

Business Understanding

Objective:

Our goal is to **create a model** that can interpret and label a message using **Natural Language Processing**. There are **37 important labels** that a message can have (for example if the message is *requesting medical help* or offering aid). In order to simplify the given dataset, I will be working only with **a single label – "aid_related".**

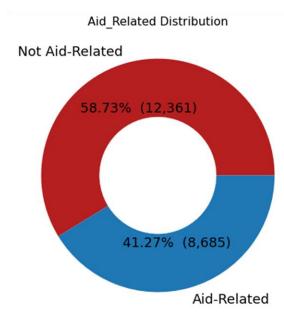
■Success criteria:

- ► 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).



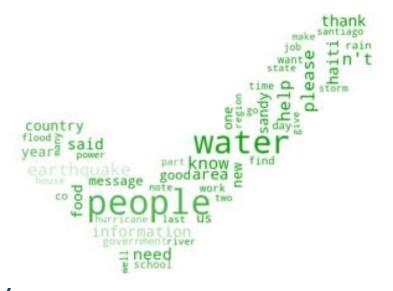
Data Understanding

- 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

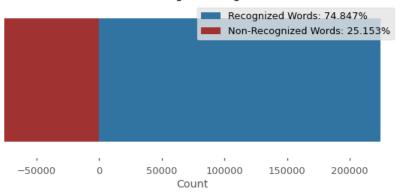
Non-Aid-Related



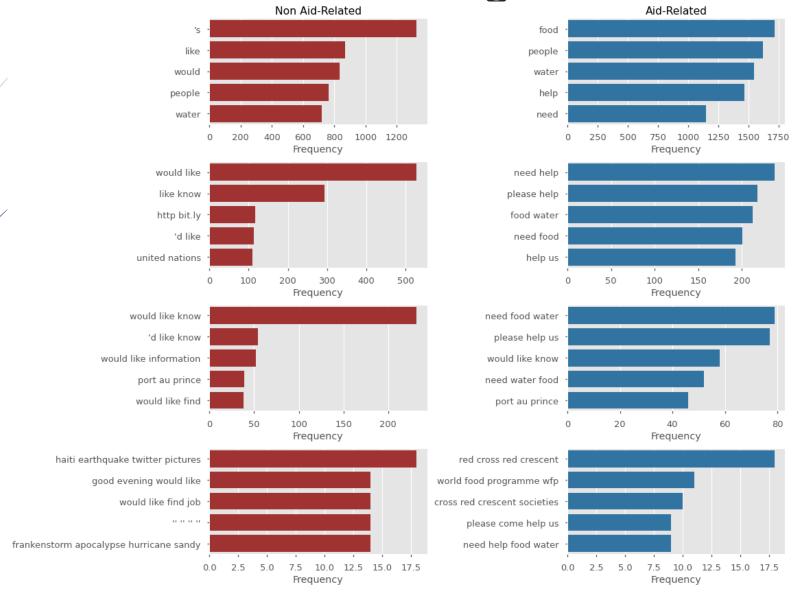
Aid-Related







Data Understanding



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

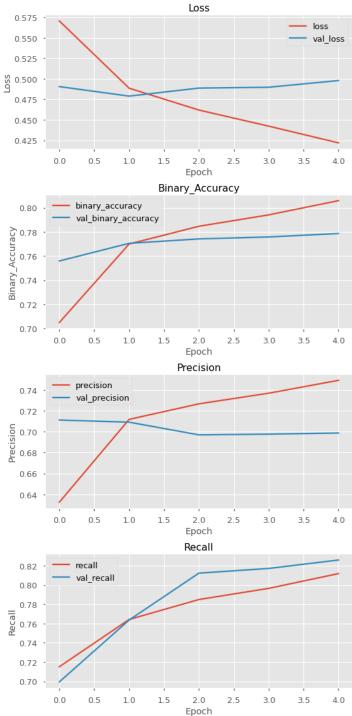
*		VILLAGE	*
****	***********	**********	*******
	Similar Words:		
1.	district		
2.	sujawal		
3.	area		
4.	region		
5.	kilometer		
6.	camp		
7. 8.	county mountain		
9.	hamlet		
10.	pir		
10.	pii		
****	********	**********	*******
*		WATER	*
****	********	*********	************
	Similar Words:		
1.	drinking		
2.	clean		
3.	potable		
4.	polluted		
5.	toilet		
6.	tarp		
7.	contaminated		
8.	chlorine latrine		
9. 10.	food		
10.	1000		
****	*******	*********	*******
*		PEOPLE	*
		**********	************
	Similar Words:		
1.	person		
2.	family		
3.	survivor		
4.	others		
5.	everyone		
6.	someone		
7. 8.	hungry resident		
8. 9.	child		
9. 10.	refuge		
10.	reruge		



Modeling Featured Model:

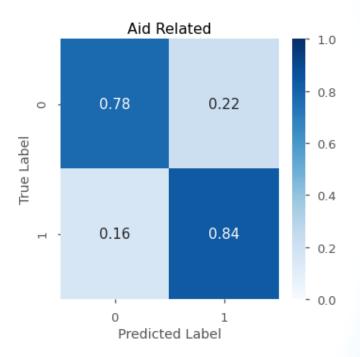
RNN -	GloVe
-------	-------

	F1	Accuracy	Recall	Precision	uracy
RNN_glove	0.794498	0.805523	0.838171	0.755151	Binary_Accuracy
SVC_glove	0.770085	0.787771	0.792436	0.748961	Bin
multilayer_model_NN_w2v	0.768911	0.780671	0.813544	0.72892	
RNN_w2v	0.764531	0.781065	0.792436	0.738525	
$multilayer_model_NN_glove$	0.756849	0.775937	0.777485	0.737281	
SVC_w2v	0.756329	0.757002	0.840809	0.687275	Precision
simple_model_NN_glove	0.75429	0.768442	0.792436	0.719649	Pre
simple_model_NN_w2v	0.752707	0.774753	0.764292	0.741468	
LOGREG_glove	0.732658	0.718738	0.859279	0.638562	
RFC_glove	0.729583	0.764892	0.707124	0.753515	
NB_w2v	0.7222	0.728994	0.7854	0.668413	_
RFC_w2v	0.712071	0.758185	0.666667	0.764113	Recall
LOGREG_w2v	0.710856	0.662722	0.924362	0.577473	
NB_glove	0.687399	0.694675	0.748461	0.635549	



Evaluation

- Overall, RNN GloVe (the RNN accompanied by the GloVe embeddings) performed clearly best overall.
 - On the test set:
 - 83.82% of aid-related messages were found.
 - ▶ 75.52% of aid-related predictions were correct.
 - 80.55% overall accuracy.



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