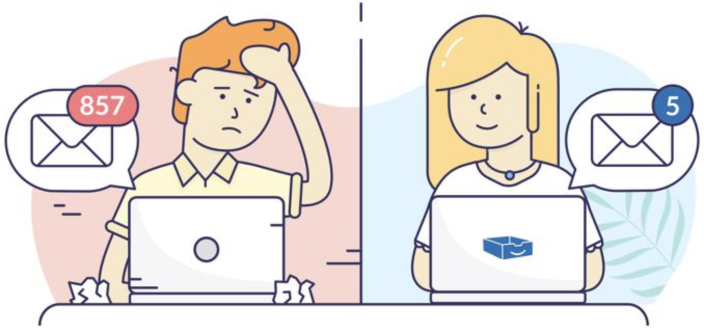
Introducing... ClutterCutter!

The smart, effective tool for your messy inbox



Introduction and Motivation

The Problem

A messy, disorganized inbox makes life more complicated than it needs to be.

Why do we need a solution?

- To save valuable time
- To make email search easier
- To increase the quality of life

The Solution Process

“

1. Provide examples of how we sort past emails
2. Apply ML to learn the patterns
3. Apply this knowledge to make predictions
4. Automated sorting for future emails!

Data & Data Processing

Data Used

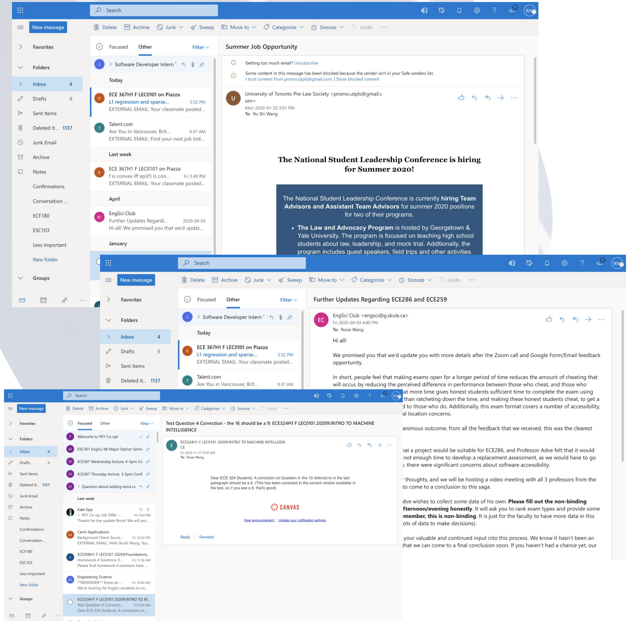
- Real-life emails from our own accounts

Data Cleaning

- Standardized text
- Removed personal info for privacy reasons

Amount of Data

- Initial: 150/class for a total of 750
- Re-sampling: Increased to ~200/class for ~1000 total
- Data Augmentation: Increased to 2850 total
  - Synonym Replacement
  - Back Translation



Software Developer Intern

Today

ECE 367H1 F LEC0101 on Piazza
L1 regression and sparse... 5:52 PM
EXTERNAL EMAIL: Your classmate posted...

Talent.com
Are You in Vancouver, Brit... 9:37 AM
EXTERNAL EMAIL: Find your next job tod...

Last week

ECE 367H1 F LEC0101 on Piazza
f is convex iff epi(f) is con... Fri 3:49 PM
EXTERNAL EMAIL: Your classmate posted...

April

EngSci Club
Further Updates Regardi... 2020-04-03
Hi all! We promised you that we'd updat...

January

Summer Job Opportunity

Getting too much email? Unsubscribe

Some content in this message has been blocked because the sender isn't in your Safe senders list.
I trust content from promo.utpls@gmail.com. | Show blocked content

University of Toronto Pre-Law Society <promo.utpls@gmail.com>
Mon 2020-01-20 3:01 PM
To: Ru Shi Wang

The National Student Leadership Conference is hiring for Summer 2020!

The National Student Leadership Conference is currently **hiring Team Advisors and Assistant Team Advisors** for summer 2020 positions for two of their programs.

- **The Law and Advocacy Program** is hosted by Georgetown & Yale University. The program is focused on teaching high school students about law, leadership, and mock trial. Additionally, the program includes guest speakers, field trips and other activities.

Further Updates Regarding ECE286 and ECE259

EngSci Club <engsci@g.skule.ca>
Fri 2020-04-03 4:40 PM
To: Rosie Wang

Hi all!

We promised you that we'd update you with more details after the Zoom call and Google Form/Email feedback opportunity.

In short, people feel that making exams open for a longer period of time reduces the amount of cheating that will occur, by reducing the perceived difference in performance between those who cheat, and those who
at more time gives honest students sufficient time to complete the exam using
than ratcheting down the time, and making these honest students cheat, to get a
d to those who do. Additionally, this exam format covers a number of accessibility,
al location concerns.

animous outcome, from all the feedback that we received, this was the clearest

at a project would be suitable for ECE286, and Professor Adve felt that it would
not enough time to develop a replacement assessment, as we would have to go
, there were significant concerns about software accessibility.

thoughts, and we will be hosting a video meeting with all 3 professors from the
to come to a conclusion to this saga.

dve wishes to collect some data of his own. **Please fill out the non-binding**
fternoon/evening honestly. It will ask you to rank exam types and provide some
member, this is non-binding. It is just for the faculty to have more data in this
ots of data to make decisions).

your valuable and continued input into this process. We know it hasn't been an
that we can come to a final conclusion soon. If you haven't had a chance yet, our

Welcome to PEY Co-op!

ESC301 EngSci MI Major Option Semi

ECE367 Wednesday lecture, 4-5pm Co

ECE367 Thursday lecture, 3-5pm Confi

Question about adding extra cs

Last week

Kate Epp
PEY Co-op Job Offer - ... Fri 3:43 PM
Thanks for the update Rosie! We will pro...

Certn Applications
Background Check Succe... Fri 12:42 PM
EXTERNAL EMAIL: Hello Rushi Wang, You...

ECE358H1 F LEC0101 20209:Foundations..
Homework 4 Solutions: E... Fri 11:58 AM
Please find homework 4 solutions here. ...

Engineering Science
REMINDER Know an ... Fri 10:40 AM
We're looking for EngSci students to no...

ECE324H1 F LEC0101 20209:INTRO TO M..
Test Question 4 Correctio... Fri 9:50 AM
Dear ECE 324 Students: A correction on ...

Test Question 4 Correction - the 16 should be a 9: ECE324H1 F LEC0101 20209:INTRO TO MACHINE INTELLIGENCE

ECE324H1 F LEC0101 20209:INTRO TO MACHINE INTELLIGENCE
Fri 2020-11-27 9:50 AM
To: Rosie Wang

Dear ECE 324 Students: A correction on Question 4: the 16 referred to in the last paragraph should be a 9. (This has been corrected in the current version available in the test, so if you see a 9, that's good).

CANVAS

View announcement | Update your notification settings

Data Statistics

8872 unique words in 2850 emails!

Below are the 5 most frequently appearing “meaningful” words for each category

Academics:

- 'course': 124, 'students': 121, 'final': 82, 'questions': 82, 'engineering': 81

Alerts:

- 'order': 218, 'email': 169, 'account': 147, 'password': 71, 'information': 70

Personal:

- 'practice': 45, 'think': 32, 'going': 20, 'other': 19, 'email': 19

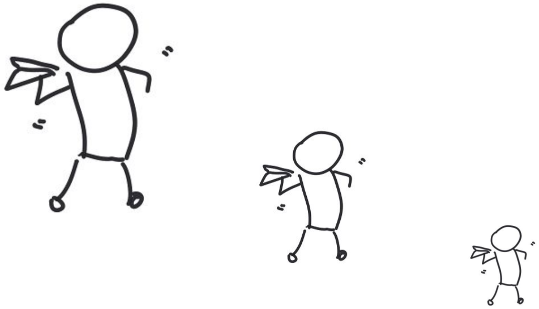
Professional:

- 'application': 177, 'interview': 145, 'position': 100, 'interest': 68, 'assessment': 59

Promotions and Events:

- 'students': 275, 'engineering': 198, 'november': 151, 'student': 144, 'university': 143

Data Statistics cont.d...



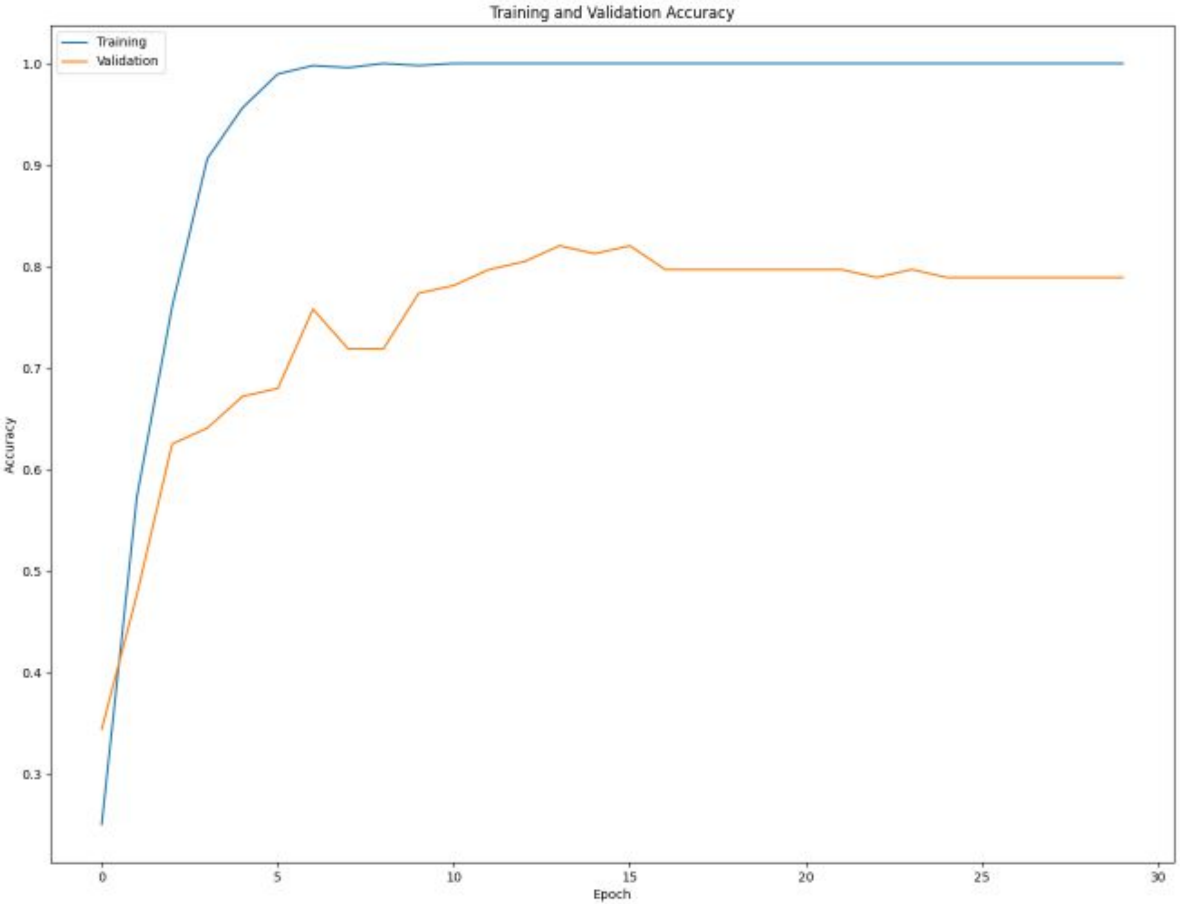
Data Splits

- Training = 0.64
- Validation = 0.16
- Test = 0.20

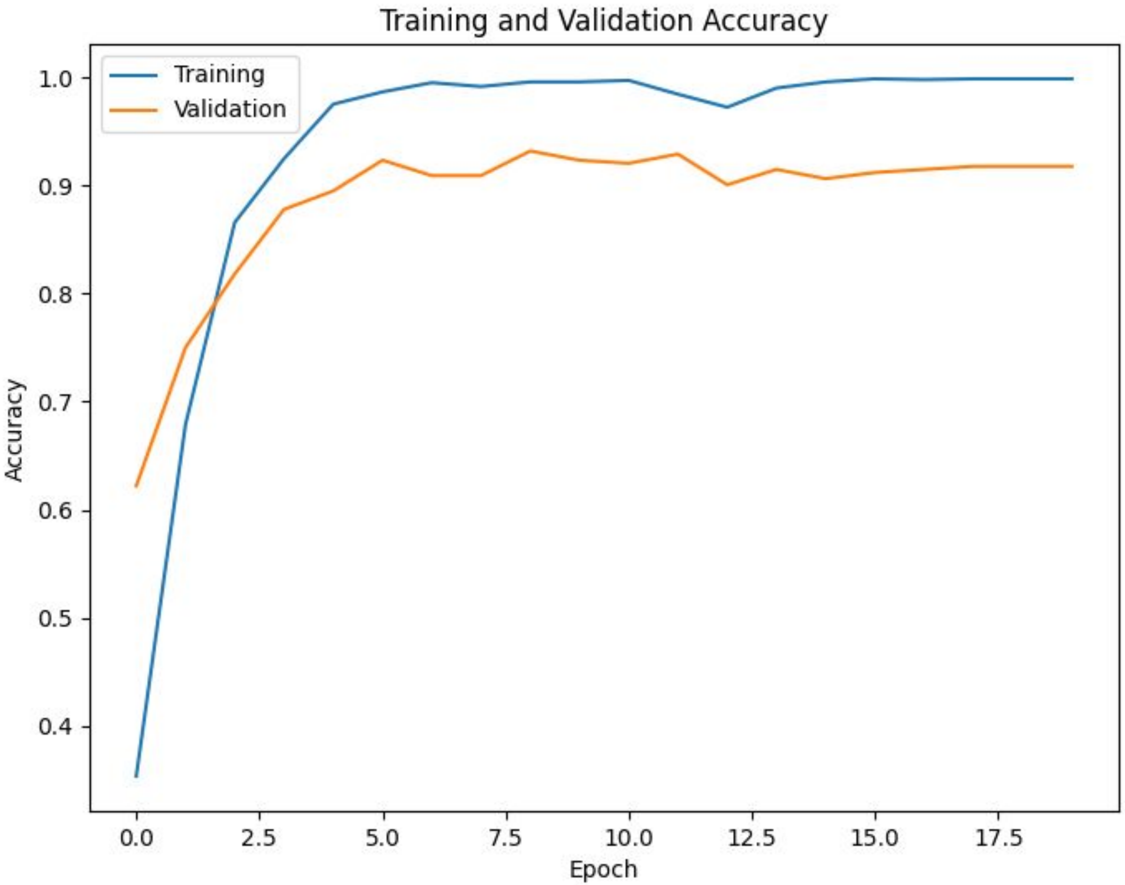
Class Balance

Labels	Train	Validation	Test
Academics	355	89	111
Alerts	370	93	116
Personal	363	91	113
Professional	376	94	118
Promotions and Events	355	88	111

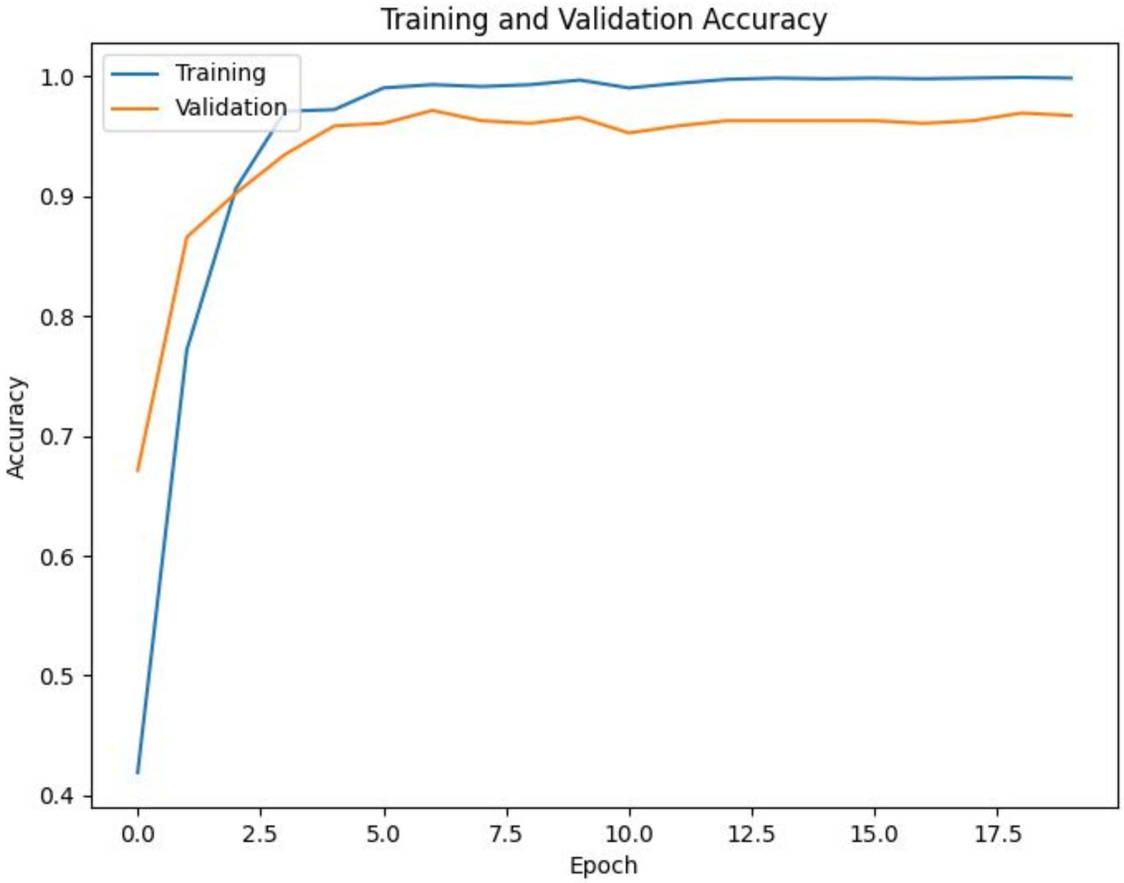
Case 1: No Data Augmentation



Case 2: 35% Data Augmentation



Case 3: 50% Data augmentation



Our Model

Vectorization

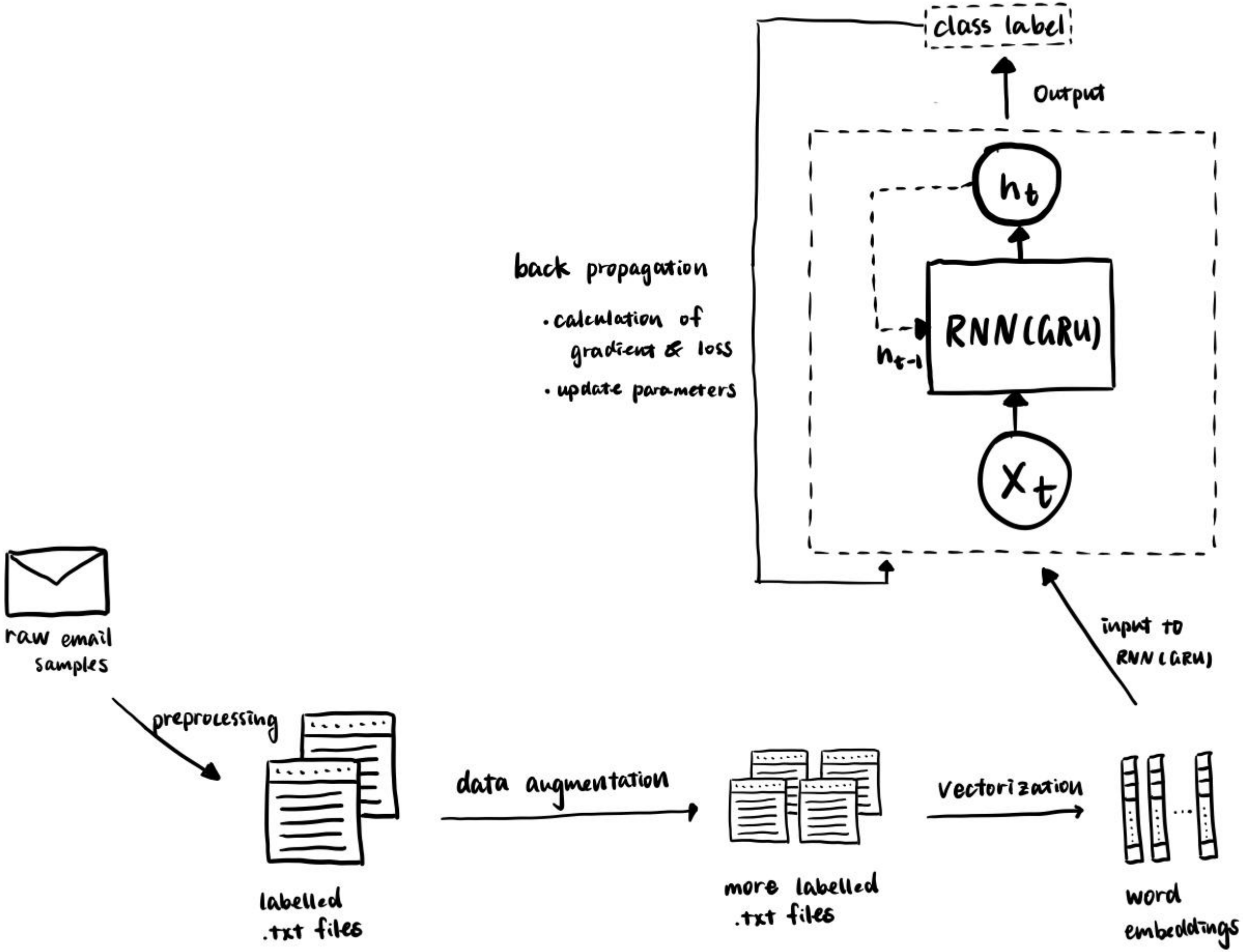
- Word embeddings using Fasttext

Baseline

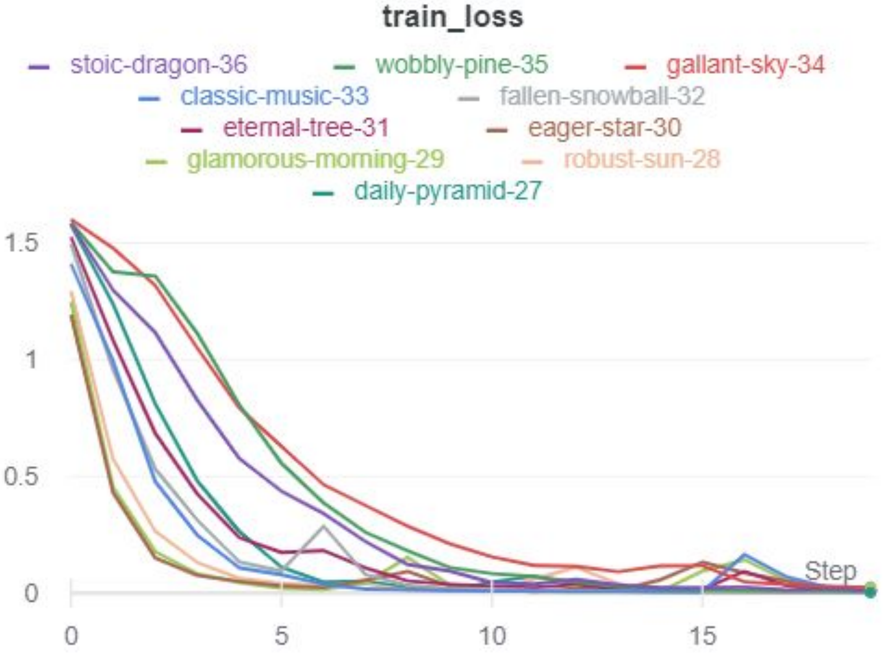
- Took average meaning of each email
- Fed to linear layer to produce 5 outputs

RNN

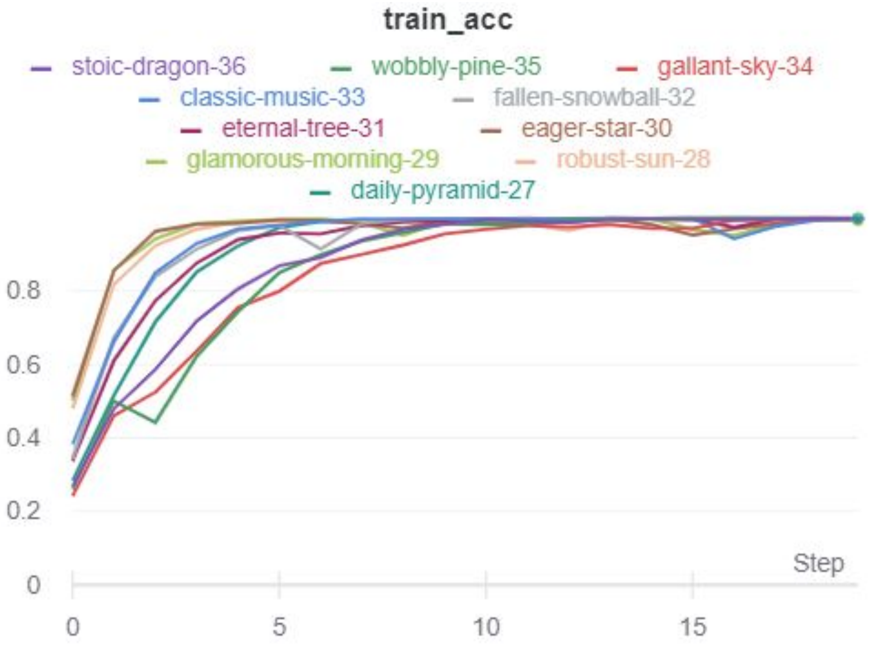
- Single GRU unit
- Initial hidden dimension = 50
- Fed to linear layer to produce 5 outputs


class label
Output
h_t
RNN(GRU)
h_{t-1}
x_t
back propagation
· calculation of gradient & loss
· update parameters
raw email samples
preprocessing
labelled .txt files
data augmentation
more labelled .txt files
Vectorization
word embeddings
input to RNN(GRU)

Hyperparameter Search: Loss and Accuracy Curves

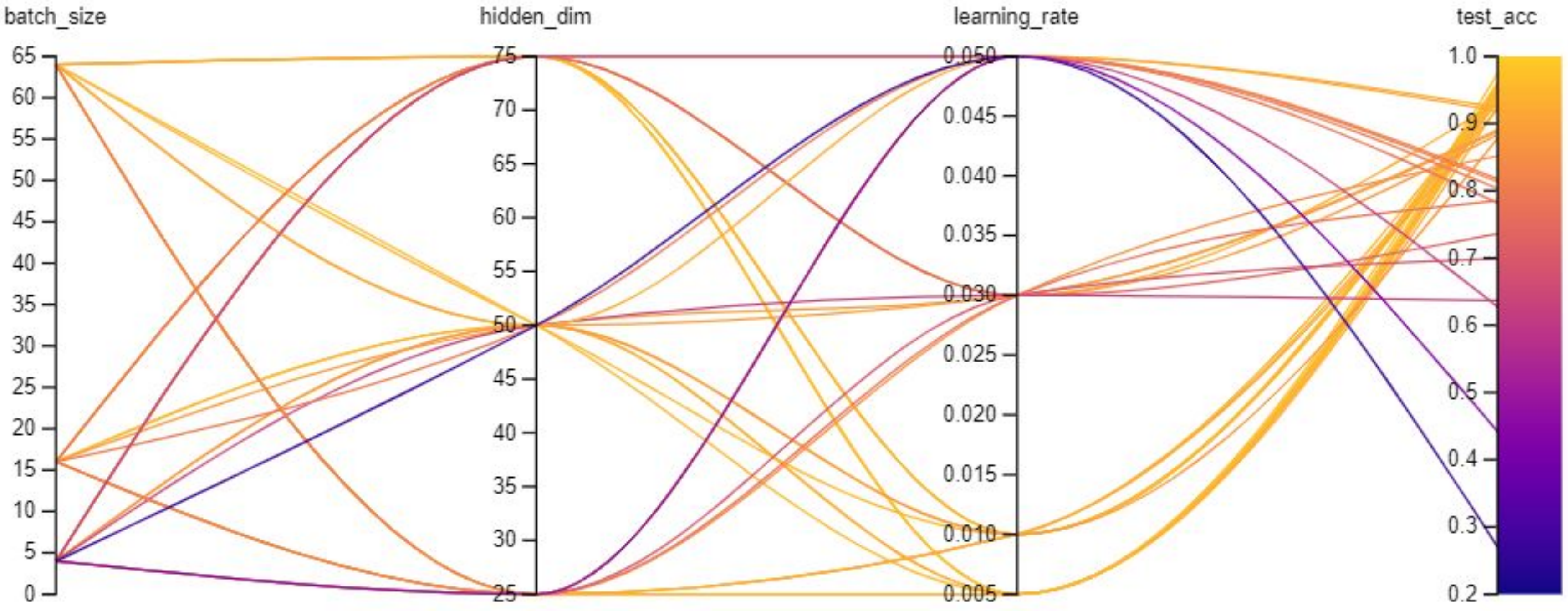


train_loss



train_acc

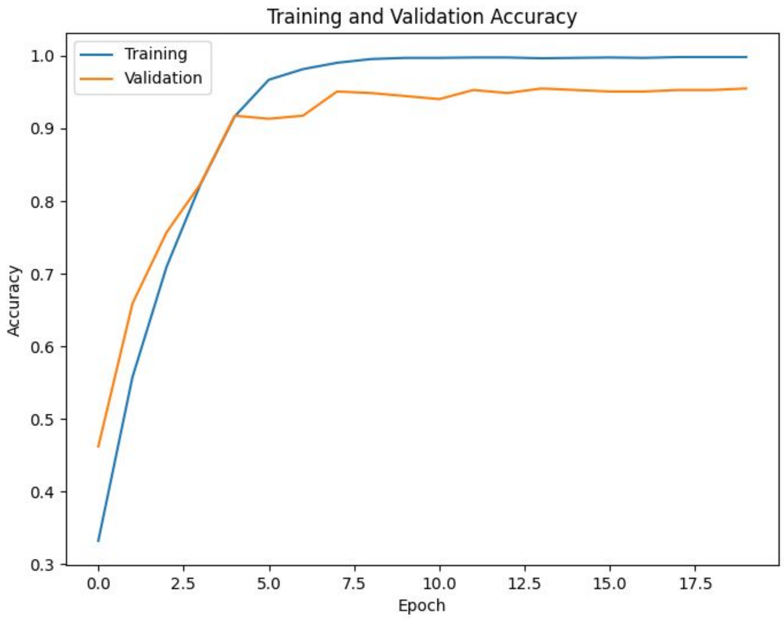
Summary Sweep of Hyperparameters

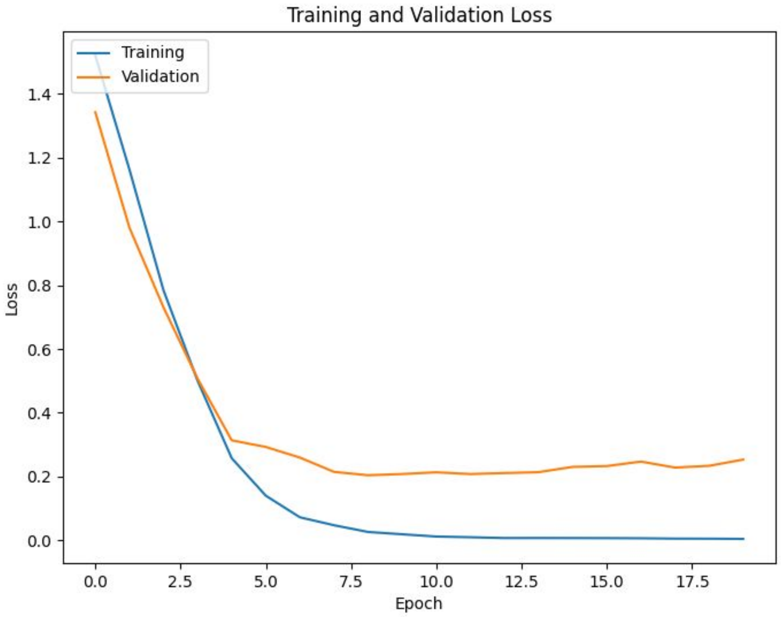


Best Model

Test Accuracy: 96.65%

Hyperparameters: Batch size - 32, Learning rate - 0.005, Hidden dimension - 75


Training and Validation Accuracy


Training and Validation Loss

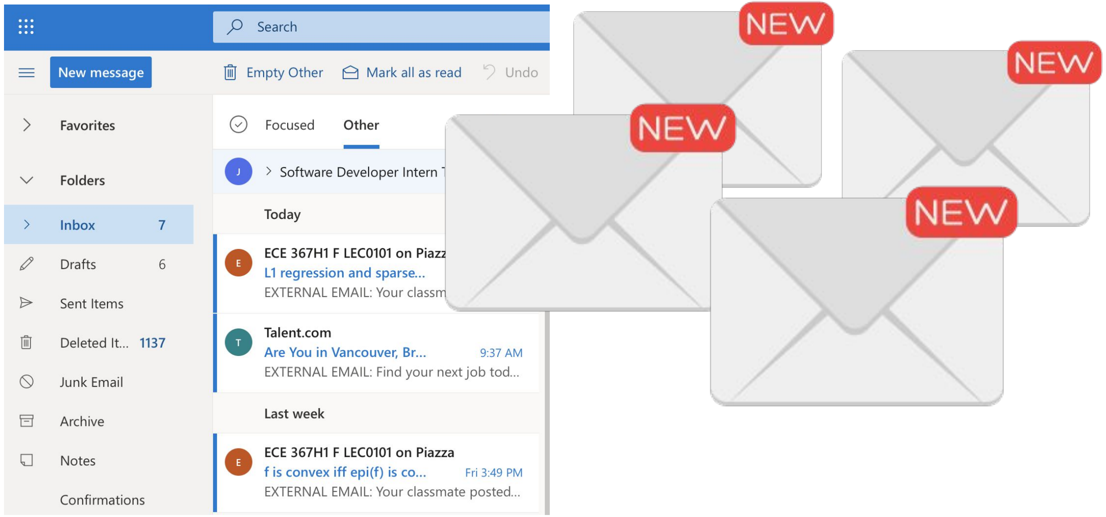
Video Demo

Video Summary

From the video we saw

- Ex 1: Question from Piazza was correctly predicted as Academics
- Ex 2: Ad for job opportunities was predicted as Promotions & Events

These examples were from new emails the model has never seen!



J Software Developer Intern

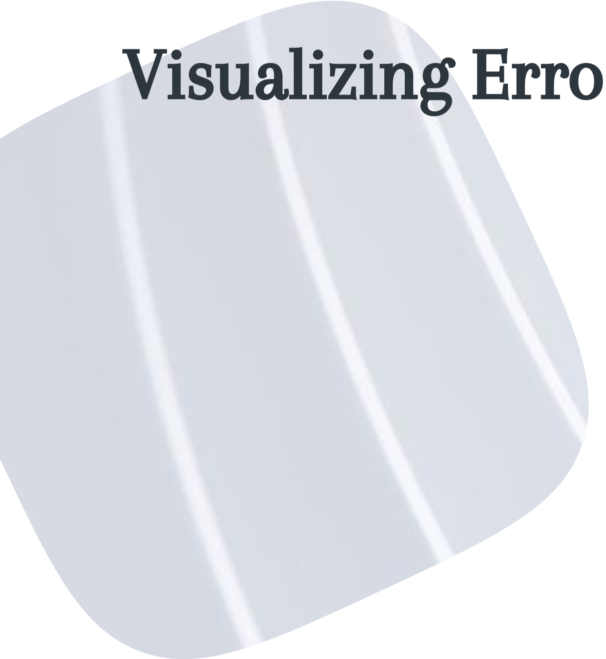
Today

E ECE 367H1 F LEC0101 on Piazz
L1 regression and sparse...
EXTERNAL EMAIL: Your classm

T Talent.com
Are You in Vancouver, Br... 9:37 AM
EXTERNAL EMAIL: Find your next job tod...

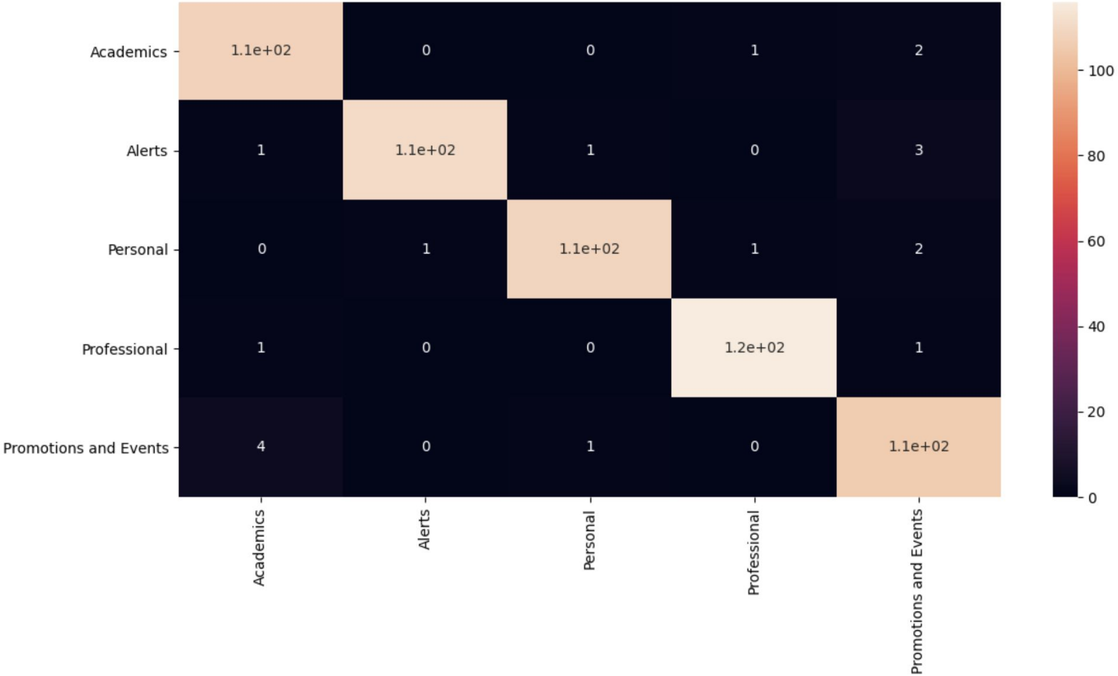
Last week

E ECE 367H1 F LEC0101 on Piazza
f is convex iff epi(f) is co... Fri 3:49 PM
EXTERNAL EMAIL: Your classmate posted...



Visualizing Errors

- The **poorest performing** class was Promotions & Events
  - Actual P&E got confused for Academics
  - Most samples falsely labeled as P&E

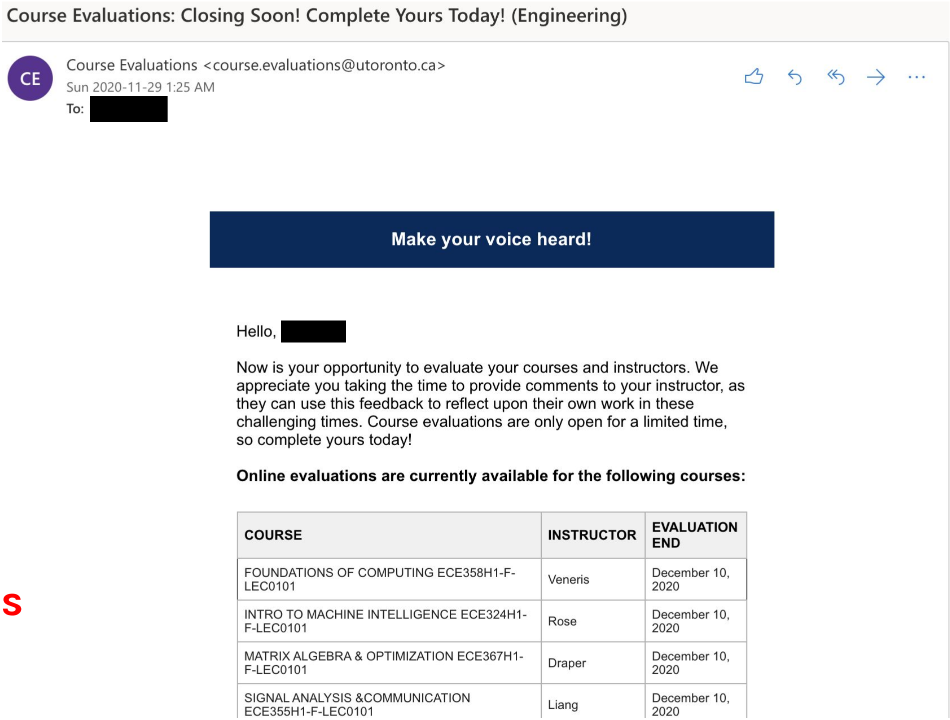


	Academics	Alerts	Personal	Professional	Promotions and Events
Academics	1.1e+02	0	0	1	2
Alerts	1	1.1e+02	1	0	3
Personal	0	1	1.1e+02	1	2
Professional	1	0	0	1.2e+02	1
Promotions and Events	4	0	1	0	1.1e+02

Examples of Incorrect Predictions 1

Prediction: Promotions & Events

Ground Truth: Academics



Course Evaluations: Closing Soon! Complete Yours Today! (Engineering)

Course Evaluations <course.evaluations@utoronto.ca>
Sun 2020-11-29 1:25 AM
To: [redacted]

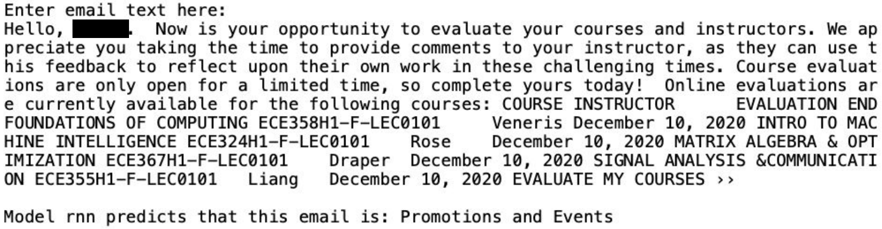
Make your voice heard!

Hello, [redacted]

Now is your opportunity to evaluate your courses and instructors. We appreciate you taking the time to provide comments to your instructor, as they can use this feedback to reflect upon their own work in these challenging times. Course evaluations are only open for a limited time, so complete yours today!

Online evaluations are currently available for the following courses:

COURSE	INSTRUCTOR	EVALUATION END
FOUNDATIONS OF COMPUTING ECE358H1-F-LEC0101	Veneris	December 10, 2020
INTRO TO MACHINE INTELLIGENCE ECE324H1-F-LEC0101	Rose	December 10, 2020
MATRIX ALGEBRA & OPTIMIZATION ECE367H1-F-LEC0101	Draper	December 10, 2020
SIGNAL ANALYSIS &COMMUNICATION ECE355H1-F-LEC0101	Liang	December 10, 2020



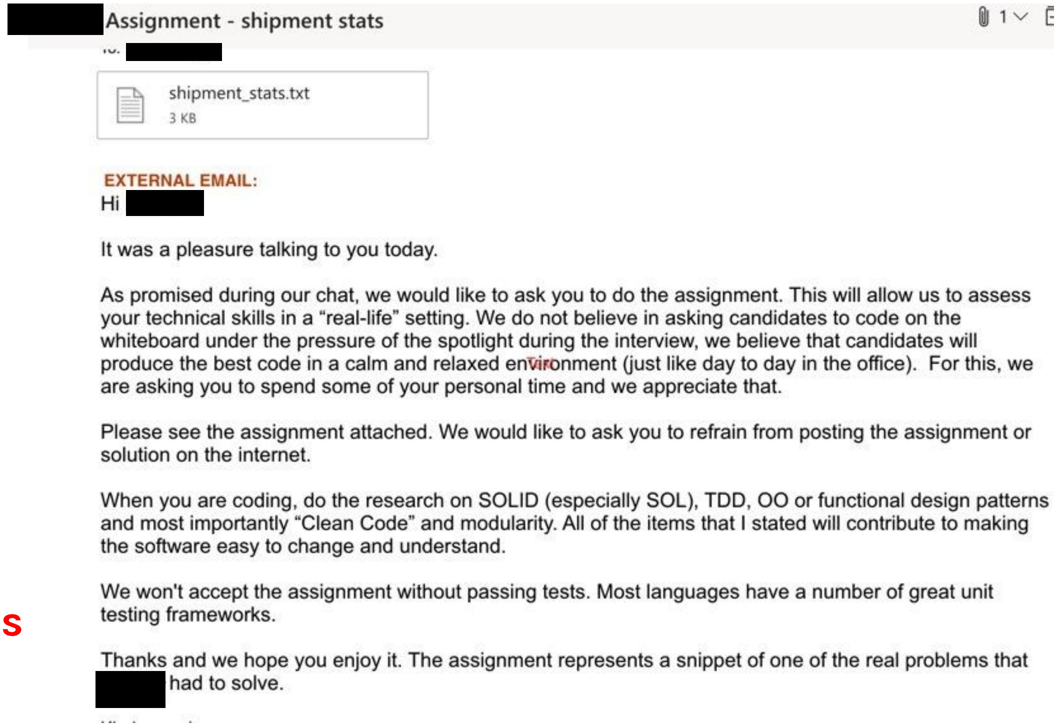
```
Enter email text here:
Hello, [redacted].  Now is your opportunity to evaluate your courses and instructors. We ap
preciate you taking the time to provide comments to your instructor, as they can use t
his feedback to reflect upon their own work in these challenging times. Course evaluat
ions are only open for a limited time, so complete yours today!  Online evaluations ar
e currently available for the following courses: COURSE INSTRUCTOR     EVALUATION END
FOUNDATIONS OF COMPUTING ECE358H1-F-LEC0101     Veneris December 10, 2020 INTRO TO MAC
HINE INTELLIGENCE ECE324H1-F-LEC0101    Rose    December 10, 2020 MATRIX ALGEBRA & OPT
IMIZATION ECE367H1-F-LEC0101    Draper  December 10, 2020 SIGNAL ANALYSIS &COMMUNICATI
ON ECE355H1-F-LEC0101   Liang   December 10, 2020 EVALUATE MY COURSES ››

Model rnn predicts that this email is: Promotions and Events
```

Examples of Incorrect Predictions 2

Prediction: Promotions & Events

Ground Truth: Professional



[redacted] Assignment - shipment stats

To: [redacted]

shipment_stats.txt
3 KB

EXTERNAL EMAIL:

Hi [redacted]

It was a pleasure talking to you today.

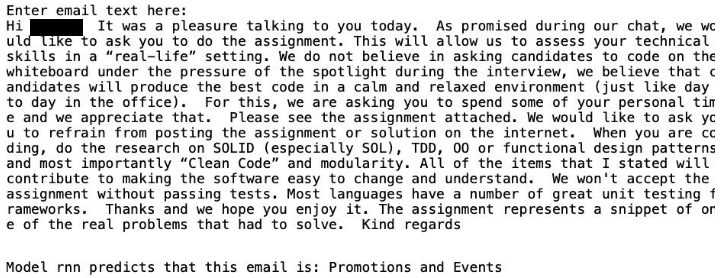
As promised during our chat, we would like to ask you to do the assignment. This will allow us to assess your technical skills in a “real-life” setting. We do not believe in asking candidates to code on the whiteboard under the pressure of the spotlight during the interview, we believe that candidates will produce the best code in a calm and relaxed environment (just like day to day in the office). For this, we are asking you to spend some of your personal time and we appreciate that.

Please see the assignment attached. We would like to ask you to refrain from posting the assignment or solution on the internet.

When you are coding, do the research on SOLID (especially SOL), TDD, OO or functional design patterns and most importantly “Clean Code” and modularity. All of the items that I stated will contribute to making the software easy to change and understand.

We won't accept the assignment without passing tests. Most languages have a number of great unit testing frameworks.

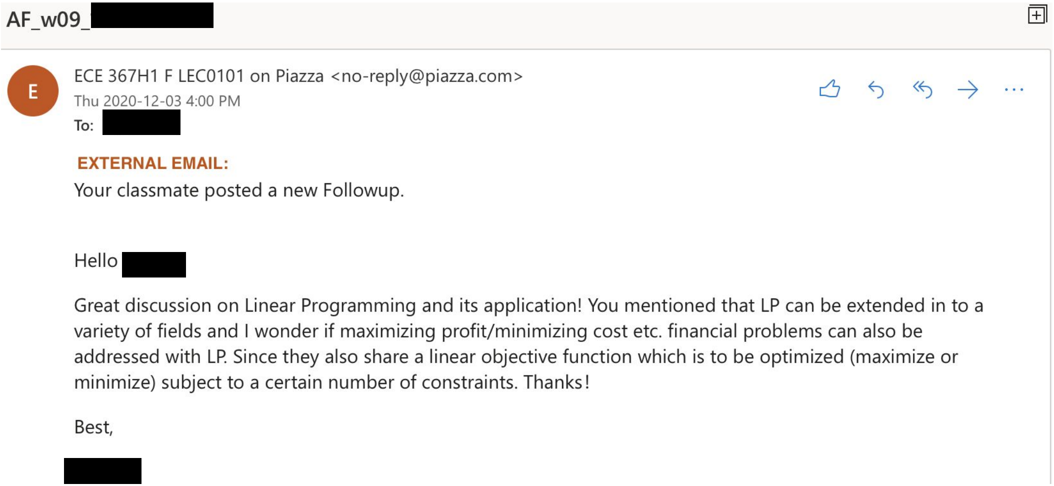
Thanks and we hope you enjoy it. The assignment represents a snippet of one of the real problems that [redacted] had to solve.



```
Enter email text here:
Hi [redacted]  It was a pleasure talking to you today.  As promised during our chat, we wo
uld like to ask you to do the assignment. This will allow us to assess your technical
skills in a “real-life” setting. We do not believe in asking candidates to code on the
whiteboard under the pressure of the spotlight during the interview, we believe that c
andidates will produce the best code in a calm and relaxed environment (just like day
to day in the office).  For this, we are asking you to spend some of your personal tim
e and we appreciate that.  Please see the assignment attached. We would like to ask yo
u to refrain from posting the assignment or solution on the internet.  When you are co
ding, do the research on SOLID (especially SOL), TDD, OO or functional design patterns
and most importantly “Clean Code” and modularity. All of the items that I stated will
contribute to making the software easy to change and understand.  We won't accept the
assignment without passing tests. Most languages have a number of great unit testing f
rameworks.  Thanks and we hope you enjoy it. The assignment represents a snippet of on
e of the real problems that had to solve.  Kind regards

Model rnn predicts that this email is: Promotions and Events
```

Ignore this



AF_w09_[redacted]

ECE 367H1 F LEC0101 on Piazza <no-reply@piazza.com>
Thu 2020-12-03 4:00 PM
To: [redacted]

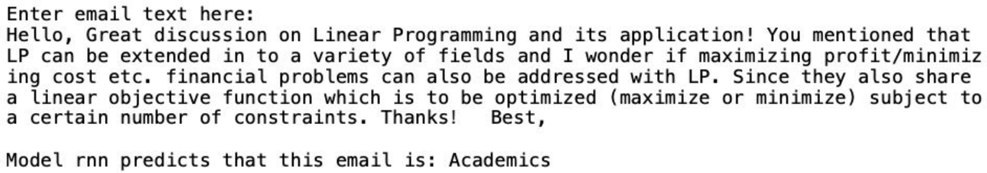
EXTERNAL EMAIL:
Your classmate posted a new Followup.

Hello [redacted]

Great discussion on Linear Programming and its application! You mentioned that LP can be extended in to a variety of fields and I wonder if maximizing profit/minimizing cost etc. financial problems can also be addressed with LP. Since they also share a linear objective function which is to be optimized (maximize or minimize) subject to a certain number of constraints. Thanks!

Best,

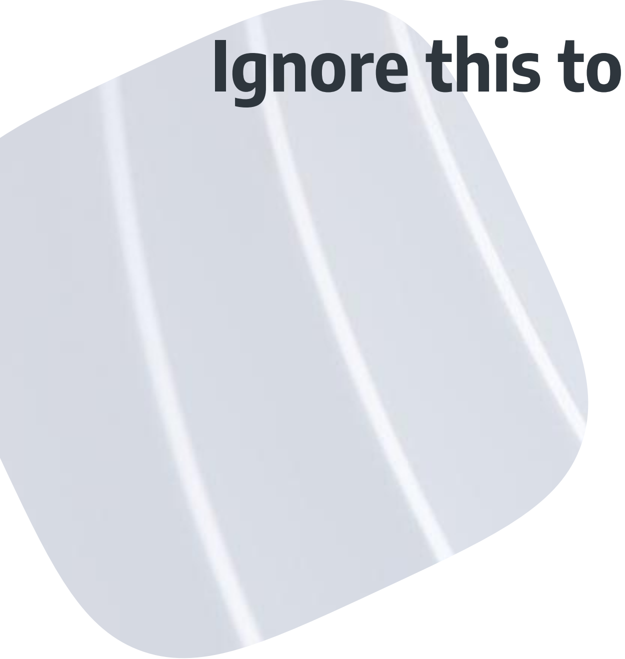
[redacted]

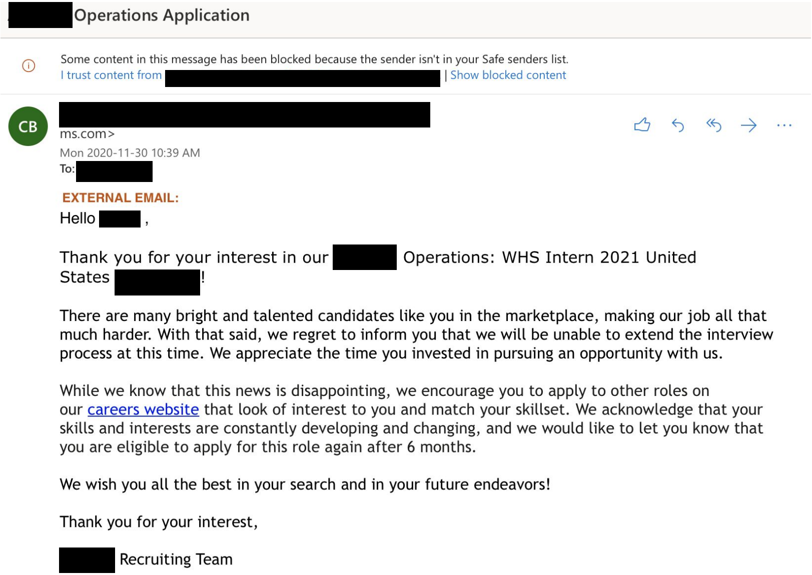


```
Enter email text here:
Hello, Great discussion on Linear Programming and its application! You mentioned that
LP can be extended in to a variety of fields and I wonder if maximizing profit/minimiz
ing cost etc. financial problems can also be addressed with LP. Since they also share
a linear objective function which is to be optimized (maximize or minimize) subject to
a certain number of constraints. Thanks!   Best,

Model rnn predicts that this email is: Academics
```

Ignore this too





[redacted] Operations Application

Some content in this message has been blocked because the sender isn't in your Safe senders list.
I trust content from [redacted] | Show blocked content

[redacted]
ms.com>
Mon 2020-11-30 10:39 AM
To: [redacted]

EXTERNAL EMAIL:

Hello [redacted] ,

Thank you for your interest in our [redacted] Operations: WHS Intern 2021 United States [redacted]!

There are many bright and talented candidates like you in the marketplace, making our job all that much harder. With that said, we regret to inform you that we will be unable to extend the interview process at this time. We appreciate the time you invested in pursuing an opportunity with us.

While we know that this news is disappointing, we encourage you to apply to other roles on our careers website that look of interest to you and match your skillset. We acknowledge that your skills and interests are constantly developing and changing, and we would like to let you know that you are eligible to apply for this role again after 6 months.

We wish you all the best in your search and in your future endeavors!

Thank you for your interest,

[redacted] Recruiting Team

```
Enter email text here:
Thank you for your interest in our [redacted] Operations: WHS Intern 2021 United States (1
[redacted])!   There are many bright and talented candidates like you in the marketplace,
making our job all that much harder. With that said, we regret to inform you that we w
ill be unable to extend the interview process at this time. We appreciate the time you
invested in pursuing an opportunity with us.    While we know that this news is disapp
ointing, we encourage you to apply to other roles on our careers website that look of
interest to you and match your skillset. We acknowledge that your skills and interests
are constantly developing and changing, and we would like to let you know that you are
eligible to apply for this role again after 6 months.   We wish you all the best in yo
ur search and in your future endeavors!   Thank you for your interest,   [redacted] Recrui
ting Team

Model rnn predicts that this email is: Professional
```

Discussion

We have high test accuracy... but:

- Data source (¾ Engineering + ¼ Artsci = 100% UofT)
- Only tested on emails in our accounts
- **May not generalize!**

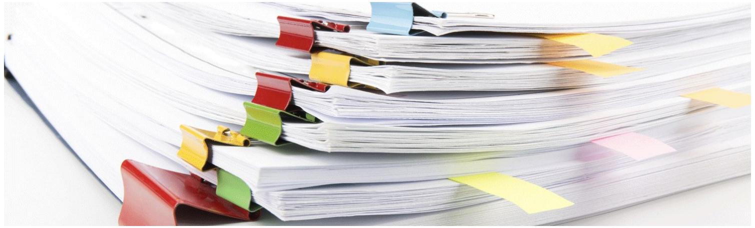
Keyword frequency implies category:

Recall the most common words in each category

- “Course” → Academics
- “Order” → Alerts
- “Application” → Professional
- “Students” → Promotion & Events
- “Practice” → Personal

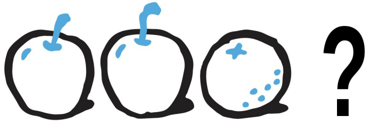
What we learned

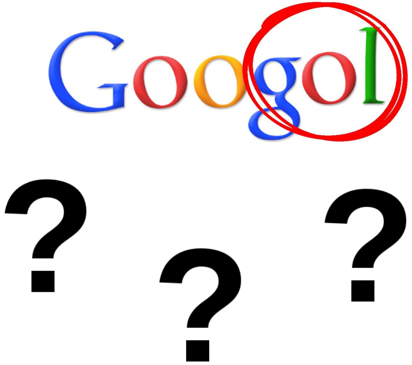
- Preparing raw data takes **time**!
  - Need enough
  - Manual cleaning
- Need to be **VERY clear** on the scope of the labels



Things to Consider

- **Generalization** for all users?
- User choice for **labels**?
- Managing **typos**?
- Incorporation of **decision heuristics** into NN design?





Any Questions?