Bachelor of Science in Computer Science & Engineering

# Intention Classification of Social Media Posts During Disasters: A Bangla Language Processing Approach

by

Md.Mohiuddin Hasan

ID: 1904125

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

October, 2024

**Chittagong University of Engineering & Technology (CUET)**
**Department of Computer Science & Engineering**
**Chattogram-4349, Bangladesh.**

---

# Thesis Proposal

Application for the Approval of B.Sc. Engineering Thesis/Project

| | | |
|---|---|---|
| **Student Name** | : Md.Mohiuddin Hasan | Session : 2022-2023 |
| **ID** | : 1904125 | |

| | |
|---|---|
| **Supervisor Name** | : Md.Mynul Hasan |
| **Designation** | : Assistant Professor<br>Department of Computer Science & Engineering |

| | |
|---|---|
| **Department** | : Computer Science & Engineering |
| **Program** | : B.Sc. Engineering |

**Tentative Title** : **Intention Classification of Social Media Posts During Disasters: A Bangla Language Processing Approach**

# Table of Contents

# List of Figures

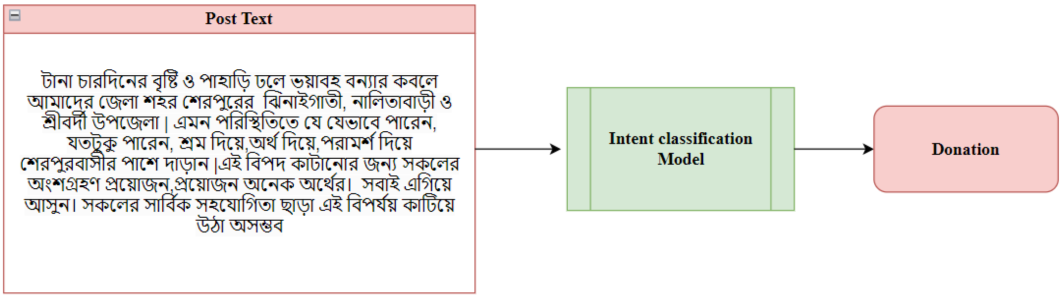
# 1 Introduction

Disasters such as floods, cyclones, and pandemics regularly affect millions of people, particularly in regions like Bangladesh. During these crises, social media platforms like Twitter and Facebook become crucial channels for communication, where affected individuals, governments, and humanitarian organizations share real-time information, request assistance, and coordinate relief efforts. At the onset of a disaster, timely access to crucial information—such as reports of injuries, urgent victim needs, and infrastructure damage—is essential for humanitarian organizations to effectively plan and execute relief operations [1]. However, the vast amount of social media data generated during such events poses significant challenges in filtering critical information and identifying the intentions behind these posts.

While substantial research has been conducted on analyzing social media posts for disaster management, most of this work has focused on posts in English. This has left a gap when it comes to analyzing social media content in other languages, particularly Bengali, which is spoken by millions in disaster-prone regions. The need to accurately classify posts written in Bengali is vital for timely disaster response, as it enables the identification of important information, such as requests for aid or reports of damage, that might otherwise be overlooked.

A significant challenge is the lack of annotated datasets in the Bengali language that can be used to train machine learning models for social media post classification during crises. Existing datasets primarily focus on English, creating a scarcity of resources for Bengali language processing in this context. To address such challenges, this research aims to create a labeled dataset of crisis-related Bengali social media posts, classifying them into categories such as advisory, urgent aid requests, damage reports, donation appeals, and non-relevant information.At the times of natural disasters, millions of people seek updates and share their thoughts and information on social media platforms. Studies have demonstrated that social media data can be beneficial for several humanitarian objectives, including enhancing situational awareness and promoting efficient responses to crises.

The study will focus on developing a system that can accurately classify these posts using both traditional machine learning methods and modern deep learning approaches. The outcome of this research will provide valuable tools for government agencies and humanitarian organizations to quickly identify and prioritize crucial information from social media posts during disasters. This can lead to faster and more effective responses, improving resource allocation and decision-making in disaster-affected areas [2] .



টানা চারদিনের বৃষ্টি ও পাহাড়ি ঢলে ভয়াবহ বন্যার কবলে আমাদের জেলা শহর শেরপুরের  ঝিনাইগাতী, নালিতাবাড়ী ও শ্রীবর্দী উপজেলা | এমন পরিস্থিতিতে যে যেভাবে পারেন, যতটুকু পারেন, শ্রম দিয়ে,অর্থ দিয়ে,পরামর্শ দিয়ে শেরপুরবাসীর পাশে দাড়ান |এই বিপদ কাটানোর জন্য সকলের অংশগ্রহণ প্রয়োজন,প্রয়োজন অনেক অর্থের।  সবাই এগিয়ে আসুন। সকলের সার্বিক সহযোগিতা ছাড়া এই বিপর্যয় কাটিয়ে উঠা অসম্ভব

Figure 1.1: An example of post intent classification

# 2 Motivation

- Natural disasters significantly impact communities, particularly in regions like Bangladesh, making timely and accurate information vital for effective response efforts.
- Despite the critical role of social media, existing research primarily focuses on English-language content, leaving a gap in understanding crisis communication in Bengali, a language spoken by millions in disaster-prone areas.
- Accurately classifying Bengali social media posts is essential for identifying urgent needs, assessing damage, and allocating resources effectively during disasters.
- The lack of annotated datasets for Bengali social media content presents a significant challenge for developing machine learning models capable of processing and analyzing crisis-related posts.

# 3 Background and Present State

In the realm of humanitarian classification for Bengali social media posts, very few studies and research works have been identified. Research in this domain remains relatively unexplored, creating a promising opportunity for advancing natural language processing for Bengali-language crisis-related content. Our investigation into this area revealed that limited research exists on Bengali humanitarian text classification, specifically targeting disaster management and crisis response. This gap highlights the need for more comprehensive work, which could significantly enhance crisis informatics and disaster management capabilities in Bengali.

Karimiziarani and Moradkhani (2023) [2] analyzed public responses to Hurricane Ian using social media analytics, employing techniques like sentiment analysis and topic classification. Their study classified over 21 million tweets into six humanitarian categories: Caution, Damage, Evacuation, Injury, Help, and Sympathy.For classification, they used CNN and GRU-based architectures, achieving a classification accuracy of 93% for humanitarian categories. Furthermore, the use of VADER for sentiment analysis provided insights into public emotional responses during the hurricane. Although this work demonstrated the utility of combining topic classification with sentiment analysis.

Alam et al. (2019) [3] developed CrisisDPS, a system designed to provide data processing services for humanitarian tasks, focusing on disaster type, informativeness, and humanitarian information type classification. CrisisDPS uses a combination of classical algorithms like Naïve Bayes, SVM, and Random Forest, alongside deep learning models such as CNN and LSTM.For humanitarian classification, CNN was found to be the most effective model, achieving an F1-score of 0.78 for classifying social media posts into humanitarian categories such as affected individuals, requests for aid, and damage reports.

Alam et al. (2021) [1] introduced CrisisBench, a benchmark framework consolidating multiple crisis-related social media datasets for humanitarian information processing. CrisisBench combined eight datasets and created a benchmark for tasks such as informativeness classification and humanitarian information type classification, offering a platform to evaluate the performance of different models.In this work, several deep learning models were evaluated, including CNN, fastText, and Transformer-based models like BERT. The best-performing model, RoBERTa, achieved an F1-score of 0.87 for humanitarian classification tasks.

Paul et al.(2022) [4] developed a framework for classifying crisis-related tweets using a combination of deep learning models like CNN and GRU. Their study aimed to categorize tweets into various humanitarian categories, such as damage reports, requests for help, and informational posts. The system showed promising results, achieving a classification accuracy of 93% for disaster-related posts. The use of deep learning models like CNN and GRU enabled the framework to handle large amounts of social media data effectively, helping disaster management teams gain valuable insights during crises.

Yang et al.(2024) [5] explored the detection and categorization of humanitarian needs from Twitter data across various crises, such as the Ukraine-Russia conflict, COVID-19, and natural disasters. Using models like Word2Vec, BERT, and CrisisTransformers, they extracted who-needs-what triples from tweets, categorizing needs into areas such as food, shelter, medical aid, and military resources. The Word2Vec model achieved an average accuracy of 81.08%, outperforming other models in need detection. Their framework for mapping needs was highly effective, with an accuracy of 88% in identifying these relationships.

Devaraj et al. (2020) [6] convolutional neural networks (CNNs) were utilized to classify urgent help requests in disaster-related tweets, achieving an F1 score of 0.87. This study specifically focused on identifying actionable tweets during emergencies, significantly enhancing the ability of first responders to locate urgent assistance requests amidst non-urgent messages.

Makkena et al. (2024) [7] focused on detecting urgency in social media messages during crises, leveraging machine learning and NLP models. They utilized a combination of CNN and BiLSTM architectures to classify disaster-related messages into multiple urgency levels. Various word embeddings, including Word2Vec, GloVe, and pre-trained models like BERT and DistilBERT, were tested on the dataset. Their CNN-BiLSTM model, with only 10% trainable parameters, achieved comparable results to the more complex BERT-based models, with an accuracy of 68% in classifying urgency levels. Their approach demonstrated efficient processing with reduced training time, making it suitable for real-time applications in crisis response systems.

Our aim is to address the existing gap in humanitarian classification for Bengali social media posts by developing an effective NLP-based framework tailored specifically for Bengali-language crisis-related content. In this effort, we will also create a dedicated dataset of Bengali social media posts related to disaster scenarios, which will serve as a valuable resource for training and evaluating classification models. By leveraging state-of-the-art machine learning and deep learning techniques, we seek to enhance the detection and classification of humanitarian information in disaster situations. This research will contribute to improving crisis informatics, enabling more efficient disaster management and humanitarian response in Bengali-speaking regions, ultimately helping to better address the needs of affected communities during emergencies.

# 4 Specific Objectives and Possible Outcomes

The work will be carried out with an aim to achieve the following objectives:

- To collect and curate a comprehensive dataset of crisis-related social media posts in Bengali.
- Exploring machine learning, deep learning & transformer based models for classifying the intention behind the posts.
- To develop an effective framework for intention classification of social media posts during crises & evaluate efficiency

# 5 Outline of Methodology

The methodology consists of the following steps for the categorization of actions in cricket videos. A model for this purpose is shown in fig. 5.3.

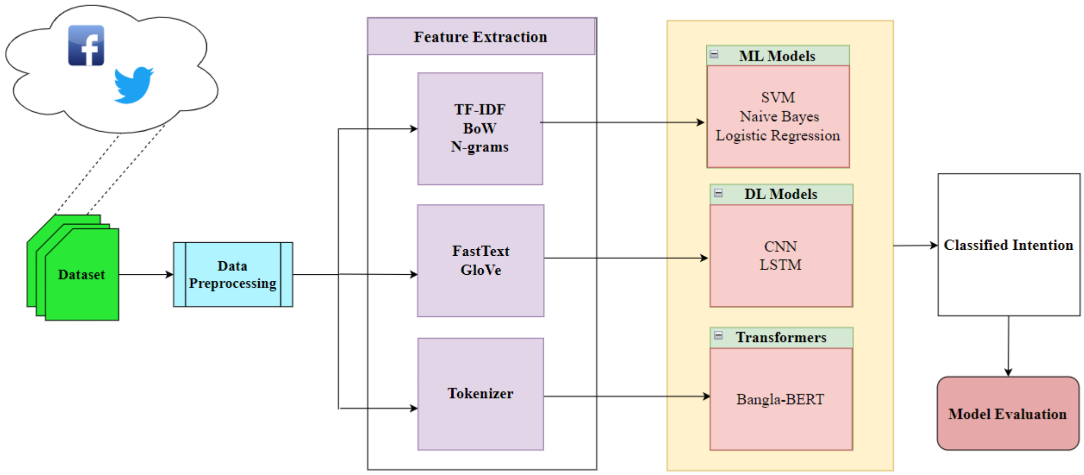


Figure 5.2: Proposed methodology of classifying intention of crisis post.

## 5.1 Dataset

For the purpose of this work, we will collect crisis-related social media posts in Bengali from various platforms, including Twitter and Facebook, as well as local news outlets. Each post will be collected carefully, identifying the intent behind the content. We will focus on categories relevant to humanitarian responses that have significant social implications.After gathering the necessary posts, we will balance the dataset to ensure representation across different intents, followed by data preprocessing.In the existing dataset of English [1], there are 11 classes of cricket actions. These include the following:Caution and advice, Disease related, Displaced and evacuations, Donation and volunteering, Infrastructure and utilities damage, Injured or dead people, Missing and found people, Not humanitarian, Other relevant information, Personal update, Physical landslide, Requests or needs, Response efforts, Sympathy and support, Terrorism related.But we need create dataset for Bengali Language and our classes will be following: advisory, donation, damage, emergency, not Relevant.

## 5.2 Data Preprocessing

Data preprocessing is an essential step in natural language processing (NLP) tasks. It involves transforming raw text data into a clean and normalized format, making it easier for different models to understand and process the text.

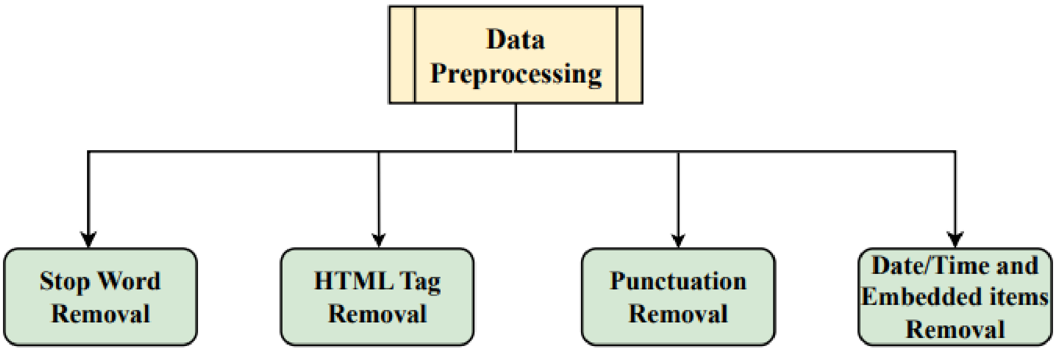


Figure 5.3: Processing of data for intent classification.

### 5.2.1 Stopward Removal

Stopwords are common words in a language that carry minimal meaning and can be safely removed from text without altering the overall meaning. Eliminating stopwords helps reduce computational complexity and noise, allowing models to focus on more significant and content-rich terms during natural language processing tasks.

### 5.2.2 HTML tag removal

HTML tag removal is a fundamental preprocessing step in natural language processing (NLP) tasks, particularly when dealing with raw text data scraped from web pages or online platforms. Web-based text often contains a variety of HTML elements, including tags for formatting (e.g., <p>, <h1>), hyperlinks, images, metadata, and other web-specific constructs that hold no inherent meaning for text analysis. These HTML elements add significant noise to the dataset and can interfere with the performance of downstream tasks such as sentiment analysis, text classification, and named entity recognition. The removal of HTML tags not only reduces noise but also decreases the computational complexity of the models, as the dataset becomes more streamlined and focused on relevant features.Overall, removing HTML tags ensures that subsequent analysis or machine learning models focus exclusively on the meaningful content, which in turn enhances both efficiency and accuracy in tasks like sentiment analysis, classification, or summarization.

### 5.2.3 Punctuation removal

Punctuation removal is a key step in data preprocessing, especially when preparing text for natural language processing (NLP) tasks. Punctuation marks such as periods, commas, question marks, and exclamation points are vital for conveying grammar and meaning in written language. However, in certain NLP applications, like crisis-related text classification or sentiment analysis, punctuation often contributes unnecessary noise rather than meaningful insights. By removing punctuation, the text becomes more focused on word content and relationships rather than grammatical structure. This is typically achieved using regular expressions or string manipulation techniques, streamlining the text for more effective analysis in tasks like ours.

### 5.2.4 Date/time and embedded items removal

Removing date/time references and embedded items is a critical step in preprocessing Bengali text data for crisis-related social media analysis. Date/time removal involves filtering out temporal references, such as dates or specific times, as they are often irrelevant to tasks focused on classifying the urgency or type of request. Regular expressions or specialized tools can be used for this task. Additionally, removing embedded items, such as URLs, usernames, or social media-specific elements, helps reduce noise and privacy risks. These elements, if left in the data, could detract from the core meaning. By employing algorithms or regular expressions to remove such irrelevant content, the text becomes more focused on its semantic meaning, improving the performance of downstream machine learning models.

## 5.3 Feature Extraction

Feature extraction is an essential step in processing text data for natural language processing (NLP) tasks. It involves transforming raw text into numerical or structured representations that capture the key characteristics of the data. This process helps convert textual information into a form that machine learning or deep learning models can analyze. Effective feature extraction ensures that relevant patterns, word relationships, or contextual information are captured, improving the model's ability to classify or predict based on the text. In the context of this work, it allows models to process Bengali crisis-related social media posts more efficiently.

### 5.3.1 TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is an essential feature extraction technique used in natural language processing to convert text data into numerical representations. It evaluates the importance of a word in a document relative to a larger corpus. The Term Frequency (TF) component measures how often a term appears in a document, while Inverse Document Frequency (IDF) assesses the term's relevance across the entire collection of documents. By multiplying TF and IDF scores, TF-IDF highlights significant terms while reducing the influence of common, less informative words. This technique is particularly beneficial in analyzing Bengali crisis-related social media posts, as it helps identify key terms that convey urgency and context, thus enhancing the performance of machine learning and deep learning models.

### 5.3.2 Bag of Words

The Bag of Words (BoW) model is a fundamental technique in natural language processing used for feature extraction from text data. It simplifies the text by treating each document as a collection of words, disregarding grammar and word order while maintaining the frequency of each word. In this model, a vocabulary is created from the entire corpus, and each document is represented as a vector based on the count of words present in it. The BoW model is particularly useful for analyzing Bengali text in crisis-related social media posts, as it enables the identification of frequently used terms that can signify urgency or specific themes. However, while BoW captures the presence and frequency of words, it does not account for context or semantic relationships, which may limit its effectiveness in capturing the nuanced meaning of the text.

### 5.3.3 N-grams

N-grams are a sequence of "n" consecutive words or tokens from a given text, commonly used in natural language processing for feature extraction and analysis. An N-gram can be a single word (unigram), a pair of words (bigram), or a longer sequence, depending on the value of "n." N-grams help in capturing word patterns and context within text data, which is useful for tasks like text classification or sentiment analysis. For example, in the context of Bengali crisis-related posts, bigrams or trigrams could help in identifying specific phrases that indicate urgency or requests for help, improving the accuracy of classification models. N-grams retain more contextual information than a simple Bag of Words model, but as "n" increases, the model becomes more computationally expensive.

### 5.3.4 FastText

FastText is another feature extraction technique commonly used for text classification tasks. It is based on word embeddings and represents words as dense vectors in a high-dimensional space. FastText takes into account the semantic information of words by considering subword information. In this step, data is transformed into FastText word embeddings which capture the contextual meaning of words. These embeddings provide rich and meaningful representations of the text allowing for more accurate aspect classification and polarity detection.

### 5.3.5 GloVe

GloVe (Global Vectors for Word Representation) is an unsupervised algorithm that generates word embeddings by analyzing word co-occurrence within a corpus. It captures both global and local contexts, producing dense vector representations where similar words have similar vectors. GloVe is effective for tasks like sentiment analysis, text classification, and machine translation.

### 5.3.6 Tokenizer

One essential feature extraction method used in natural language processing (NLP) is the tokenizer for transformer models. It transforms unprocessed text into numerical tokens that transformer-based models like BERT (Bidirectional

Encoder Representations from Transformers) or GPT (Generative Pre-trained Transformer) can comprehend and interpret. The tokenizer divides the input text into smaller units, such as words or subwords, and gives each token a distinct numerical ID. By effectively processing and comprehending the text's contextual information, the transformer model is able to capture word dependencies and provide meaningful embeddings for later NLP tasks like text categorization, question-answering, and machine translation. Tokenizers are essential for transformer models to handle natural language input well and operate at the cutting edge.

### 5.4 Classification Model

The modeling phase of this study will involve the application of various machine learning (ML) and deep learning (DL) techniques to classify the intentions of Bengali social media posts related to humanitarian crises. For traditional ML models, algorithms such as Support Vector Machines (SVM), Naive Bayes, and Logistic Regression will be employed to classify the extracted features like TF-IDF, Bag of Words (BoW), and N-grams. These models are known for their simplicity and efficiency in handling text classification tasks. Additionally, deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks will be utilized to leverage the semantic context and sequence of words captured through embeddings like FastText and GloVe. Lastly, a transformer-based model, Bangla-BERT, specifically fine-tuned for the Bengali language, will be employed to capture complex linguistic nuances and context within the social media posts. Each model's performance will be evaluated, and comparisons will be made to select the most effective approach for accurately classifying intention behind the post in the context of humanitarian crises.

# 6 Impact Identification

Here are the potential impacts identified based on the methodology and applications outlined:

## 6.1 Enhanced Disaster Response

The classification system aims to facilitate real-time analysis of social media posts, helping disaster response teams prioritize urgent information. By accurately categorizing posts related to damage reports, emergency needs, and other crisis indicators, agencies can allocate resources more effectively and accelerate response times.

## 6.2 Support for Humanitarian Efforts

Aid organizations will benefit from structured insights into the type and urgency of needs communicated through social media. The framework allows them to identify specific needs—such as requests for food, shelter, or medical aid—enabling them to distribute resources more precisely during crises.

## 6.3 Contribution to Crisis Informatics and Bengali NLP

This research will contribute to the relatively underdeveloped field of Bengali-language crisis informatics. The creation of a dedicated dataset and the application of machine learning and deep learning models in Bengali text classification will support future advancements in NLP for low-resource languages, promoting further research and development.

## 6.4 Data-Driven Decision-Making for Governmental Agencies

Government agencies tasked with disaster management can use this framework to make data-driven decisions, prioritizing actions based on the analyzed social media data. This could result in better-preparedness, more responsive actions, and more efficient recovery efforts following a crisis

# 7 Required Resources

An ideal laptop configuration for the thesis would include a powerful processor (Intel Core i7 or AMD Ryzen 7), ample RAM (16 GB or more), and a solid-state drive (SSD) with a minimum of 512 GB capacity.

## 7.1 Required Tools

- Operating System (ex. Windows or Linux)
- Jupyter notebook or Google Collaboratory

## 7.2 Required Language and Modules

- Python 3.7 64-bits
- Keras or Pytorch(for Deep Learning models)
- Transformers library (for Bangla-BERT)
- Scikit-learn (for traditional machine learning algorithms)
- Gensim (for word embeddings like FastText or GloVe)

# 8 Cost Estimation

Presented below is an example of Cost estimation, which can be tailored and adjusted as per the specific project requirements and needs.

a. Cost of Materials :

| | |
|---|---|
| • Software(Windows 11) | Tk 12000 |
| Total | Tk. 12000 |

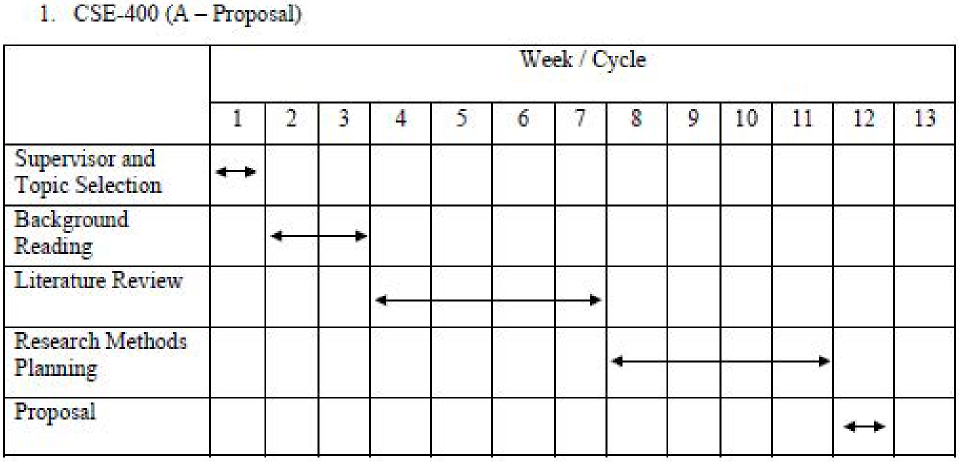
b. Drafting & Binding :

| | |
|---|---|
| • Paper | Tk 500 |
| • Drafting | Tk 800 |
| • Printing | Tk 500 |
| • Binding | Tk 400 |

| | |
|---|---|
| Total | Tk. 2200 |
| Miscellaneous | Tk. 1000 |
| Grand Total | Tk. 17400 |

## 8.1 Time Management

Gantt Chart for the entire timeline of the thesis proposal is provided below, outlining the scheduled tasks and milestones.



1. CSE-400 (A – Proposal)

| | Week / Cycle | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| Supervisor and Topic Selection | ↔ | | | | | | | | | | | | |
| Background Reading | | ↔ | ↔ | | | | | | | | | | |
| Literature Review | | | | ↔ | ↔ | ↔ | ↔ | | | | | | |
| Research Methods Planning | | | | | | | | ↔ | ↔ | ↔ | ↔ | | |
| Proposal | | | | | | | | | | | | ↔ | |

Figure 8.4: The gantt chart for the timeline of the proposal.

# References

[1] F. Alam, H. Sajjad, M. Imran and F. Ofli, 'Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing,' in *Proceedings of the International AAAI conference on web and social media*, vol. 15, 2021, pp. 923–932 (cit. on pp. 1, 3, 6).

[2] M. Karimiziarani and H. Moradkhani, 'Social response and disaster management: Insights from twitter data assimilation on hurricane ian,' *International journal of disaster risk reduction*, vol. 95, p. 103 865, 2023 (cit. on pp. 2, 3).

[3] F. Alam, F. Ofli and M. Imran, 'Crisisdps: Crisis data processing services.,' in *ISCRAM*, 2019 (cit. on p. 3).

[4] N. R. Paul, D. Sahoo and R. C. Balabantaray, 'Classification of crisis-related data on twitter using a deep learning-based framework,' *Multimedia Tools and Applications*, vol. 82, no. 6, pp. 8921–8941, 2023 (cit. on p. 4).

[5] P. Yang, L. Dinh, A. Stratton and J. Diesner, 'Detection and categorization of needs during crises based on twitter data,' in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, 2024, pp. 1713–1726 (cit. on p. 4).

[6] A. Devaraj, D. Murthy and A. Dontula, 'Machine-learning methods for identifying social media-based requests for urgent help during hurricanes,' *International Journal of Disaster Risk Reduction*, vol. 51, p. 101 757, 2020, ISSN: 2212-4209. DOI: `https://doi.org/10.1016/j.ijdrr.2020.101757`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S2212420920312590` (cit. on p. 4).

[7] N. Makkena, A. R. Islam, C. Varol and M. K. An, 'Urgency detection in social media texts using natural language processing,' in *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, IEEE, 2024, pp. 156–163 (cit. on p. 4).

## CSE Undergraduate Studies (CUGS) Committee Reference :

**Meeting No :** **Resolution No :** **Date :**

**Signature of the Student**

**Signature of the Supervisor**

**Signature of the Head of the Department**