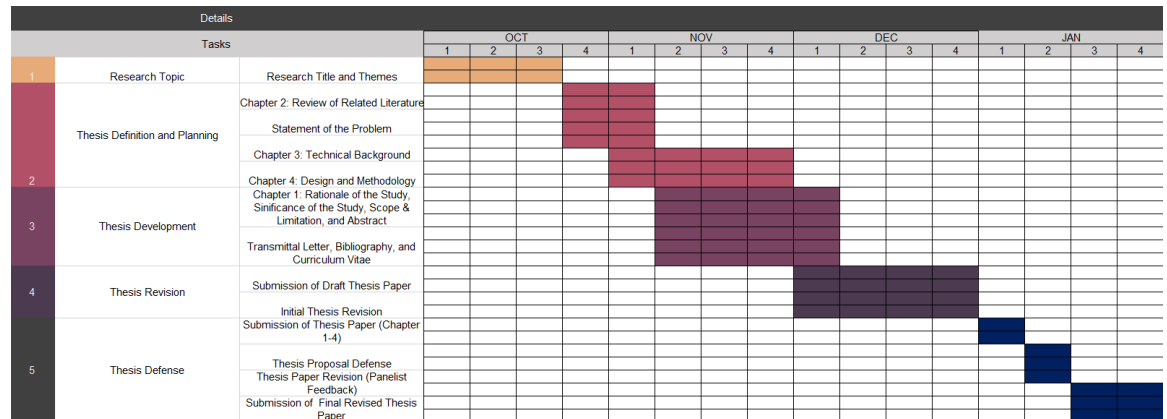
## 4.7 Project Management

This section will discuss the schedule and timeline, tasks, responsibilities, budget and cost management of the team in developing the research project. Each chapter covers both its goal and the reason behind the decision made during project management.

### 4.7.1 Schedule and Timeline

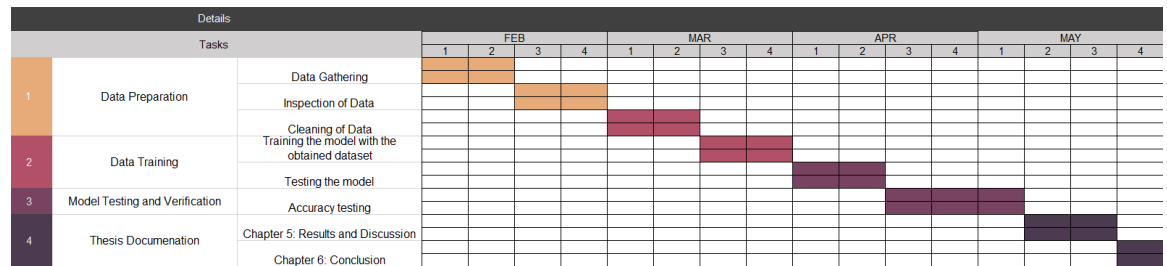
Table 11 shows the tasks that the researchers planned out, and schedules per task when it should be accomplished. Table 11 has the list of tasks on the left, and the timeline of these tasks on the right, with each task spanning a certain amount of numbers (weeks) with the given month on top. Table 11 is the tasks to be completed during the first half of thesis completion.

**Table 11**  
**Gantt Chart of Activities, First Semester, A.Y. 2022 - 2023**



The next table shows the schedule for the second half of the completion of the thesis. Table 12 will be the development of the proposed research methods such as training the model created with the help of the Sentiment dataset, and testing it. Verification and validation of the created model will then be measured.

**Table 12**  
**Gantt Chart of Activities, Second Semester, A.Y. 2022 - 2023**



#### 4.7.2 Responsibilities

Table 13 which lists down the proponents of the study, roles, responsibilities, and task assignments in order to accomplish the objectives of the research.

Table 13

Table of Roles and Responsibilities

Member	Roles	Assignment
John Daves S. Baguio	Researcher and Developer	Understand various literature concepts of climate change and text classification.
		Collection of data to be used in the study.
		Understand core concepts of FastText, Word Embeddings, and LDA.
		Understand core concepts and methods of Natural Language Processing and Text Classification.
Billy A. Lu	Researcher and Developer	Understand core concepts of FastText, Word Embeddings, and LDA.
		Understand core concepts and methods of Natural Language Processing and Text Classification.
		Preprocessing the data to be used in the dataset.
		Train and test the text classification model.
		Evaluating the performance of the text classification model

#### 4.7.3 Budget and Cost Management

Table 14 shows the list of total expenses for conducting the study. The list includes items that the researchers used for this study, such as gadgets with each type of gadget's approximate cost. All items listed down are used during the duration of the study. Items listed are generalized and

are core items, and each kind of item is different for each researcher.

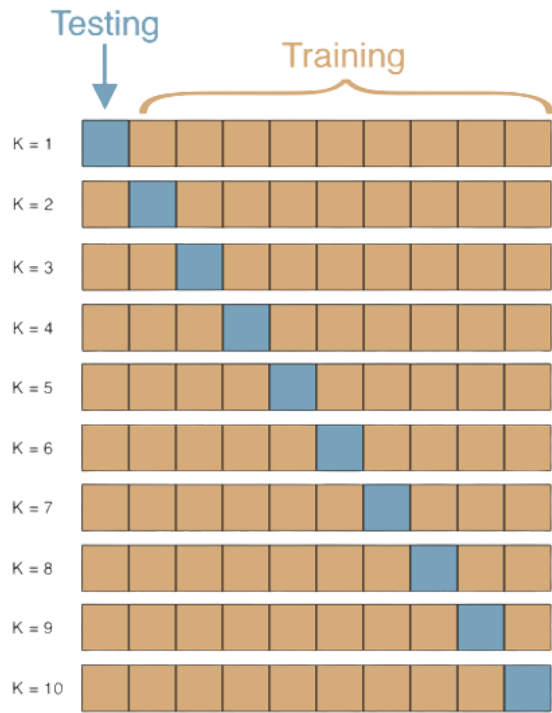
*Table 14*

*Table of Expenses*

Item/Commodity	Cost
Laptop/PC	~ 50,000.00 ₱
Mobile Phone	~ 10,000.00 ₱
Internet	~ 2,000.00 ₱
Miscellaneous	~ 10,000.00 ₱
<b>Total</b>	<b>~ 72,000.00 ₱</b>

#### **4.8 Verification, Validation and Testing**

The text classification model is validated and verified to ensure the overall quality of the model itself. Verification and validation is done by using performance metrics such as K-fold cross validation, wherein the dataset is divided into K subsets, in which each subset will be a testing set, and compared to the remaining subsets to properly measure the accuracy of the model. The letter “K” is a number, which is the number of subsets to be used in the test. This study will be making use of 10-fold cross validation. The accuracy, precision, recall, and f1 score of each of the tests will then be averaged to see the average metrics of the model created.



*Figure 13:* 10-fold cross validation (Dantas, 2020)

## BIBLIOGRAPHY

### Book

- Alpaydin, E. (2020). *Introduction to Machine Learning, fourth edition*
- Joshi, P. (2017). *Artificial Intelligence with Python*
- Mishra, B.K. (2020). *Natural Language Processing in Artificial Intelligence*
- Goodrich, M. (2013). *Data Structures and Algorithms in Python*

### Dictionary

- Atmosphere. (2022). *Merriam-Webster Dictionary*.  
<https://www.merriam-webster.com/dictionary/atmosphere>
- Weather. (2022). *Merriam-Webster Dictionary*.  
<https://www.merriam-webster.com/dictionary/weather>

### Journal Article

- Aimiuwu, E. (2017). Efficacy of Social Media to Promote Green Technology Use.  
<https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=5133&context=dissertations>
- Al-Rawi A., Bizimana, A.J., Kane, O., & O'Keefe D. (2021). Twitter's Fake News Discourses Around Climate Change and Global Warming.  
<https://doi.org/10.3389/fcomm.2021.729818>
- Anderson, A. (2017). Effects of Social Media Use on Climate Change Opinion, Knowledge, and Behavior. *Oxford Research Encyclopedia of Climate Science*. <https://doi.org/10.1093/acrefore/9780190228620.013.369>
- Anderson, A., & Huntington, H. (2017). Social Media, Science, and Attack Discourse: How Twitter Discussions of Climate Change Use Sarcasm and Incivility. *Science Communication*.  
<https://doi.org/10.1177/1075547017735113>
- Altinel, B., & Ganiz, M.C. (2018). Semantic text classification: A survey of past and recent advances. <https://doi.org/10.1016/j.ipm.2018.08.001>
- Bollettino V., Stevens-Alcayna T., Sharma M., Dy P., Pham P., & Vinck P. (2020). Public perception of climate change and disaster preparedness: Evidence

- from the Philippines. *Climate Risk Management*.  
<https://doi.org/10.1016/j.crm.2020.100250>
- Brüggemann, M., Elgesem, D., Beinzeisler, N., Gerts, H.D., & Walter, S. (2020). Mutual Group Polarization in the Blogosphere: Tracking the Hoax Discourse on Climate Change. <https://ijoc.org/index.php/ijoc/article/view/11806>
- Cody, E., Reagan, A., Mitchell, L., Dodds, P.S., & Danforth, C. (2015). Climate Change Sentiment on Twitter: An Unsolicited Public Opinion Poll. *PLOS ONE*. <https://doi.org/10.1371/journal.pone.0136092>
- Dahal, B., Kumar, S.A.P. & Li, Z. (2019). Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining*. <https://doi.org/10.1007/s13278-019-0568-8>
- Effrosynidis, D., Karasakalidis, A., Sylaios, G., & Arampatzis, A. (2022). The climate change Twitter dataset. *Expert Systems with Applications*. <https://doi.org/10.1016/j.eswa.2022.117541>
- Fagbola, T.M., Abayomi, A., Mutanga, M.B., & Jugoo, V. (2022). Proceedings of the 13th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2021). [10.1007/978-3-030-96302-6\\_60](https://doi.org/10.1007/978-3-030-96302-6_60)
- Fownes, J., Yu, C., & Margolin, D. (2018). Twitter and climate change. *Sociology Compass*. <https://doi.org/10.1111/soc4.12587>
- Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. <https://doi.org/10.48550/arXiv.1607.01759>
- Habbat, N., Anoun, H., & Hassouni, L. (2021). Topic Modeling and Sentiment Analysis with LDA and NMF on Moroccan Tweets. *Innovations in Smart Cities Applications Volume 4*. [10.1007/978-3-030-66840-2\\_12](https://doi.org/10.1007/978-3-030-66840-2_12)
- Hill, R.K. (2016). What an Algorithm Is. *Philosophy Technology*. <https://doi.org/10.1007/s13347-014-0184-5>
- Kabir, M.I., Rahman, M.B., Smith, W., Lusha M.A., Azim, S. & Milton, A.H. (2016). Knowledge and perception about climate change and human health: findings from a baseline survey among vulnerable communities in

- Kahan, D., Peters, E., Wittlin, M., Slovic P., Ouellette, L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. <https://doi.org/10.1038/nclimate1547>
- Laureiro, M., & Allo, M. (2020). Sensing climate change and energy issues: Sentiment and emotion analysis with social media in the U.K. and Spain. *Energy Policy*. <https://doi.org/10.1016/j.enpol.2020.111490>
- Medhat, W., Hassan, A., & Korashy, H. (2014). *Ain Shams Engineering Journal*. <https://doi.org/10.1016/j.asej.2014.04.011>
- Mikolov, T., Chen K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. <https://doi.org/10.48550/arXiv.1301.3781>
- Mikolov T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. <https://doi.org/10.48550/arXiv.1310.4546>
- Mitrani, A. (2019). An Overview of NLP. <https://medium.com/@amitrani/an-overview-of-nlp-fe597ed7e8b6>
- Mucha, N. (2018). Sentiment analysis of global warming using twitter data. <https://library.ndsu.edu/ir/bitstream/handle/10365/28166/Sentiment%20Analysis%20of%20Global%20Warming%20Using%20Twitter%20Data.pdf?sequence=1&isAllowed=y>
- Neogi, P.P., Das, A.K., Goswami, S., & Mustafi, J. (2019). Topic Modeling for Text Classification. [10.1007/978-981-13-7403-6\\_36](https://doi.org/10.1007/978-981-13-7403-6_36)
- Noble, W.S. (2006). What is Support Vector Machine?. *Nature Biotechnology*. <https://doi.org/10.1038/nbt1206-1565>
- Öztürk, N., & Ayvaz, S. (2018). Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics*. <https://doi.org/10.1016/j.tele.2017.10.006>
- Renear, A., Sacchi, S., & Wickett, K. (2011). Definitions of *dataset* in the scientific and technical literature. *Proceedings of the American Society for*



<https://doi.org/10.1002/meet.14504701240>

Rezaeinia, S.M., Ghodsi, A., & Rahmani, R. (2017). Improving the Accuracy of Pre-trained Word Embeddings for Sentiment Analysis. *Cornell University*.

<https://doi.org/10.48550/arXiv.1711.08609>

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, S., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*.

<https://doi.org/10.1080/19312458.2018.1455817>

Sham, N.M., & Mohamed, A. (2022). Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches. *Sustainability*.

<https://doi.org/10.3390/su14084723>

Shofiya, C. & Abidi, S. (2021). Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data. *International Journal of Environmental Research and Public Health*.

<https://doi.org/10.3390/ijerph18115993>

Sutor, P., Aloimonos, Y., Fermüller, C., & Stay, D.S. (2019). Metaconcepts: Isolating Context in Word Embeddings. [10.1109/MIPR.2019.00110](https://doi.org/10.1109/MIPR.2019.00110)

Taufek, T.E., Fariza, N., Jaludin, A., Tiun, S., & Lam, K.C. (2021). Public Perceptions on Climate Change: A Sentiment Analysis Approach. *GEMA Online® Journal of Language Studies*. [10.17576/gema-2021-2104-11](https://doi.org/10.17576/gema-2021-2104-11)

Yang, X., Macdonald, C., & Ounis, L. (2016). Using Word Embeddings in Twitter Election Classification. *Cornell University*.

<https://doi.org/10.48550/arXiv.1606.07006>

Yehia, M., Abulkhair M., & Fattouh, L. (2019). Text Mining and Knowledge Discovery from Big Data: Challenges and Promise. *International Journal of Computer Science Issues*. [10.20943/01201603.5461](https://doi.org/10.20943/01201603.5461)

## Web Article

Agile Alliance. (2021). *What is Kanban?*.  
<https://www.agilealliance.org/glossary/kanban/#q=~>

About Twitter | Our company and priorities. (n.d.). *About Twitter | Our company and priorities*. <https://help.twitter.com/en/resources/new-user-faq>

Aslam S. (2022). Twitter by the Numbers: Stats, Demographics & Fun Facts. Omnicore. <https://www.omnicoreagency.com/twitter-statistics/>

Atmosphere. (n.d.). National Geographic Society.  
<https://education.nationalgeographic.org/resource/atmosphere>

Bell J., Poushter J., Fagan M., & Huang C. (2021). In Response to Climate Change, Citizens in Advanced Economies Are Willing To Alter How They Live and Work. *Pew Research Center*.  
<https://www.pewresearch.org/global/2021/09/14/in-response-to-climate-change-citizens-in-advanced-economies-are-willing-to-alter-how-they-live-and-work/>

Dantas, J. (2020). The importance of k-fold cross-validation for model prediction in machine learning.  
<https://towardsdatascience.com/the-importance-of-k-fold-cross-validation-for-model-prediction-in-machine-learning-4709d3fed2ef>

Earth's Atmosphere. (n.d.). National Geographic Society.  
<https://spaceplace.nasa.gov/atmosphere/en/>

*Earth's Atmosphere: A Multi-layered Cake – Climate Change: Vital Signs of the Planet.* (2019). NASA.  
<https://climate.nasa.gov/news/2919/earths-atmosphere-a-multi-layered-cake/#:~:text=Earth's%20atmosphere%20has%20five%20major,Troposphere>

Fagan M., & Huang, C. (2019). A look at how people around the world view climate change. *Pew Research Center*.  
<https://www.pewresearch.org/fact-tank/2019/04/18/a-look-at-how-people-around-the-world-view-climate-change/>

How to Tweet – what is a Tweet, keyboard shortcuts, and sources. (n.d.). *Help Center*. <https://help.twitter.com/en/using-twitter/how-to-tweet>

Krogh, A. (2008). What are artificial neural networks?. <https://people.binf.ku.dk/~krogh/publications/pdf/Krogh08.pdf>

Jay, A., Reidmiller, D.R., Avery, C.W., Barrie, D., DeAngelo B.J., Dave, A., Dzaugis M., Kolian, M., Lewis K.L.M., Reeves, K. & Winner, D. (2018). Overview. In Impacts, Risks, and Adaptation in the United States: Fourth National Climate Assessment, Volume II. *NCA4*. <https://nca2018.globalchange.gov/chapter/1/>

Littman, J., & Wrubel, L. (2019). Climate Change Tweets Iids. *Harvard Dataverse*. <https://doi.org/10.7910/DVN/5QCCUU>

MonkeyLearn Blog. (2021). Topic Modeling: An Introduction. <https://monkeylearn.com/blog/introduction-to-topic-modeling/>

New user FAQ. (n.d.). *Help Center*. <https://help.twitter.com/en/resources/new-user-faq>

Quian, E. (2019). Twitter Climate Change Sentiment Dataset. *Kaggle*. <https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset>

Reddy, S. (2019). GloVe and fastText — Two Popular Word Vector Models in NLP. <https://blogs.sap.com/2019/07/03/glove-and-fasttext-two-popular-word-vector-models-in-nlp/#:~:text=fastText%20is%20another%20word%20embedding,an%20n%2Dgram%20of%20characters>.

Scaled Agile Framework. (2022). *Team Kanban*. <https://www.scaledagileframework.com/team-kanban/#:~:text=Team%20Kanban%20is%20a%20method,and%20continuously%20improving%20their%20process>

Summer, E. (2021). DocNow/hydrator: Turn Tweet IDs into Twitter JSON & CSV from your desktop. *GitHub*. <https://github.com/DocNow/hydrator>

Udelson, A. & Nathalia. (2021). Visualizing the global #ExtremeWeather conversation on Twitter.

<https://developer.twitter.com/en/blog/industry-team-news/2021/visualizing-the-global-extremeweather-conversation-on-twitter>

*What is Text Classification.* (n.d.). MonkeyLearn.

<https://monkeylearn.com/what-is-text-classification/>

Wikipedia. (2022). FastText. <https://en.wikipedia.org/wiki/FastText>

## Reports

Houghton, J. (2005). Global Warming. *Reports on Progress in Physics*.  
<https://doi.org/10.1088/0034-4885/68/6/R02>

Matthews, J.B.R., Möller, R, Diemen, Fuglestedt, J.S., Masson-Delmotte V., Méndez C., Semenov S., & Reisinger A. (2021). IPCC, 2021: Annex VII: Glossary. *In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*.  
[10.1017/9781009157896.022](https://doi.org/10.1017/9781009157896.022)

Pörtner, H.-O, Roberts, D.C., Poloczanska, E.S., Mintenbeck, K., Tignor M., Alegria A., ..., & Okem, A. (2022). Summary for Policymakers. Climate Change 2022: Impacts, Adaptation and Vulnerability. *Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*.  
[https://www.ipcc.ch/report/ar6/wg2/downloads/report/IPCC\\_AR6\\_WGII\\_SummaryForPolicymakers.pdf](https://www.ipcc.ch/report/ar6/wg2/downloads/report/IPCC_AR6_WGII_SummaryForPolicymakers.pdf)

## **CURRICULUM VITAE**

### **CONTACT INFORMATION**

Name: John Daves S. Baguio

Address: Purok Talong, Catarman, Liloan, Cebu, 6000

Telephone: None

Cell Phone: 09773440268

Email: johndaves6240@gmail.com



### **PERSONAL INFORMATION**

Birthday: June 24, 2000

Religion: Roman Catholic

Civil Status: Single

### **EDUCATION**

University of San Carlos

Bachelor of Science in Computer Science

Tertiary Level (2019 - 2023)

De La Salle Andres Soriano Memorial College

Secondary Level (2013 - 2019)

Catarman Elementary School

Primary Level (2007 - 2013)

### **TECHNICAL SKILLS**

C, Java, Javascript, Python, PHP

### **WORK EXPERIENCE**

None

### **TRAINING**

None

**CONTACT INFORMATION**

Name: Billy A. Lu

Address: 17 Sepulveda Corner, Sikatuna Streets, Cebu City,  
Cebu, 6000

Telephone: None

Cell Phone: 09993665493

Email: billylu\_10@ymail.com

**PERSONAL INFORMATION**

Birthday: September 8, 1999

Religion: Roman Catholic

Civil Status: Single

**EDUCATION**

University of San Carlos

Bachelor of Science in Computer Science

Tertiary Level (2019 - 2023)

Colegio de la Inmaculada Concepción - Cebu

Letran de Davao Incorporated

Secondary Level (2013 - 2019)

Mary Immaculate Child Development Academy - The Kiddie Math Science Grade  
School

Primary Level (2007 - 2013)

**TECHNICAL SKILLS**

C, Java, Python, C++, C#, PHP, Javascript

**WORK EXPERIENCE**

None

**TRAININGS**

None