# COMPSCI 4NL3 Project Proposal

Group 20

## 1. Team Members

- Raymond Ke (ker)
- Jasraj Johal (johalj11)
- William Clubine (clubinew)
- Neaha Bijo (bijon)

## 2. Task Title and Overview

### 2.1 Task Title

Email Category Classification

### 2.2 Overview

The goal of this project is to build an email classification system that automatically assigns meaningful category labels to emails based on their content. Email classification is a common and useful natural language processing task that is widely used in applications such as spam filtering and inbox organisation into promotions, transactions, work and personal. As people receive a large number of emails every day, being able to automatically organize emails into categories can help users manage information more efficiently.

This task is challenging because emails are often written in an informal and unstructured way or with generative AI. In many cases, a single email may belong to more than one category, such as an email that is both personal-related and transaction-related. For this, the task is framed as a multi-label classification problem where each email can have multiple labels. This involves designing clear annotation guidelines so that team members can label emails consistently, making the task realistic, while still feasible.

## 3. Task Definition

The task is a text classification problem where the input consists of the sender, subject line, and body of an email, and the output is one or more category labels that describe the email's intent or content. The data is unstructured text, and the goal is to automatically predict the appropriate label(s) for each email based on its textual content.

This project is framed as a multi-label classification task, since a single email may belong to more than one category at the same time. The proposed label set includes a fixed number of categories, such as Work, Personal, Transactions, Promotions, Spam, and Social. Each email can be assigned zero, one, or multiple labels depending on its content. This setup reflects real-world email usage and introduces additional complexity compared to single-label and spam/not spam classification.

## 4. Data Sources and Data Collection Plan

https://www.cs.cmu.edu/~enron/

https://www.kaggle.com/datasets/wcukierski/enron-email-dataset/data

### 4.1 Where the data comes from

The dataset for this project will be collected from publicly available email corpora, such as the Enron Email Dataset, which contains real-world emails exchanged in a professional context. We have many potential datasets, from Kaggle and other sources, that we intend to use for verification and in circumstances of lack of diversity. From the full corpus, we will select a subset of emails that are suitable for text classification. Emails that are duplicated, extremely short, or contain no meaningful textual content will be excluded. Only the email sender, subject line, and body text will be used.

### 4.2 Annotation Plan

Emails will be annotated using a pre-defined set of category labels. Multiple labels may be assigned to a single email when appropriate, as the task is framed as a multi-label classification problem. To ensure consistency across annotators, the team will first label a small shared subset of emails and refine the annotation guidelines based on initial disagreements. Inter-annotator agreement will be measured, and any disagreements will be resolved through group discussion before finalizing the labelled dataset.

## 5. Expected Dataset Size

Rather than using the entire email corpus, this project will initially focus on a carefully selected subset of 3,000 emails from the dataset, which will be manually annotated by the team. We anticipate that this subset will be large enough to support meaningful model training and evaluation while being manageable for four manual annotators, but could increase the number of emails selected if needed.

### 5.1 Dataset Plan

This corresponds to several hours of annotation work distributed across all team members, making the annotation process feasible. Each annotated email is estimated to take a minimal amount of time, depending on its length and complexity. In addition to the email text, basic metadata such as the subject line and email length will be retained when available. Preprocessing steps will include removing duplicate emails, normalizing text (e.g., lowercasing), and removing unnecessary formatting or email headers. No external features beyond the email content and basic metadata will be used at this stage.

### 5.2 Example Data Points

**Example 1:**

**Subject:** Confirming your New Subscription to Paperloop.com !

**Email Body:** This message is to confirm your order to Paperloop.com. ….

**Assigned Labels:** Transaction, Spam

**Example 2:**

**Subject:** Expense Report Receipts Not Received

**Email Body:** You are only allowed 2 reports with receipts outstanding. Your expense reports will not be paid until you meet this requirement.

**Assigned Labels:** Work

**Example 3:**

**Subject:** PhD in Finance

**Email Body:** Ronald Melicher advised me to contact you about the PhD program.

I am interested in the field of finance and I wanted to learn about the

PhD finance program at the University of Colorado.

**Assigned Labels:** Personal

## 6. Team Contract

### 6.1. Team Purpose and Mission

Our team is committed to completing a rigorous Natural Language Processing project with technical excellence, clear documentation, and collaborative professionalism. We aim to deliver all milestones on time while ensuring that each member actively contributes to both the technical and analytical components of the project. We will support each other's learning throughout the semester by sharing knowledge, reviewing work constructively, and maintaining a respectful and inclusive team environment.

### 6.2. Duties and Roles

1. All Team Members:

- Complete assigned tasks by agreed-upon deadlines.
- Communicate reasonably well in advance if progress is delayed or if blockers arise.
- Participate in discussions and decision-making via Discord.
- Review teammates' work when requested and provide constructive feedback.
- Maintain professional, respectful, and inclusive communication at all times.

Rotating Leadership (Every 2 Milestones):

- Project Lead: Coordinates meetings when needed, tracks overall progress, ensures alignment with course requirements, and manages final submissions.
- Data Lead: Oversees data collection, scraping, preprocessing, and ensures compliance with ethical and legal standards.
- Annotation Lead: Designs annotation guidelines, coordinates annotation efforts, and monitors inter-annotator agreement and quality.
- Model Lead: Implements baseline models, manages experiments, and ensures results are properly documented and reproducible.

### 6.3. Communication and Decision-Making

Discord will be used as the primary communication platform for all project-related discussions, announcements, and coordination. Team members are expected to check Discord regularly and respond within a reasonable timeframe, especially when deadlines are approaching.

The team is not required to meet weekly. Meetings will be held biweekly or on an ad-hoc basis as needed, and may take place via Discord call or in person. A common meeting time will be communicated and agreed upon in advance.

The team will strive for consensus on major decisions. If consensus cannot be reached after reasonable discussion, a simple majority vote will be used. All members agree to respect and support the final decision.

### 6.4. Conflict Resolution

Level 1 – Direct Communication: Address concerns directly with the involved team member(s) respectfully and promptly.

Level 2 – Team Discussion: If unresolved, the issue will be discussed with the full team and facilitated by the Project Lead.

Level 3 – Resolution: The team will agree on a clear path forward through consensus or majority vote and move forward constructively.

### 6.5. Team Standards

- Work Quality: All code must be readable and documented. Experimental results and analyses must be clearly reported.
- Accountability: Each member is responsible for completing assigned tasks on time and communicating early if assistance is needed.
- Collaboration: Team members will provide constructive feedback, share knowledge, and support one another.
- Academic Integrity: All work will comply with McMaster University academic integrity policies.

### 6.6. Key Dates

- January 20: Project Proposal
- TBD: Data Collection & Annotation Guidelines
- TBD: Annotated Dataset & Agreement Metrics
- TBD: Baseline Models & Benchmark Setup
- TBD: Final Model Submissions & Project Completion

### 6.7. Team Signatures

By signing below, each team member agrees to uphold this contract and work collaboratively and professionally throughout the semester.

| Team Member Name | Signature | Date |
|---|---|---|
| Raymond Ke | RK | 2026/01/20 |
| Jasraj Johal | JJ | 2026/01/20 |
| William Clubine | WC | 2026/01/20 |
| Neaha Bijo | N.B | 2026/01/20 |