Using an LLM to better train text classification models

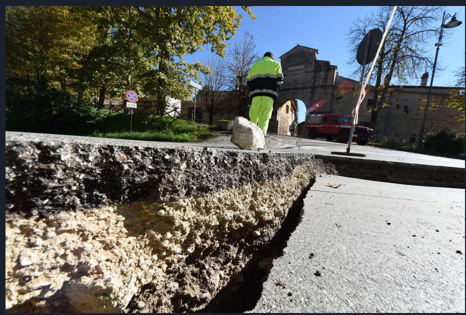
Michael Szczepaniak, April 2024

Scenario: FEMA is modernizing their central monitoring facility...

- Insight Solutions (IS, fictitious company, govt. contractor) receives a request for proposal (RFP)
  - Build system to monitor X (formerly Twitter) tweets for those related to disasters
  - System tasked with assigning probabilities to tweets being about a real disaster
  - Personnel notified if probability > 50%
- Expectations
  - Too much data for a human to monitor, so some kind of ML model is expected, > 80% accurate
  - Would like to see generative AI (GAI) utilized if and where it makes sense

Scenario: FEMA is modernizing their central monitoring facility (cont.)

- IS anticipated RFP three months earlier and initiated a project to explore gen. AI feasibility (aka feasibility project)
  - Data set similar to what FEMA will be monitoring was identified (kaggle Disaster Tweets, see References)
- Management needs from Data Science (DS)
  - Is > 80% accuracy feasible? If so, propose how to demo
  - Proposal for GAI utilization
  - Describe the overall system in just enough detail to properly quote the solution and explain it to FEMA
  - These DS deliverables are the focus of this presentation



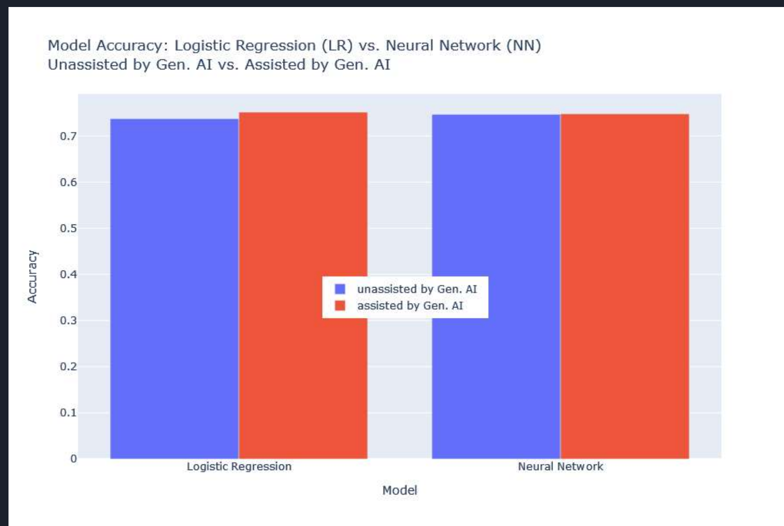
Feasibility of accuracy requirement

- Feasibility project goal focused on GAI requirement...
  - Accuracy was not the goal, but was used to evaluate generative AI utility
  - Two types of models: logistic regression and a single hidden-layer neural network
- Baseline model - logistic regression (LR)
  - Two LR models built and compared:

    model 1: unassisted by GAI
    model 2: assisted by GAI *(how GAI assisted described later)*
  - model 1: 73.7% accuracy on unlabeled kaggle test set
  - model 2: 75.1% accuracy on unlabeled kaggle test set

Feasibility of accuracy requirement (cont.)

- Validation model - neural network (NN)
  - simplest architecture: single hidden-layer
  - hidden units and activation function determined by CV (100, ReLU)
- NN model expected to outperform LR model, but did not
  - Two NN models built and compared
  - model 1: unassisted by GAI
    model 2: assisted by GAI *(GAI assist described later)*
  - model 1: 74.7% accuracy on unlabeled kaggle test set
  - model 2: 74.8% accuracy on unlabeled kaggle test set



Factors inhibiting accuracy

- Mislabeled training data (biggest inhibitor)
  - Degrades performance of both models
  - Roughly 1 in 4 DISASTER class labels should be NOT DISASTER (96 of 400 random samples, see examples below and Appendix A). Very few NOT DISASTER label problems (2 in 400).
- Better validation model (much less of a factor, see Appendix B for details)
- We'll discuss how to address these factors in later slides

id	text	target	notes
4882	Kendall Jenner and Nick Jonas Are Dating and the World Might Quite Literally Explode http://t.co/pfvzVPxQGr	1	two celebs dating
8880	I get to smoke my shit in peace	1	probably not explosive
4142	We happily support mydrought a project bringing awareness to the LA drought. Track your water⑨Û_ https://t.co/2ZvhX41I9v	1	support for a disaster project ≠ disaster
4973	ITS A TIE DYE EXPLOSION ON IG HELP ME. IM DROWNING IN TIE DYE	1	a mess for the textile artist ≠ disaster

What is Generative AI?

Let’s let a generative AI model speak for itself...

prompt: *explain generative AI in 30 words or less*

ChatGPT 3.5 response: Generative AI creates new data resembling original input. It generates content like images, text, or music, often using neural networks to mimic patterns and create novel outputs.

Utilizing Generative AI

- Many, many models available (see References)
- Selection depend on use case
  - open source: free, but need to develop skills to use well, support from online communities
  - commercial: not free, designed to be easier to use, typically better support than open source
- Audio
  - speech generation: Bark, Coque
  - music generation: Harmonai (open source), MusicLM (Google)
- Images
  - DALL-E-2 (closed source)
  - Stable Diffusion (open source)
- Text
  - Gemini (Google, closed source)
  - ChatGPT (Open AI, closed source)

Utilizing Generative AI (what about our task?)

- Our task: classify tweets
- Suitable Generative AI: Large Language Models (LLMs)
  - generate text from a prompt
- How can an LLM help us with our task?
  - Need labeled data to train tweet classifier
  - Using human to label data is expensive
- Could an LLM generate labeled data to train our classifier?

Utilizing Generative AI (+ example)

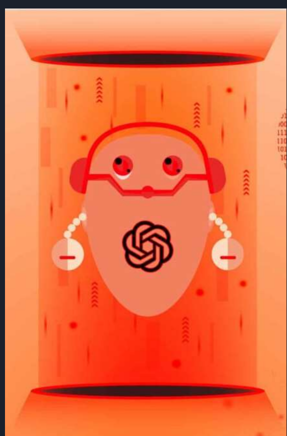
Prompt to create a disaster tweet from an original disaster tweet:

Write me a tweet similar to this one in length and content, under 141 characters, does not contain double quotes but refers to a different disaster and location:

Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all

ChatGPT 3.5 response:
Witnessing the devastation caused by the powerful hurricane in the Caribbean region. Praying for the safety and well-being of all those affected. #hurricane #Caribbean

Utilizing Generative AI (– example)



Prompt to create a NOT disaster tweet from an original disaster tweet:

Write me a tweet similar to this one in length and content, under 141 characters, does not contain double quotes but refers to a different activity, feeling and location:

What's up man?

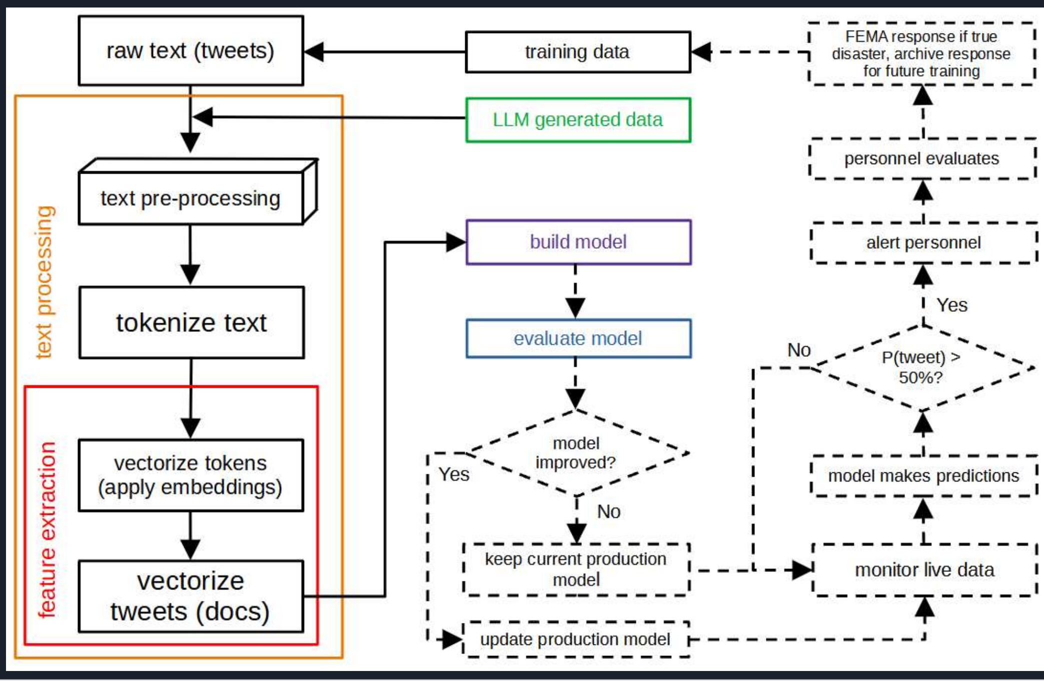
ChatGPT 3.5 response:
Hey there! How are you feeling after that intense workout at the gym? #fitnessgoals #gymlife

So it looks like we've found a good way to put Gen AI to work for us!

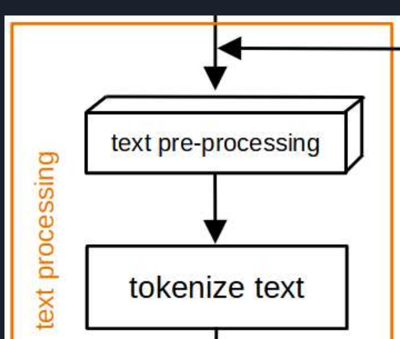
Accuracy path forward...

- Manually or build model to check the labels on all kaggle train.csv samples
  - Start with the 3188 DISASTER (target = 1) tweets
  - Take a different sample of 400 from the 4297 NOT DISASTER (target = 0) tweets
    - If still ~0.5% labeling errors, don't worry about these.
- Regenerate augmented data from corrected samples
- Replace samples generated from mislabeled tweets with newly generated augmented data
- Rerun pre-processing and vectorization (described later)
- Retrain models and evaluate results (NN should be noticeably better)

Feasibility Study vs. FEMA System (overview)



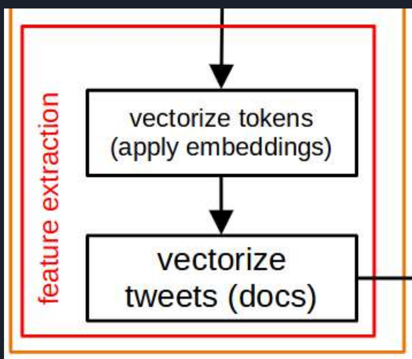
Feasibility Study vs. FEMA System (text processing)



- Before building a model that takes text as inputs…
  - must convert text into numbers (vectors)
  - before this conversion, text must be “cleaned up”
    - analogy: cleaning vegetables before cooking them
- Steps in the clean up
  - normalize URLs and special twitter tokens (# and @)
  - expand contractions: I’m ➔ I am, can’t ➔ can not, etc.
  - remove stop words, singletons, punctuation, OOV words/tokens
  - See Appendix C for detailed step-by-step example
- Tokenize = break text into logical pieces (tokens ≈ words, “chopping the veggies”)

➔ **original tweet**	*I can't bloody wait!! Sony Sets a Date For Stephen King@Ûªs @Û÷The Dark Tower@Ûª #stephenking #thedarktower http://t.co/J9LPdRXCDE @bdisgusting*
➔ **processed tweet**	*not bloody wait sony set date stephen dark <hashtag> <hashtag> <url> <user>*
➔ **tokenized tweet**	not | bloody | wait | sony | set | date | stephen | dark | <hashtag> | <hashtag> | <url> | <user>

Feasibility Study vs. FEMA System (feature extraction)



- Converting tokenized text to numbers (vectors)
  - Each token (word) represented by vector of 50 numbers (can be more or less)
  - Mapping between words and vectors = embedding
  - How this mapping is created is interesting, but not important here
  - What is important: After this conversion, text is in a form that can be used to train our models or to use our models to make predictions
- Example - first 2 words of tweet from the previous slide: not bloody wait
  - not = [0.49427, 0.13234, -0.023199, ... 47 more numbers]
  - bloody = [0.42033, 0.30658, 0.80744, ... 47 more numbers]

First 100 dimensions of our example tweet vector:

not | bloody = [0.49427, 0.13234, -0.023199, … 47 more numbers, 0.42033, 0.30658, 0.80744, … 47 more numbers, ...]

Summary & Conclusions

- ❑ Initial feasibility accuracy on kaggle tweet data ~75%
- ❑ Primary factor preventing higher accuracy: mislabeled training data
  - Estimated approximately 1 in 4 DISASTER class tweets mislabeled
  - Estimated 0.5% of NOT DISASTER class tweets mislabeled
- ❑ High confidence that > 80% requirement can be reached
  - Need to fix training data
  - Models retrained on fixed data should significantly improve performance
  - Demo on randomly selected tweets with new model (redesigned, tuned NN) on fixed train data

Summary & Conclusions (cont.)

- Generative AI in the form of an LLM is a great fit for this RFP.
  - Meets a customer “want” and makes sense in terms of our business
- Using an LLM to augment training data for our models will save us significant time & money.
  - Every hand-labeled training sample can be amplified many times over
- Using an LLM to augment training data will help make our models more accurate
  - This project demonstrated improved accuracy on two types of models

Summary & Conclusions (cont.)

- Experience with generative AI is becoming a competitive advantage
  - Strategic investments can be made to strengthen our expertise while simultaneously enhancing our technology portfolio
- A proposed system design was presented which satisfy FEMA requirements
- Further details available in feasibility project report
  - Last item in References

Appendices

Appendix A - Mislabeled kaggle train tweets

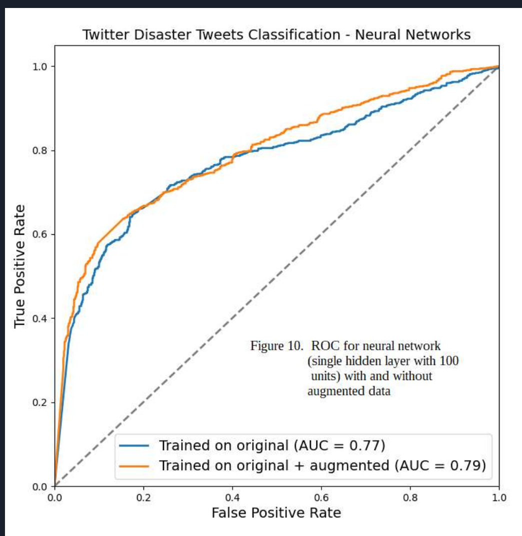
Summary of results from sample of 400 taken from each class:

target class	judged correct	judged incorrect	unsure
0	391	2	7
1	264	96	40

Examples where target class was considered “unsure”:

id	text	target	notes
5356	It may seem like our fire has been a little burnt out....	0	word sense is not clear here
9077	@sirtophamhat @SCynic1 @NafeezAhmed @jeremyduns and of course you don't have to melt the steel in order to cause structural failure.	0	could have been referencing an active fire
3066	500 deaths a year from foodborne illness... @frackfreelancs dears... @DECCgovuk @frackfree_eu @tarleton_sophie http://t.co/JSccX8k0jA	1	ambiguous reference
6187	Governor allows parole for California school bus hijacker who kidnapped 26 children in 1976. http://t.co/hdAhLgrprl http://t.co/Z1s3T77P3L	1	don’t see how this can be considered a disaster

Appendix B - neptune.ai RNN vs. Single hidden-layer NN



RNN trained on the same kaggle Disaster Tweets dataset.

	Without Data Augmentation	With Data Augmentation
ROC AUC score	0.775	0.785

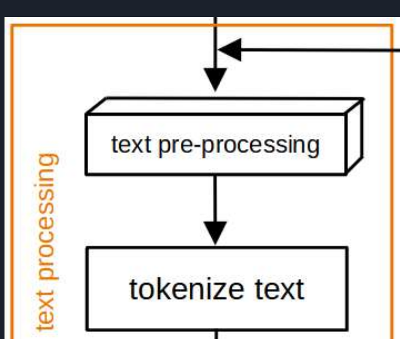
https://neptune.ai/blog/data-augmentation-nlp

Appendix C - Text processing example

I can't bloody wait!! Sony Sets a Date For Stephen King⑨Ûªs ⑨Û÷The Dark Tower⑨Ûª #stephenking #thedarktower http://t.co/J9LPdRXCDE @bdisgusting

➔ normalize URLs

I can't bloody wait!! Sony Sets a Date For Stephen King⑨Ûªs ⑨Û÷The Dark Tower⑨Ûª #stephenking #thedarktower <url> @bdisgusting



➔ normalize twitter special chars

I can't bloody wait!! Sony Sets a Date For Stephen King⑨Ûªs ⑨Û÷The Dark Tower⑨Ûª <hashtag>stephenking <hashtag>thedarktower <url> <user>bdisgusting

➔ expand contractions

I can not bloody wait!! Sony Sets a Date For Stephen King⑨Ûªs ⑨Û÷The Dark Tower⑨Ûª <hashtag>stephenking <hashtag>thedarktower <url> <user>bdisgusting

➔ remove stop words

~~I can~~ not bloody wait!! Sony Sets ~~a~~ Date ~~For~~ Stephen King⑨Ûªs ⑨Û÷The Dark Tower⑨Ûª <hashtag>stephenking <hashtag>thedarktower <url> <user>bdisgusting

➔ remove punc, lemmatize, lower case, remove singletons and OOV words

not bloody wait!! Sony Sets Date Stephen King⑨Ûªs ⑨Û÷The Dark Tower⑨Ûª <hashtag>stephenking <hashtag>thedarktower <url> <user>bdisgusting

not bloody wait sony set date stephen dark <hashtag> <hashtag> <url> <user>

➔ tokenize

not | bloody | wait | sony | set | date | stephen | dark | <hashtag> | <hashtag> | <url> | <user>

References

- Kaggle Disaster Tweets:
  https://www.kaggle.com/competitions/nlp-getting-started/data
- Nice list of generative AI models:
  https://github.com/steven2358/awesome-generative-ai
- RNN trained on kaggle Disaster Tweets data:
  https://neptune.ai/blog/document-classification-small-datasets
- Large Language Model Assisted Model Development project code, data, notebooks
  https://github.com/MichaelSzczepaniak/llmamd
- Feasibility project - final report
  https://github.com/MichaelSzczepaniak/llmamd/blob/main/docs/Final_paper.pdf