Disaster Tweets Classification

Team Member:

Seyed Ali Sadat Akhavani – sadatakhavani.s@husky.neu.edu

Objectives and significance:

The goal of this project is to create a model that identifies whether a tweet is talking about a real-world disaster or not. Doing this work is not easy because a lot of people are using twitter for different purposes. It's not always clear whether a person's words are announcing a disaster. It is not enough to Look for a word such as "disaster" to find out if that person is talking about a real disaster or not. Because that tweet may be a joke, or even that person is complaining about his/her day and says it was a disaster for that person. So, I think that creating a model to identify the type of the tweet is interesting and also challenging.

Working on a problem related to social media is interesting for me. So, I decided to pick a project that is related to that. Also, natural language processing (NLP) is one of the famous subjects in Machine Learning that I like to work on it. Twitter is one of the most famous social media applications that a huge number of people around the world are using to express their feelings, talk about news, disasters, and a lot of other things. It has also become an important communication channel in times of emergency. So, I chose Twitter as my target for this project.

Background:

In this project, we are working on tweets so the main thing that we are going to face is natural language processing (NLP). Because we have to do a lot of process on tweets which are consist of texts. One of the famous models that is being used in most of NLP methods is called Bag-of-Words [1] model. The bag of words model is a simplifying representation used in natural language processing and information retrieval. In this model, a text is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity. We are going to use this model in our method.

Twitter is a famous social media, so there exists lots of different machine learning studies on different parts of it that include lots of topics. I found some projects that had similar ideas for tweet classification but there is going to be a lot of difference between my work and theirs.

One similar project is about tweet stance classification.[2] They are using a method called transfer learning which is a technique where instead of training a model from scratch, they use models pre-trained on a large dataset and then fine-tune them for specific natural language tasks. We are not going to use this method in our project but this is one of the famous methods in NLP. They used their model on different datasets about these topics:

- Atheism
- Climate change is a concern
- Feminist movement
- Hillary Clinton
- Legalization of abortion

They have found interesting results too. In this model, each tweet can be in one of these three classes:

- Favor
- Against
- No Stance

This is an interesting work but it is not useful in our topic and has some fundamental differences with what I am going to do in my project.

There is also another work on tweet classification. In this project, they try to find out about customers' stance in tweets.[3] Through sentiment analysis, companies can automatically process what their consumers write in natural language and get valuable insights in order to take decisions. So, they are trying to create a model to automatically understand whether a customer is happy or not by reading his/her tweet. Each tweet in this project is considered as one of these types:

- Positive

- Neutral
- Negative

They use different methods such as Naive Bayes, Logistic Regression, Convolutional Neural Network, and Recurrent Neural Network. In the end, they evaluate all of these methods and find the best one for their work. I will use some ideas that are presented in this work. But not all of the work is related to my project and I have to use some different ideas to fit them in my topic.

As I said before, there may exist some previous work on tweet classification. But first, our work is different from them and none of them are working on such a thing like this. Because in this project, user exaggeration may easily lead to a wrong result. But for example, in a customer review tweets, it is so rare to see a tweet that is joking about his/her opinion on the product. But lots of people may use words such as disaster or similar words for their everyday lives. So, in our case, we have to use more parameters to detect the real intention of the user.

Also, we are going to use some tweet metadata in our learning too. I did not find much works that include tweet links, location, or images in their learning method. But I am going to try and combine tweet text with links, images and location data in my learning model and try to find out a well-designed model.

Proposed approach:

Dataset:

For this project, we have the training dataset and the test dataset. I have collected the dataset from a Kaggle competition which is called “Real or Not? NLP with Disaster Tweets”.[4] This dataset contains 10,000 tweets that were hand classified. Each tweet can have two target values. If a tweet is talking about a real disaster is has target value of 1. And if the tweet is not talking about a real-world disaster its value is 0.

Each sample in the training and test dataset has the following information:

- The text of a tweet
- A keyword from that tweet (can be blank)

- The location the tweet was sent from (may also be blank)

We are going to predict whether a given tweet is about a real disaster or not. If so, predict a 1. If not, predict a 0.

These are the columns in our dataset:

- id - a unique identifier for each tweet
- text - the text of the tweet
- location - the location the tweet was sent from (may be blank)
- keyword - a particular keyword from the tweet (may be blank)
- target - in training dataset only. This denotes whether a tweet is about a real disaster (1) or not (0)

Method:

I am going to use these methods in order to create a model for this data. As I said before, the data contains some missing values so first, I have to fill those missing values in those features with no_keyword and no_location.

After that, I have to find out more data about each tweet. Some things that I think are important in tweet evaluation are:

- Word Count
- Hashtag Count
- Punctuation Count
- Exclamation Mark Count
- URL Count

So, I have to measure all of these values for each tweet and find out whether they are important for our target value or not. And I have to use the important features in my model creation.

After that we have to clean tweet texts. A lot of tweets are messy because they are written in a person’s everyday life. Tweets are different from academic journals and we

have to do a lot of cleaning before trying to process them. For the cleaning step, I am considering these works:

- Removing Unwanted Words
- Transforming Words to lowercase
- Removing Stop Words (such as “the”, “a”, “an”, “in”)
- Stemming Words (reducing words to their word stem)

I am going to create a sparse matrix that is used in the Bag of Words method. After that, I am going to use different models such as Naive Bayes, Logistic Regression, and Support Vector Machine and evaluate all of them. We use the F1 score to evaluate our model. In order to compute the F1 score for our model, first we have to compute Precision and Recall.

Precision is the number of True Positives divided by the number of True Positives and False Positives. Put another way, it is the number of positive predictions divided by the total number of positive class values predicted.

Recall is the number of True Positives divided by the number of True Positives and the number of False Negatives. Put another way it is the number of positive predictions divided by the number of positive class values in the test data. It is also called Sensitivity or the True Positive Rate.

The F1 Score is the 2*((precision * recall) / (precision + recall)). It is also called the F Score or the F Measure. Put another way, the F1 score conveys the balance between the precision and the recall.

Also, since this dataset is from a Kaggle competition, the evaluation is easy. I just need to submit my work in Kaggle and see my F1 score. They have a big dataset that evaluates our model and generates an F1 score for our model. At the moment, the median value for the F1 score in this Kaggle competition is around 0.78. My goal is to achieve an F1 score higher than this.

If I find out that the images and links data are not enough and there are a lot of tweets without this kind of data, I have to put more focus on the text recognition part. In the first look, I see that there exist lots of tweets that do not have location data. But I need to do more exploration of the data in order to recognize whether we can rely on images in tweets or not.

Individual tasks:

Since I am doing this project individually, there are not any other team member and I am doing all the works for this project. I have to clean data, figure out a way to handle unlabeled data, find a machine learning solution and implement it.

References:

[1] - https://machinelearningmastery.com/gentle-introduction-bag-words-model/

[2] - https://towardsdatascience.com/transfer-learning-in-nlp-for-tweet-stance-classification-8ab014da8dde

[3] - https://medium.com/@martinpella/customers-tweets-classification-41cdca4e2de

[4] - https://www.kaggle.com/c/nlp-getting-started/overview