Predicting Gasoline Shortage using Tweets during Disasters

Abhinav Khare
Department of Industrial & Systems Engineering
University at Buffalo
Buffalo, NY 14260
abhinavk@buffalo.edu

Samarth Kanabar
Department of Computer Science
University at Buffalo
Buffalo, NY 14260
skanabar@buffalo.edu

Rajiv Ranjan
Department of Computer Science
University at Buffalo
Buffalo, NY 14260
rajivran@buffalo.edu

Abstract

Shortage of gasoline is a common phenomenon during the onset of forecasted disasters like hurricanes. Identifying gasoline shortage can guide agencies in pushing supplies to the correct regions and mitigating the shortage. In this work, we developed a classifier to classify and identify tweets about gasoline-shortage from a corpus of tweets. To develop the classifier we used tweets generated in Florida during Hurricane Irma in the September 2017 (when gasoline shortage was observed). In our work, we tested various Convolutional Neural Network models (CNN), Long-Short Term Memory (LSTM) Recurrent Neural Network Models, and a hybrid model on the Irma data. We found that the LSTM method had the highest accuracy. Therefore, we further varied its hyperparameters and word embedding sources to find the best LSTM classifier. Our best model produced an F1-score =

1 Introduction

Recently, social media is transforming the way people communicate not only in daily lives, but also during disasters. There is a surge in usage of social media during an emergency in the affected regions. Nowadays, many people are willing to share the disaster information through social media. Public uses social media to communicate, seek information, raise concerns and express sentiments, and responders use it to plan and communicate important massages to the public. As a result, there is a keen interest in employing social media for disaster management. Consequently, multiple social media data analysis techniques have been developed in the context of a disaster, ranging from tools for event detection, prediction and warning; impact assessment; situation awareness;disaster tracking; and response planning.

However, there is scant literature exploiting social media data for detection and prediction of demand and shortage of essential commodities during disasters. We aim to fill this research gap by focusing on using social media data to predict the shortage of gasoline one day in advance (everyday)during the onset and post landfall of foreseen disasters like hurricanes. Our motivation comes from the fact that people have been found to use Twitter to tweet about shortages and needs during a disaster.For instance, during gasoline shortage in Florida in the onset of Irma, the following kinds of tweets were observed:

"The shelters are full, there is no gas. Tornados could happen, and storm surge is predicted. So what are people supposed to do? Irma "
"Insane..95 percent of Florida trying to leave at one time. Roads r slammed. No gas. No hotels available. Scared to see my neighborhood after irma"
"Gas stations out of gas, water shelves empty, stores and airports closed. Stocked up on food and wine, waiting on irma"

The natural question that arises is that can social media be used to detect and predict gasoline shortage. The first challenge associated with this process is *"How to identify tweets about shortage"*. Social media data, especially from twitter, is difficult to process and classify as it is unstructured, noisy and contains a plethora of information (large number of tweets). Also, a single tweet contains a maximum of 140 characters, is informal and contains abbreviations and spelling mistakes. Interpreting the semantics of such a short message and classifying it is a hard problem. There are methods in the literature that have classified tweets generated during crisis into caution/advise, information source, people, casualties and damage, pre-disaster or post-disaster(tweet4act), tweets reporting casualty or damage (Tweedr) , information, preparation and movement. However, classifying tweets for a specific problem like identifying gasoline shortage has never been done. Identifying important features for this classification task is a novel and unique question.

If these issues are resolved, then one can identify tweets about gasoline shortage and treat them as sensors for shortage.To address this challenge, we developed that classifies and identifies tweets about gasoline-shortage from a corpus of tweets. To develop the classifier we used tweets generated in Florida during Hurricane Irma in the September 2017 (when gasoline shortage was observed). In our work, we tested Convolutional Neural Network models (CNN), A Long-Short Term Memory (LSTM) Recurrent Neural Network Models, and a hybrid model on the Irma data. We found LSTM methods yielded the highest accuracy. So, we further varied its hyperparameters and word embedding sources to find the best LSTM model.

2 Data

Our data set had roughly one million tweets from Florida during the period 6-15 September 2017. The data covered a data frame in R with 1048575 rows and 41 columns that include TWEET ID, TWEET TEXT, USER ID, DATE, HASHTAG, LATITUDE, and LONGITUDE. Summary statistics of the tweet data is in Table 1.

Table 1: Summary statistics of tweet data

Summary Statistic	**Values**
Number of Tweets Collected	1,048,575
Number of Unique Twitter Users	111,801
Period of Data Collection	6th Sept 2017- 15th Sept 2017
Date of Irma Landfall in Florida	9th Sept 2017
Number of tweets prior to Irma landfall in Florida	456,530
Number of tweets during Irma in Florida	151,792
Number of tweets post Irma in Florida	440,253
Number of tweets about gasoline shortage	2594

The data was filtered using data filtering techniques described in Section 4. Following that, the data was labeled and it was found that there were 4070 gasoline-related tweets. Out of the 4070 gasoline-related tweets, 2594 were "gasoline-shortage" tweets i.e they were talking about gasoline shortage in their area at that time. Rest of the gasoline-related tweets were "non-gasoline-shortage" tweets. Sample of our final labaled data is in Table 2.

3 Related Work

There have been some work in the areas of using machine earning on tweets during disasters. There are several tools that have been used for disaster detection, disaster impact assessment, situation awareness and response planning. There are classifier that classify posts into caution/advise, information source, people, casualties and damage [4], pre-disaster or post-disaster(tweet4act) [2], tweets reporting

Table 2: Sample of labeled tweet data

SNO	TWEET TEXT	Labels
27751	Can someone explain to me why everyone is filling up their gas tanks?	1
27755	The drive and wait I had to make for gas is crazy af	1
27762	In the trunk of your car you need two gas cans case of water first aid kit and a road kit at least. Just in case of emergency	0
27806	If the hurricane hits and everyone tries to leave it will be impossible for everyone to leave. Yet we all have gas	0
27816	I need to stop driving like I have a race car since gas stations are running out of gas	1

casualty or damage (Tweedr) [1], information, preparation, movement etc. [6], into user-defined categories (AIDR) [3].

4 Data Filtering

we filtered out "gasoline-related" tweets from the compendium of tweets generated in the affected area. We do this by keyword search in both the content and hashtags of each tweet. In the case of gasoline any word which has the letters "gas" as part of the word is a possible keyword. We use regular expressions to identify these key words.The regular expression @gas finds words starting with "gas" and also finds words that contain the string "gas" (e.g. the word "nogas"). For searching tweets, we use the regular expression//bgas to look for words in a sentence starting with"gas". These regular expressions are the ones used with grep() function in R. Next, we identified the relevant keywords filtered through the regular expressions and retain the tweets containing those keywords. Finally,We combined the tweets curated from searching hashtags and tweets by using inner join,to account for duplicate tweets. Even after this filtering, we have many tweets that are very noisy and need to be cleaned so as to facilitate further processing. We achieved this cleaning by removing user names, links, punctuation, tabs, general whitespaces, stopwords, and numbers. We further changed words to stem words and to lower case. The hashtags and words that we found using the regular expressions includedgasoline, gas, gasinmiami, gaspricefixing, gasstation, gasservice, gas-tateparks, gasshortage, gasoil, gastation, gaswaste, nogas, outofgas, findgas.

5 Methodology and Experiments

To build the most accurate classifier we did 3 kinds of experiments:

1. Testing different kinds of Classifiers
2. Varying the Hyper-parameters of the most accurate classifier
3. Varying the Embedding Matrix

These experiments were conducted using Keras. In all these experiments the optimzer used was ADAM, the train/validation ratio was 60/40, and early-stopping was used for regularization (to avoid overfitting).

5.1 Testing different kinds of Classifiers

We tested 4 different kinds of ANN's:

1. LSTM
2. CNN with a Convolutional Layer
3. CNN with a Convolutional and a Maxpool Layer
4. LSTM + CNN with a Convolutional and a Maxpool Layer

The architectures and the parameters of these networks are described in detail in Figure 1 and Figure 2.

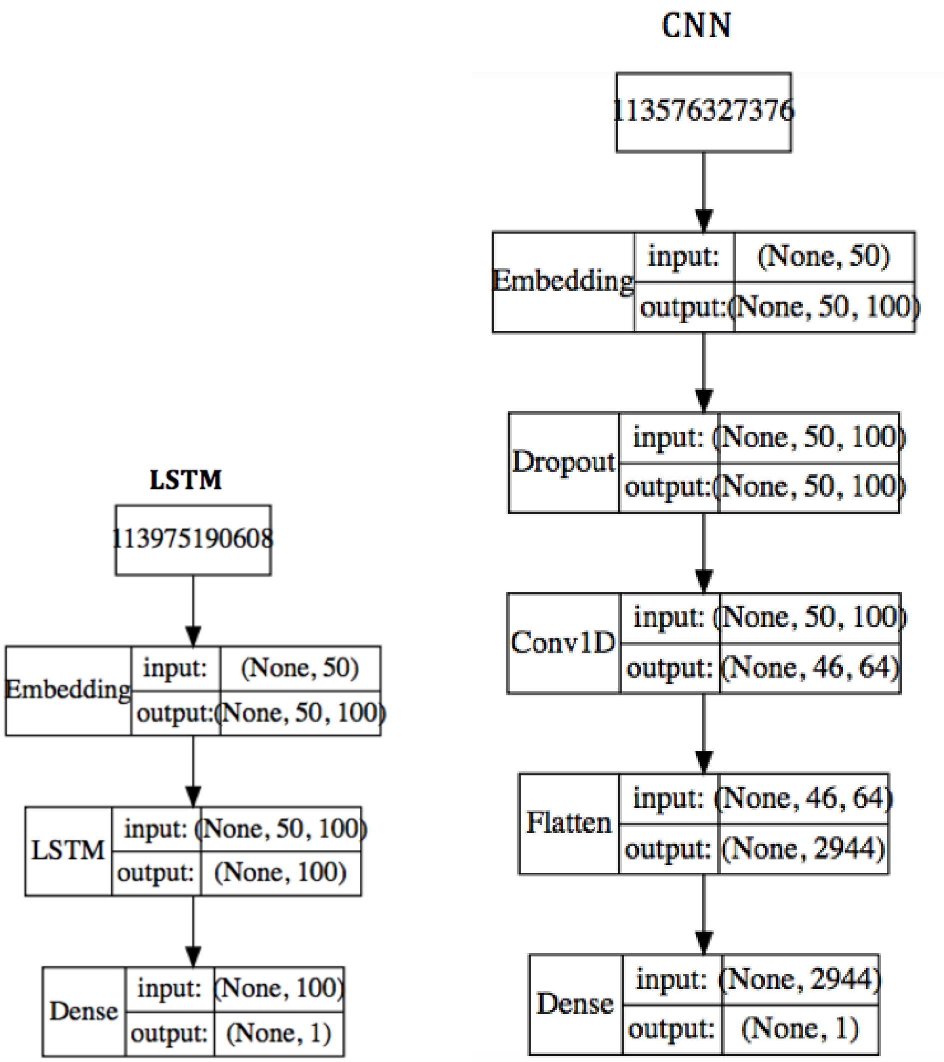


Figure 1: Architectures of different LSTM and CNN tested

5.2 Varying the Hyper-parameters of the most accurate model

We found that LSTM was the best performer amongst all the classifiers as described in the result section. Hence to find the most accurate LSTM classifier, we varied the following hyper parameters:

1. Number of units in the LSTM Hidden Layer
2. Dropout Ratio
3. Learning Rate of Adadelta Optimizer

5.3 Varying the Word Embeddings

We used three different kinds of Word Embedding:

1. Self-trained Word Embeddings in Keras
2. Word Embeddings from CrisisNLP's github repository [7]
3. Word Embedings from Glove [5]

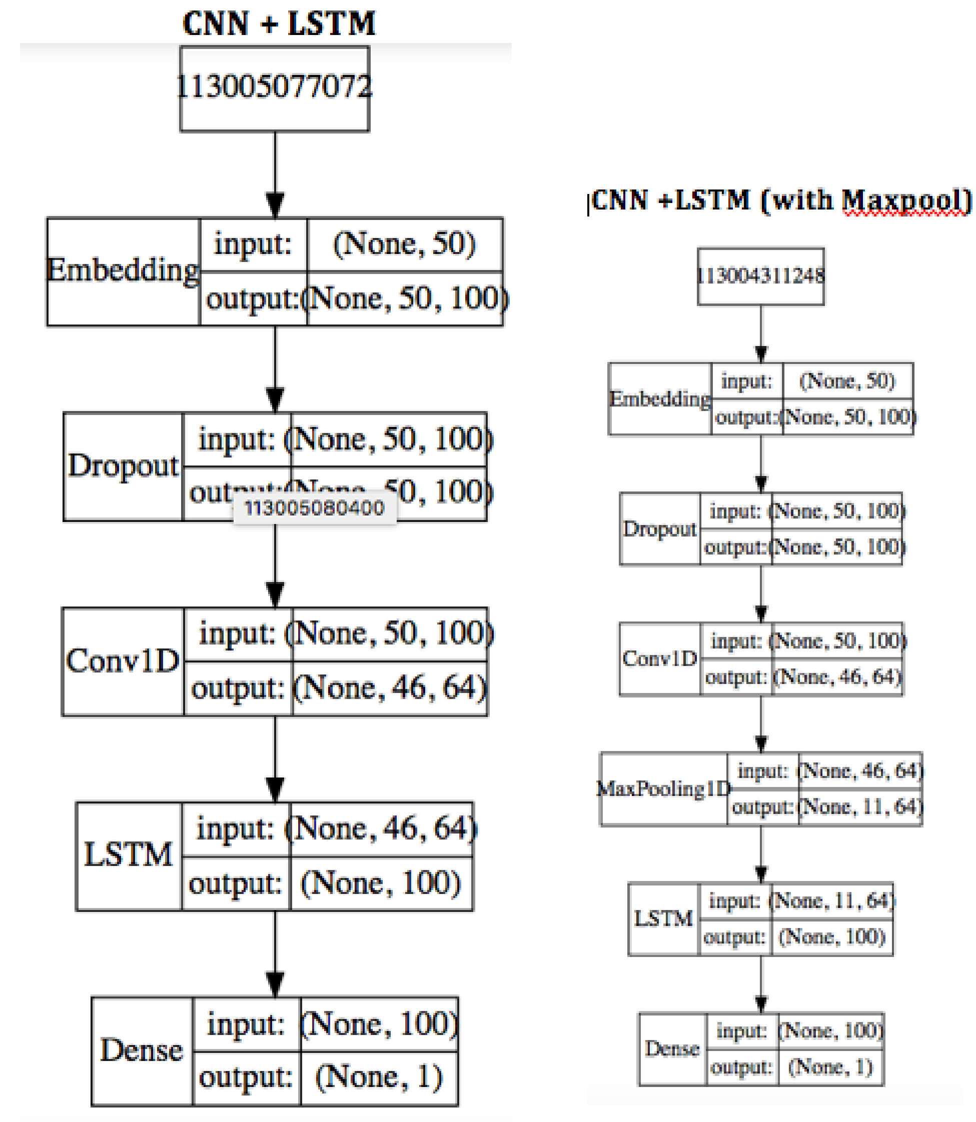


Figure 2: Architectures of Hybrid Models

6 Results and Discussion

The results have been discussed in three subsection for each experiment type. In all the experiments the optimzer used was ADAM, the train/validation ratio was 60/40, and early-stopping was used for regularization (to avoid overfitting)

6.1 Results on different Classifiers

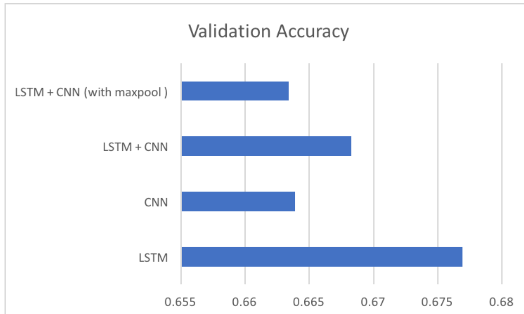
Table 3 shows the training and validation accuracy for the different classifiers. It is clear that the LSTM. It is very clear that LSTM is the best performer with highest validation accuracy.

6.2 Results on LSTM with varying Hyper-parameters

Table 4 shows the training and validation accuracy as we vary the number of parallel units of LSTM (dimensionality of the output space), the dropout ratio and the learning ratio.

Table 3: Performance of different classifiers

Model	Accuracy	Validation Accuracy
LSTM	0.9382	0.6769
CNN	0.9889	0.6639
LSTM + CNN	0.9709	0.6683
LSTM + CNN (with maxpool)	0.9533	0.6634



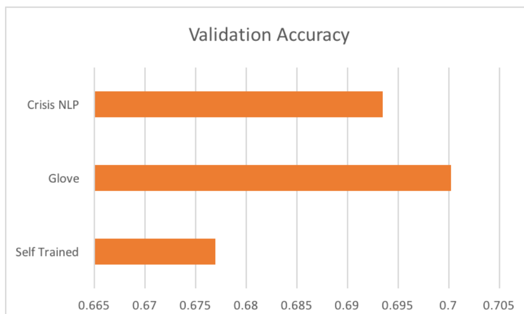


Figure 3: Validation Accuracy variation with models and embeding matrix

The highest **validation accuracy = 0.6947** is achieved for **units = 25, drop out ratio = 0.0 , learning rate = 1** and **units = 100, dropout ratio = 0.5, learning rate = 1.** . There is no monotic trend of accuracy with change in learning rate. It can be seen the validation accuracy is significantly poor at 0.5405 for learning rate = 0.1 but the performace in gerneral is good at learning rate = 0.1. With increase in droput ratio it increases monotonically increases. With change in units it reaches its optimal value around a mid range value of 25.

Table 4: Performance of LSTM with varying Hyper-Paramters

Units	Drop out ratio	Learning Rate	Accuracy	Validation Accuracy
5	0	1	0.9681	0.6935
25	0	1	0.9603	0.6947
50	0	1	0.9713	0.6873
100	0	1	0.9382	0.6769
100	0.1	1	0.9664	0.6713
100	0.2	1	0.9578	0.6898
100	0.5	1	0.955	0.6947
100	0	0.001	0.9787	0.6744
100	0	0.01	0.9934	0.6517
100	0	0.1	0.7248	0.5405

6.3 Results on different Embedding Matrix

Table 5 shows the training and validation accuracy for different word Embeddings. It is very clear from the table that the best performance is observed for the Glove embedding.

Table 5: Performance of LSTM with different Word Embeddings

Word Embedding	Accuracy	Validation Accuracy
Self Trained	0.9382	0.6769
Glove	0.8354	0.7002
Crisis NLP	0.8595	0.6935

References

[1] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta. Tweedr: Mining twitter to inform disaster response. *ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management*, (May):354–358, 2014.

[2] Roy Chowdhury, Soudip Roy Chowdhury, and Carlos Castillo. Tweet4act : Using Incident-Specific Profiles for Classifying Crisis-Related Messages. *Proceedings of the 10th International ISCRAM Conference*, (May):834–839, 2013.

[3] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. AIDR: Artificial intelligence for disaster response. *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, (October):159–162, 2014.

[4] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1021–1024. ACM, 2013.

[5] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[6] Kevin Stowe, Michael J Paul, Martha Palmer, Leysia Palen, and Kenneth Anderson. Identifying and categorizing disaster-related tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 1–6, 2016.

[7] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. Rapid classification of crisis-related data on social networks using convolutional neural networks. *arXiv preprint arXiv:1608.03902*, 2016.