

G6 Project Proposal

Teammates: Akshay Dongare (adongar), Ayush Gala (agala2), Vidhisha Kamat (vskamat)

Project Title: The Professional Filter : Machine Learning Approach to Work Email Detection

Dataset description: The Enron Email Dataset [<https://www.cs.cmu.edu/~enron/>] is a collection of approximately 500,000 emails from 150 Enron Corporation employees (primarily senior executives) made public during the company's collapse and subsequent federal investigation.

Project Idea: The goal of the project is to classify emails as either 'work-related' or 'non-work-related' using the Enron dataset. Our own study and the exploratory analysis done by Klimt et al. [1] and Alkhoreyf et al. [3] on the Enron emails revealed that a significant proportion of emails contain jokes, banter, and forwards alongside professional correspondence. Thus, by building a classification model to distinguish work emails from others, we can potentially improve email management and productivity. We plan to explore both supervised and unsupervised learning techniques to model the data which includes topic modeling, neural networks, and decision tree based methods. Since we are working with textual data, we plan to leverage NLP techniques to engineer semantic features that will further help us to classify the email samples.

Our project workflow will consist of 6 stages. Since the emails are in the MIME format they are not immediately appropriate for machine learning tasks. Hence, we will 1) begin by writing a script to parse the emails in the dataset to transform them into an appropriate format. This will be followed by 2) Feature Engineering based on domain knowledge, going beyond the already supplied features in the dataset and extracting relevant features such as the a) number of attachments, b) no. of recipients, c) hyperlinks, d) email length, etc. We will then undertake the crucial task of 3) labeling each email as either 'work-related' or 'non-work related'. To do so we may employ both manual and automated methods to ensure accuracy and speed. After these steps, we will 4) perform data preprocessing as required, including sampling techniques, normalization and transformation of features to meet model requirements and determine the splits for training, validation, and test sets. 5) We plan to train multiple models and 6) evaluate and analyze their performance on key binary classification metrics like accuracy, recall, precision, and F-1 score.

Software we will need to write:

Based on the above description, some of the softwares we will need to write are as follows:

1. An MIME email parser to suitably format the emails in the dataset.
2. A feature engineering module to extract and calculate relevant features from the parsed emails. We have provided a few examples of features in our idea description.
3. A labeling tool to manually label the email samples or automate the labeling process.
4. The main project code to orchestrate the entire workflow from data loading to model evaluation. i.e main python notebook.

Papers to read:

1. "The Enron Corpus: A New Dataset for Email Classification Research" by Klimt and Yang [https://link.springer.com/chapter/10.1007/978-3-540-30115-8_22]. This paper introduces the Enron corpus as a new benchmark dataset for email classification research, analyzing its suitability for email folder prediction and providing baseline results using Support Vector Machines.
2. "Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora" by Bekkerman et al. [<https://ciir.cs.umass.edu/pubfiles/ir-418.pdf>]. This paper presents benchmark experiments on automatically categorizing email into folders using the Enron and SRI email corpora, comparing several classification algorithms and proposing new evaluation methods for email foldering.
3. "Work Hard, Play Hard: Email Classification on the Avocado and Enron Corpora" by Alkhereyf and Rambow. [http://www.cs.columbia.edu/nlp/papers/2017/alkhereyf_email_classification.pdf]. This paper presents an empirical study on classifying emails as business or personal using lexical and social network features, training on the Enron corpus and testing on both Enron and Avocado corpora.

Work Division & Timeline:

- Recurring Meet Link : ALDA Project Meet
(4:15pm to 5:15pm; Monday, Wednesday, and Sunday)
- We will work individually on the components and discuss each aspect together in regular meetings. We will incorporate changes and integrate all the modules collaboratively
- Attendance Report and Notes from Meetings : Notes - ALDA Project Meet

Work Division and Responsibilities:

- Akshay Ashutosh Dongare (adongar) :
 - Exploratory Data Analysis, Data preparation and pre-processing
 - Handling class imbalance if present
 - Implementing, training, evaluating, and testing an Artificial Neural Network for work email classification
- Ayush Gala (agala2) :
 - Labeling emails in the dataset as work and non-work related
 - Engineering additional features for classification and a feature selection/extraction pipeline
 - Model the problem using decision tree-based techniques
- Vidhisha Kamat (vskamat) :
 - Labeling emails in the dataset as work and non-work related
 - Implementing topic modeling using Latent Dirichlet Allocation (LDA)
 - Performing a comparative analysis of the different models and approaches, including artificial neural networks, topic modeling, and decision tree models.