DATA CHALLENGE II

Plan – Kaggle case

20-11-2021

Kristina Krasteva

Lia Boyadzhieva

Table of Contents

1. Description

This Data Challenge is inspired by a prediction competition in Kaggle – Natural Language Processing with Disaster Tweets. Here is a link to the competition page and the dataset - https://www.kaggle.com/c/nlp-getting-started/overview

Nowadays, Twitter has become an important communication channel in times of emergency. The pervasiveness of smartphones enables people to announce an emergency they are observing in real-time. Because of this, more agencies such as disaster relief organizations and news agencies, are interested in programmatically monitoring Twitter. But it is not always clear whether a person's words are announcing a disaster. For instance, a person can tweet about the sky being ablaze. However, here the word "ablaze" is used metaphorically. For the people reading this post is clear what the author meant but is not that clear for a machine to distinguish figuratively meanings.

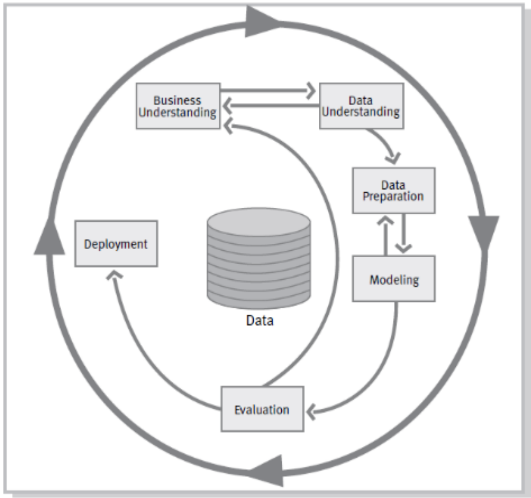
2. Goal of the challenge

The goal for this challenge is to build a machine learning model that predicts which Tweets are about actual disasters and which ones are not real disasters by predicting (1) for disaster and (0) for not. The model should distinguish those tweets containing words usually regarded as disasters (i.e., earthquake, flood, pandemic, etc.) which are used in a metaphorical way are not disasters. For the purpose Natural Language Processing will be applied.

The dataset that is going to be used consist of 10,000 tweets that were hand classified. The columns in both the train and test csv files are "id" - a unique identifier for each tweet, "keyword" - a particular keyword from the tweet (may be blank), "location" - the location the tweet was sent from (may be blank), "text" - the text of the tweet. Furthermore, the train dataset has a column "target" which designate whether a tweet is about a real disaster (1) or not (0).

3. Approach

For this Data Challenge the standard IBM CRISP-DM methodology will be used. There are 6 phases to be considered – **Business understanding, Data understanding, Data preparation, Modeling, Evaluation and Deployment**. These stages are linked to one another to help us solve the particular data science problem.



The data to be used is the one published by Kaggle where a train and test csv files are provided.

Additionally, we intent to focus mainly on the first 5 phases since there we can be more flexible when it comes to the project's iteration. Below a more in-depth description of what are the major task per stage is shown.

3.1. Business understanding

Before implementing any solution, the problem needs to be understood and the goal to be determined. This step is of great importance as it will result in a more reliable starting point for the actual modeling.

For this part the following things have to be done:

- **Determine the Business Objectives** – Background (what is the influence and impact of disaster tweets), Business Success Criteria (aimed for Disaster Relief Organizations/News Agencies)
- **Assess Situation** – Inventory of resources (both students working on the Data Challenge, taking into account how qualified we are), Constraints and Risks tables

3.2. Data understanding

When the business part of the process has been clarified, the data understanding stage arrives. It is an extremely essential phase since it will provide us with first insights about the data.

For this part the following things have to be done:

- **Collect Initial Data** – Gathering (or accessing) the data chosen from Kaggle. List the datasets acquired, together with their locations, the methods used to acquire them, and any problems encountered.

- **Describe Data** – Describing the collected data (the format of the data, the quantity of data such as the number of records and fields in each table).
- **Explore Data** –Exploring the gathered data by querying, visualizations (EDA).
- **Verify Data Quality** – To verify the data questions like "Is the data complete?", "Are there any missing values?", "Does the data contain errors and, if there are errors, how common are they?" must be answered and reported.

3.3. Data preparation

This stage is about preparing the final dataset that will be fed into the model. Practice shows that the data preparation tasks are likely to be performed multiple times.

For this part the following things have to be done:

- **Select Data** – Decision on what data will be used for the analysis. The criteria to be considered includes relevance to the end goals, quality, and technical constraints.
- **Clean Data** – Raising the quality of the collected data by removing the irrelevant data.
- **Construct Data** – Description of any new data records that were made so to make the data more appropriate for the future model.
- **Integrate Data** – Describe any instances where the data is combined from multiple tables or records to create new records or values.
- **Format Data** – Syncing the data when all of the above-mentioned techniques are applied.

3.4. Modeling

When the data is ready, the modeling happens where a number of different techniques are selected and applied, and their parameters are tuned in search for the optimal solution.

For this part the following things have to be done:

- **Select Modeling Techniques** – We will select several modeling approaches based on the data, the requirements, the wanted results.
- **Build Model** – Run the model on the dataset that was prepared from the phases above.

- **Assess Model** – When the model was executed with the specified parameters, assess the output, and tune those parameters to obtain better results.

3.5. Evaluation

In this phase, the final evaluation and reviewing of the models is done. Here, the importance lays on the business value again and if everything has been sufficiently considered.

For this part the following things have to be done:

- **Evaluate Results** – This step is a bit similar to the "Assess Model" one from the previous phase, but here we will focus on comparison between the different approaches (confusion matrix, plot accuracy and loss, etc.).

- **Review Process** – After the results from the models are satisfactory, here if we have the chance to examine there are some other factors that had lower priority or were overlooked.

3.6. Deployment

Usually, the final step is to launch the model. However, as we are more focused on experimenting with the different approaches, our deployment phase would be more like an advice which model is the best for our particular case.

For this part the following things have to be done:

- **Produce Final Report** – A final report is required at the end of the project.

4. Timeline

The duration of this data challenge is 7 weeks (from week 11 to week 18).

What	Duration	When
Data Challenge II Plan	1 week	Week 11
Business Understanding	1 week	Week 12
Data Understanding (+EDA)	1 week	Week 13

Data Preparation (+Text Cleaning, Tokenization, Lemmatization)	2 weeks	Week 15
Modeling (LSTM, TF-IDF) + Evaluation	1 week	Week 16
Modeling (BERT, GloVe) + Evaluation	1 week	Week 17
Final report	7 weeks (throughout the whole challenge)	Week 18
Final presentation	1 week	Week 18

5. Expected results

Throughout the challenge we are going to experiment with various models and hyperparameters in order to achieve as better accuracy score as possible. We expect to achieve high accuracy score for at least one model to be approximately 90%.

As for personal goal we expect to gain more knowledge for Natural Language Processing. Therefore, we can achieve a higher level of progression for the learning outcome "Algorithms". Moreover, with this challenge we will contribute for better progression for few other outcomes such as "Dataset", "Structured approach" and "Co-working".