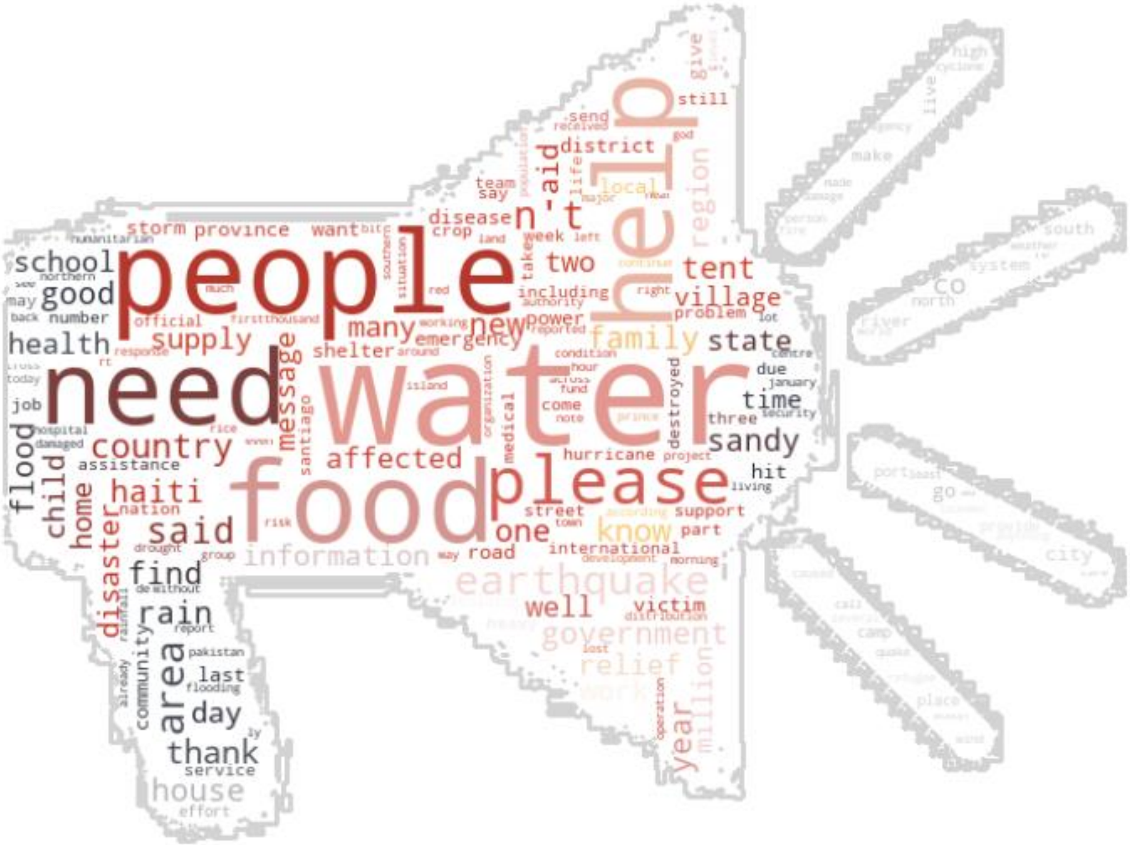
Labeling Disaster-Related Messages
Using Natural Language Processing

Business Understanding

- **Objective:**
  - Our goal is to **create a model** that can interpret and label a message using **Natural Language Processing**.
  - Messages are either:
    - ***direct*** *(messages sent from person-to-person)*
    - ***news*** *(headlines or clippings)*
    - ***social*** *(social media)*
  - In order to simplify the given dataset, I will be working only with **a single label – "*aid_related*".**
- **Success criteria:**
  - How well the model finds all the *aid_related* messages *(Recall)*.
  - How accurate the model is when it predicts an *aid_related* message *(Precision)*.
  - How accurate the model is overall *(Accuracy)*.

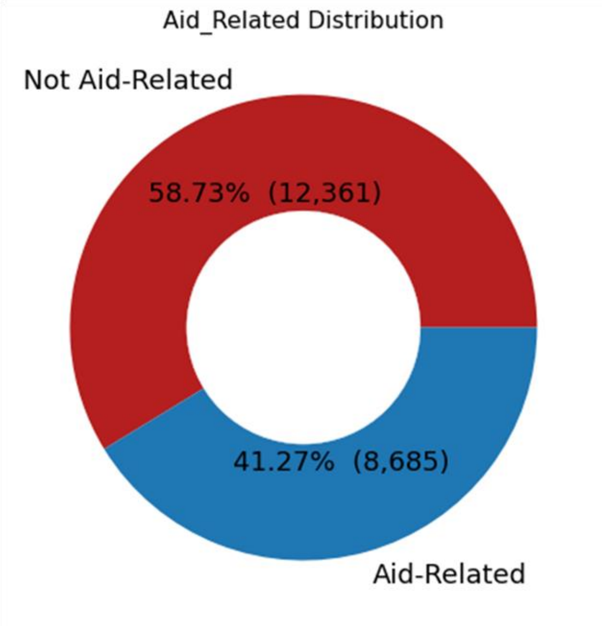


Data Understanding

- The relevant columns of the dataset are ***message*** and ***aid_related***.

  - ***message*** *(our predictive data) is a string of text, e.g.:*
    - *"Weather update – a cold front from Cuba that could pass over Haiti"*
    - *"There's nothing to eat and water, we starving and thirsty."*

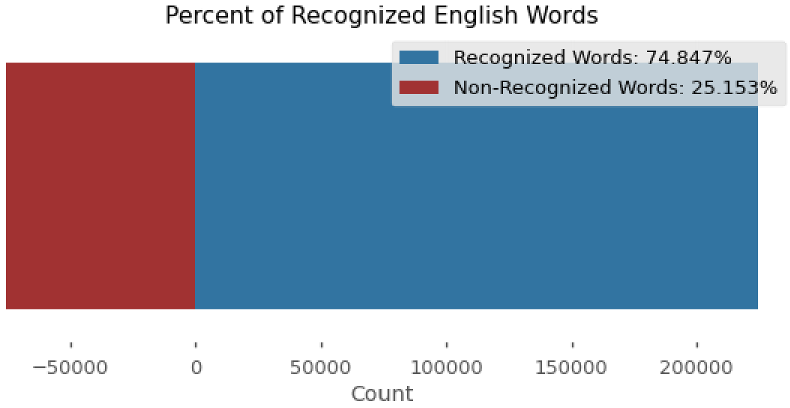
  - ***aid_related*** *(our target) is a binary column, i.e.:*
    - *Is the message aid related? 1=yes, 0=no.*



Data Understanding

Non-Aid-Related

Aid-Related



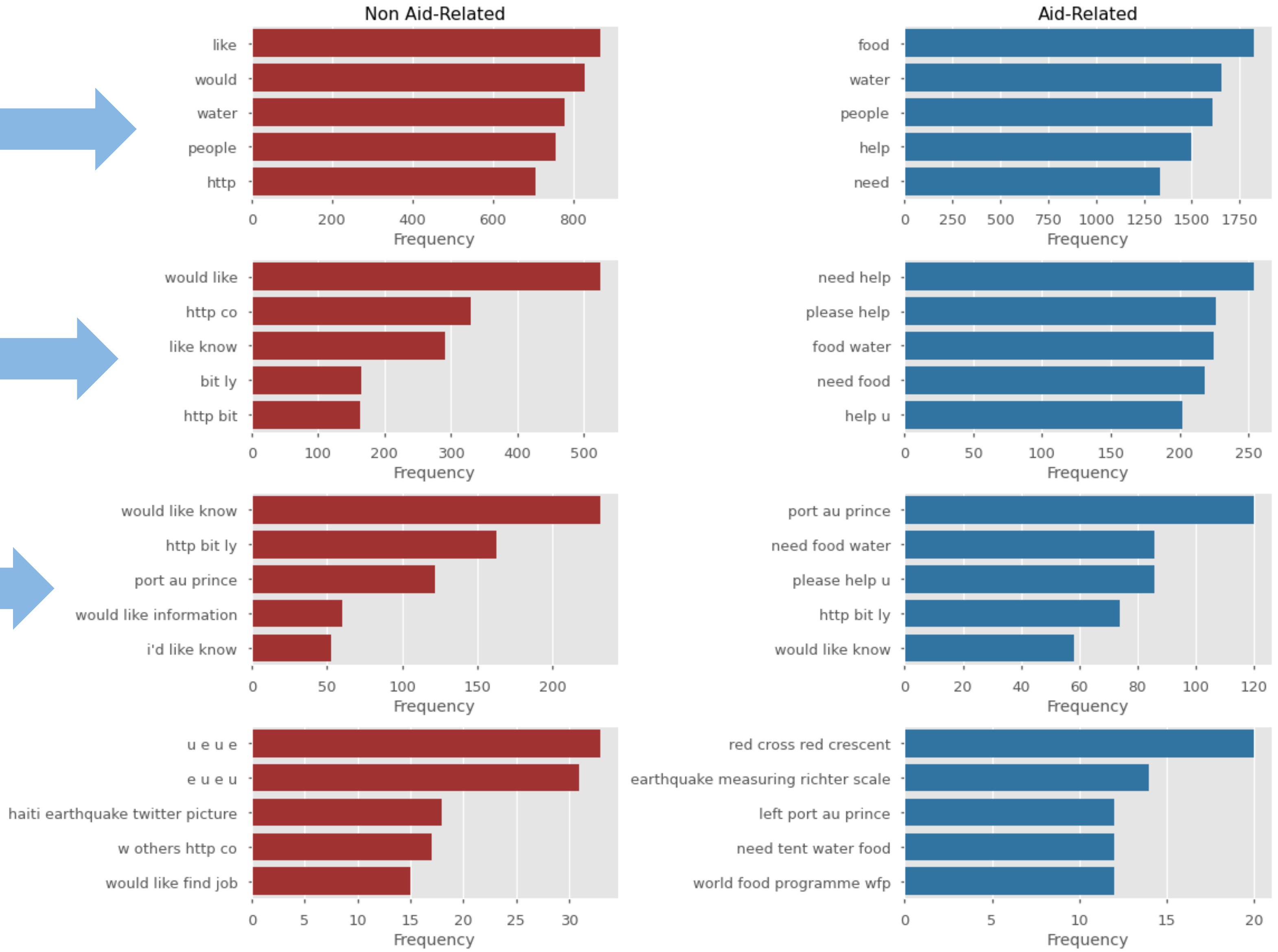
Data Understanding

One-word sequences

Two-word sequences

Three-word sequences

Four-word sequences



Data Preparation

- Text Preparation
  - Cleaning abnormalities (unusual html characters),
  - Removing *stop words ("the", "is", "and") & punctuation,*
  - Lemmatizing *(feet -> foot; running -> run)*
- Vectorizing
  - Premade Vectorizer - GloVe model *(Global Vectors for Word Representation)*
    - https://nlp.stanford.edu/projects/glove/
  - Homemade Vectorizer - Gensim Word2Vec model

Data Preparation

Homemade Word Vectors Trained on Training Data

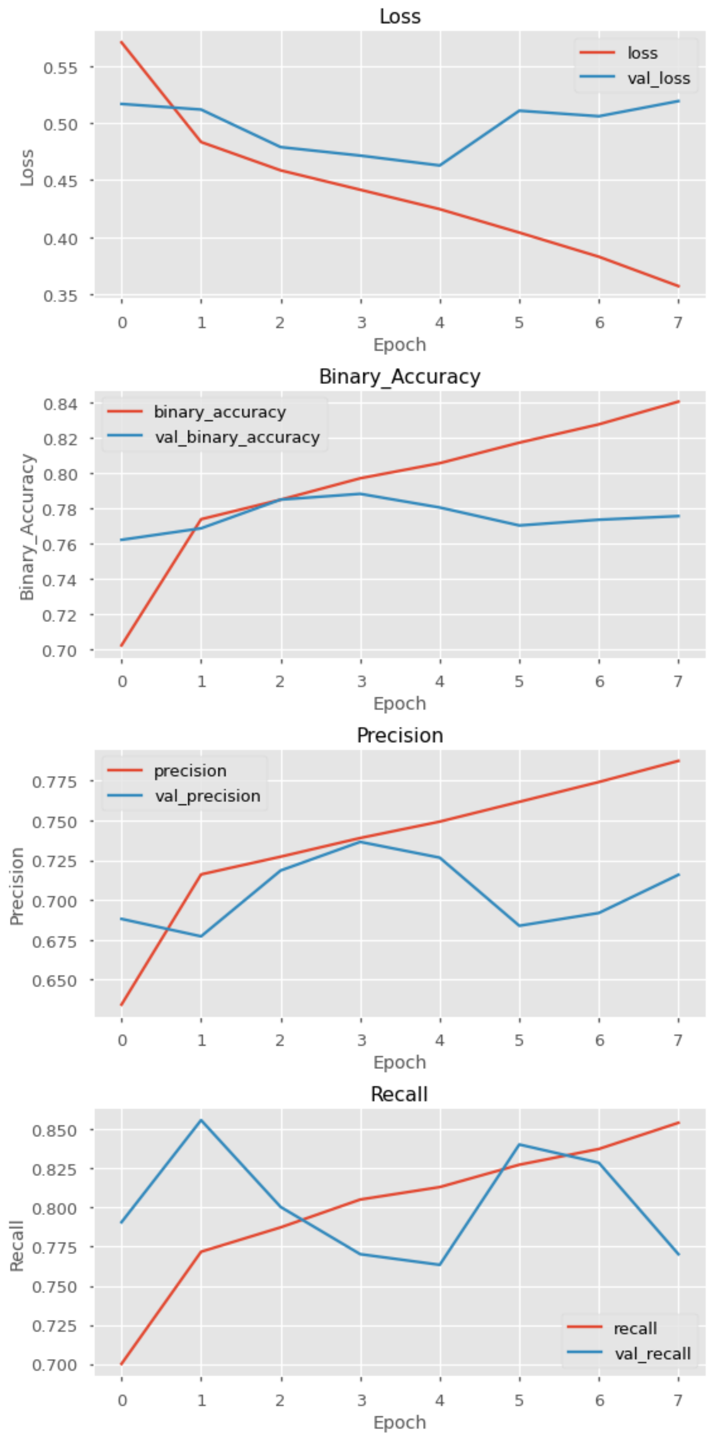
```
********************************************************************************
*                                   VILLAGE                                    *
********************************************************************************
Most Similar Words:
1.      district
2.      area
3.      tahsil
4.      hill
5.      city
6.      basti
7.      wala
8.      distt
9.      embankment
10.     house

********************************************************************************
*                                    WATER                                     *
********************************************************************************
Most Similar Words:
1.      drinking
2.      cloths
3.      toilet
4.      wells
5.      shelter
6.      toilets
7.      latrines
8.      contaminated
9.      rainwater
10.     tablets

********************************************************************************
*                                    PEOPLE                                    *
********************************************************************************
Most Similar Words:
1.      families
2.      survivors
3.      those
4.      refugees
5.      children
6.      residents
7.      villagers
8.      victims
9.      persons
10.     students
```

Modeling
Featured Model:
RNN - GloVe

[Try the WebApp on StreamLit here!]

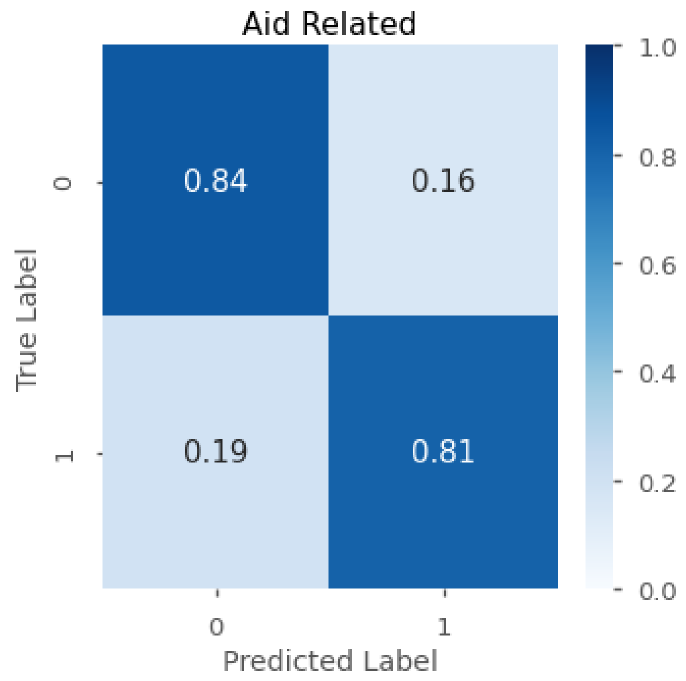
Modeling
Featured Model:
RNN - GloVe

	F1	Accuracy	Recall	Precision
RNN_glove	0.807064	0.826862	0.806708	0.80742
multilayer_model_NN_glove	0.77246	0.785261	0.812004	0.736589
SVC_glove	0.769165	0.787639	0.788173	0.751051
simple_model_NN_glove	0.764513	0.771791	0.825243	0.71211
multilayer_model_NN_w2v	0.762413	0.770602	0.819947	0.712423
simple_model_NN_w2v	0.754296	0.773376	0.774934	0.734728
RNN_w2v	0.747631	0.767829	0.766108	0.730025
LOGREG_glove	0.733208	0.718304	0.862312	0.637728
RFC_glove	0.73288	0.76664	0.713151	0.753731
LOGREG_w2v	0.729776	0.751189	0.748455	0.712007
RFC_w2v	0.722543	0.771791	0.661959	0.795334
NB_w2v	0.707668	0.709984	0.781995	0.646244
SVC_w2v	0.707053	0.731775	0.721094	0.693548
NB_glove	0.689233	0.698098	0.745808	0.640637

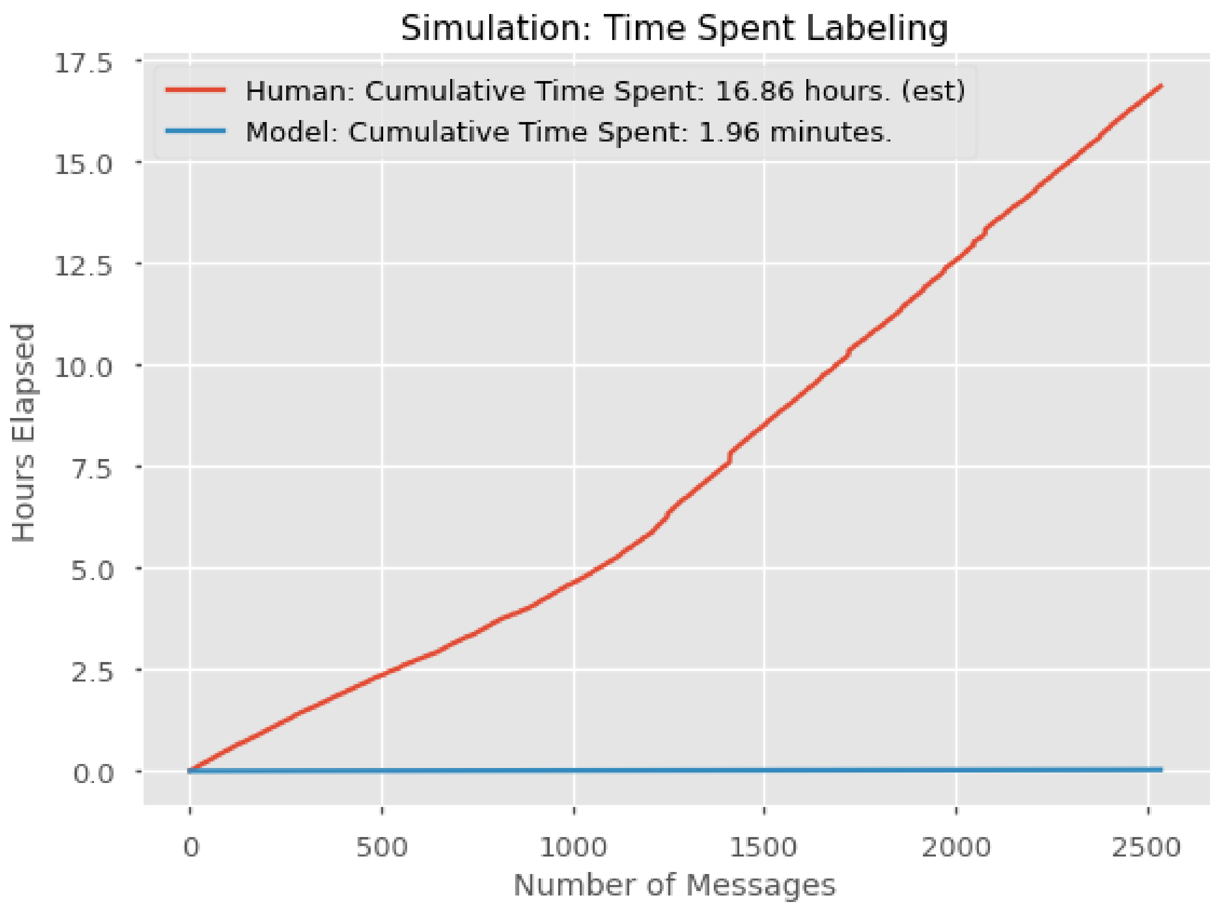


Evaluation

- Overall, RNN - GloVe (the RNN accompanied by the GloVe embeddings) performed clearly best overall.
  - On the test set:
    - 80.67% of `aid-related` messages were found.
    - 80.74% of `aid-related` predictions were correct.
    - 82.69% overall accuracy.



Evaluation



- This model, if used in the field, would save hours of man-power.
  - With approximately 2500 messages, the model would save approximately 15 hours of time that would have been spent with a human-labeler.

Model Recommendations

- If the priority is **overall accuracy, confidence in positive predictions**, and **balance** (F1):
  - The **Recurrent Neural Network** with GloVe embeddings scored significantly best – 81% of aid-related predictions were correct, and 83% of its overall predictions were correct.
- If the priority is to **find the most aid-related messages** (at the expense of mislabeling many messages as aid-related):
  - **Logistic Regression** with the homemade Vectorizer scored the best – finding 86% of all aid-related messages.

	F1	Accuracy	Recall	Precision
RNN_glove	0.807064	0.826862	0.806708	0.80742
multilayer_model_NN_glove	0.77246	0.785261	0.812004	0.736589
SVC_glove	0.769165	0.787639	0.788173	0.751051
simple_model_NN_glove	0.764513	0.771791	0.825243	0.71211
multilayer_model_NN_w2v	0.762413	0.770602	0.819947	0.712423
simple_model_NN_w2v	0.754296	0.773376	0.774934	0.734728
RNN_w2v	0.747631	0.767829	0.766108	0.730025
LOGREG_glove	0.733208	0.718304	0.862312	0.637728
RFC_glove	0.73288	0.76664	0.713151	0.753731
LOGREG_w2v	0.729776	0.751189	0.748455	0.712007
RFC_w2v	0.722543	0.771791	0.661959	0.795334
NB_w2v	0.707668	0.709984	0.781995	0.646244
SVC_w2v	0.707053	0.731775	0.721094	0.693548
NB_glove	0.689233	0.698098	0.745808	0.640637

Future Work

- Include the other 36 target labels to further classify the messages.
  - (Multilabel Classification)
- Add other languages to the model rather than just English translation.
- Continue to explore the complexity of the neural network architecture and create a larger network.

Thank You!

- Data
  - **Appen Datasets**
    - https://appen.com/datasets/combined-disaster-response-data/
- Flatiron School
  - James Irving – DS Instructor