2022

Natural Language Processing for Disaster Tweets

Akinyemi D. Apampa
CUNY New York City College of Technology

Nan Li
CUNY New York City College of Technology

NATURAL LANGUAGE PROCESSING FOR DISASTER TWEETS

AKINYEMI APAMPA, NAN LI

DEPARTMENT OF MATHEMATICS

ABSTRACT

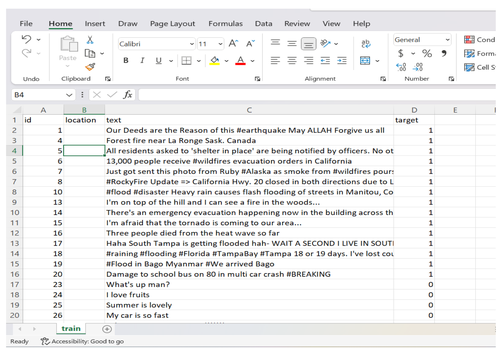
Our goal is to establish an automatic model that identifies which tweets are about natural disasters based on the content of the tweets. Our method is to construct a decision tree based on keyword searching. We will construct the model using 7,613 tweets and test our model on 3,263 tweets.

INTRODUCTION

Twitter has become an essential and effective medium to disseminate information, especially in emergencies. Due to this fact, many people and organizations monitor Twitter activity to locate tweets about certain significant events, like weather emergencies. This project aims to use machine learning to determine which tweets are about natural disasters and which are not.

METHODS AND RESULTS

- We are given a dataset comprising of 7613 tweets, the variables are:
  1. Id: A unique integer assigned to each tweet.
  2. Location: the location from where the tweets were sent.
  3. Text: The list of sample tweets.
  4. Target: A list of predictions for the tweets, the model constructed is to match these predicted values as closely as possible. Since this is the data which we are using to train our model, we will be using the tweets to formulate our model.



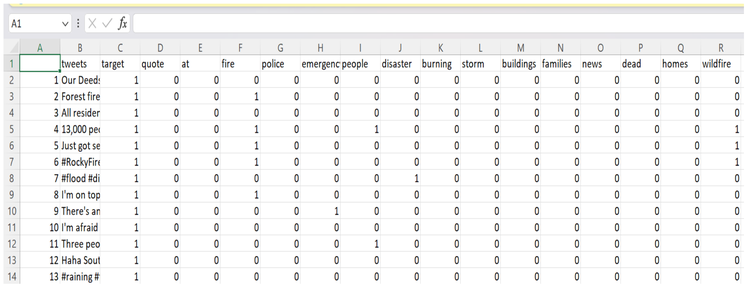
id	location	text	target
1		Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all	1
4		Forest fire near La Ronge Sask. Canada	1
5		All residents asked to 'shelter in place' are being notified by officers. No ot	1
6		13,000 people receive #wildfires evacuation orders in California	1
7		Just got sent this photo from Ruby #Alaska as smoke from #wildfires pour	1
8		#RockyFire Update => California Hwy. 20 closed in both directions due to L	1
10		#flood #disaster Heavy rain causes flash flooding of streets in Manitou, Co	1
13		I'm on top of the hill and I can see a fire in the woods...	1
14		There's an emergency evacuation happening now in the building across th	1
15		I'm afraid that the tornado is coming to our area...	1
16		Three people died from the heat wave so far	1
17		Haha South Tampa is getting flooded hah- WAIT A SECOND I LIVE IN SOUT	1
18		#raining #flooding #Florida #TampaBay #Tampa 18 or 19 days. I've lost cou	1
19		#flood in Bago Myanmar #We arrived Bago	1
20		Damage to school bus on 80 in multi car crash #BREAKING	1
23		What's up man?	0
24		I love fruits	0
25		Summer is lovely	0
26		My car is so fast	0

- **Step 1. Generating a Frequency Table:** We formulate a word frequency table. This is a table that displays the frequency of words that appear in the tweets. We do this by:
  1. Searching through every tweet in the dataset.
  2. Counting the total number of times that every word appears.
  3. Putting them in an ordered list.

The words in this table could indicate the feature of the tweet, if it is about disasters or not. A portion of the frequency table, shown below, displays the different words that appear in the tweets as well as the total number of times that they appear in the tweets.

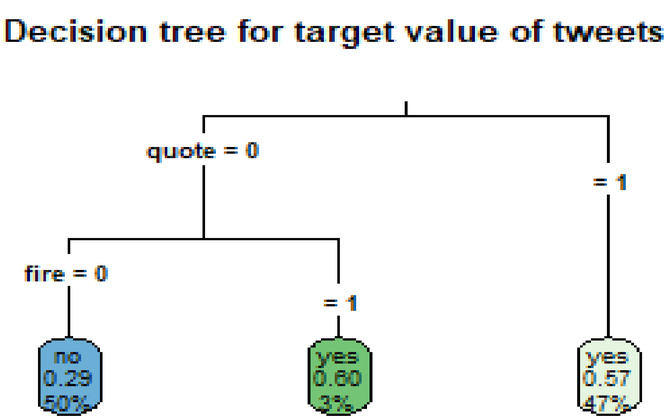
	A	B	C
1		Var1	Freq
2	25911	the	3194
3	6188	a	2127
4	18258	in	1943
5	26172	to	1932
6	21357	of	1813
7	6820	and	1397
8	18079	i	1325

- **Step 2. Keyword Selection:** We select a few keywords from the frequency table as our variables. We then classify these variables as presented or not in the given tweets, 1 being yes and 0 being no. We chose these variables because they were the most frequent meaningful keywords, that is, words that are not pronouns, in our frequency table.



	tweets	target	quote	at	fire	police	emergenc	people	disaster	burning	storm	buildings	families	news	dead	homes	wildfire
1	Our Deed	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Forest fire	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	All resider	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	13,000 pe	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1
5	Just got se	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
6	#RockyFir	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
7	#flood #di	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
8	I'm on top	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
9	There's an	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
10	I'm afraid	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	Three peo	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
12	Haha Sout	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	#raining #	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- **Step 3. Formulating a model:** The model(as shown in the diagram below) is constructed by minimizing our misclassification error over all possible splitting. The splitting stops when the improvement of the error is not significant enough.



- **Step 4. Interpretation of Results**
  - The decision tree looks through the keywords into two groups, 1 if it is present and 0 if it is not present. The keyword quote signifies tweets that are quoting another tweet.
    1. In our decision tree, if quote = 0 and fire = 0 then it is not a disaster tweet, if quote = 0 and fire =1, then it is a disaster tweet, and if quote =1 and fire =1, then it is also a disaster tweet.
    2. Looking at the decision tree, the value 0.29 signifies that 71% of the data is predicted correctly, the value 0.60 means that 60% of the data is predicted correctly, the value 0.57 signifies that 57% of the data is predicted correctly.
    3. The misclassification error is computed by comparing our model to the target values. We take the number of predicted values that are not consistent with the target values and divide that by the sum of the target values to find the misclassification error percentage. Our misclassification error percentage is 34%.
  - We also see the following features in the decision tree:
    1. Our model searches for quoted tweets first then the keyword "fire" and it searches for these two variables first. This indicates that they have more influence on the class compared to other variables.
    2. Since this is a competition project on Kaggle, we are provided with a testing data to test our model. Our testing error is 36%. The errors are consistent so the model works.

FUTURE WORKS

This model still has room to improve and it could be refined in the future to improve its accuracy.

REFERENCES

- Li, Nan. MAT 4800 Lecture Notes.
- "Natural Language Processing with Disaster Tweets." Kaggle, www.kaggle.com/competitions/nlp-getting-started/overview.

ACKNOWLEDGEMENT

I would like to thank Professor Nan Li for his guidance and knowledge on this project. I would also like to thank Undergraduate Research for giving me the opportunity to use this platform to gain this invaluable knowledge and experience.