Data Challenge ii

Final Report – Kaggle case

11-12-2021

Kristina Krasteva

Lia Boyadzhieva

Table of Contents

[1. Approach 6](#_Toc90126680)

[1.1. Business understanding 6](#_Toc90126681)

[1.2. Data understanding 6](#_Toc90126682)

[1.3. Data preparation 7](#_Toc90126683)

[1.4. Modeling 7](#_Toc90126684)

[1.5. Evaluation 8](#_Toc90126685)

[1.6. Deployment 8](#_Toc90126686)

[2. Timeline 8](#_Toc90126687)

[3. Expected results 9](#_Toc90126688)

1. **Data Challenge II Plan**

The Plan took till the end of week 11 to get finished. The first thing we as a team had to do was to look up the challenges in Kaggle. When we selected several of them we had a discussion what is the potential value of each challenges regarding our learning progression. For instance, one of our options was the Google Landmarks challenge but we decided against it as its focus is on CNN which we are already actively applying in our group project. And since both of us are interested in NLP but have only used it in one Core Program exercise, we wanted to improve our skills and expand our knowledge. We both liked the Disaster Tweets NLP challenge and started to brainstorm a more specific goal for this challenge.

The plan was prepared by both of us – Lia writing the description, goal and expected results, and Kristina taking care of the approach and the timeline. After we were finished we did a peer review of each other’s chapters and suggested what could be improved.

Throughout this week we asked and received feedback multiple times – verbally on our initial idea, and written feedback on the plan document itself which we intend to apply in the upcoming weeks.

1. **Business Understanding**

In week 12, we were mainly focused on researching what is the influence on Twitter and the impact that disaster tweets can have.

Twitter was brought to the horizon of the Internet in 2006 and ever since then it has become one of the most famous social website with millions of users around the globe. Despite the fact that other medias such as Instagram, Tiktok and Facebook have far more users, the traffic of information which Twitter accumulates is still impressive and relevant. What differentiates Twitter from the other social networks mentioned is that it has this kind of openness to everyday news and information. While Facebook is more focused on the people one is acquainted to, Tiktok is based on algorithms showing short videos of one’s interests and Instagram is mainly about sharing photos, Twitter spreads information which can be discovered with the help of hashtags.

In recent years, this has led to Twitter becoming an integral part of the information flow. The news are no longer only controlled by the newspapers and tv, but also by the Twitter users. They can post, comment, and share about local and global events almost in real time. Exactly here we are faced with the challenge of determining which disaster tweets are real and which are not.

With this data challenge, we aim to explore the power of AI and specifically NLP and test if the models such as BERT, XGBoost, and LSTM are able to distinguish between tweets with real information about disasters and ones where it was meant metaphorically.

Additionally, we will also prepare Constraints and Risks tables which can be seen below.

* Constraints

The purpose of the below mentioned project constraints is to outline what restricts or dictates the actions of this data challenge. There are some general constraints that can occur in almost every project, but we have also included some more specific ones which are tightly related to the NLP Disaster Tweets challenge.

|  |  |
| --- | --- |
| Constraints Table | |
| What | **Why** |
| Time constraint | The Deep Learning Challenge must be finished within 7 weeks. It starts on November 16, 2021, and is expected to be completed on January 21, 2022. |
| Dataset quality | The dataset provided is from Kaggle and even though it has quite a lot of records, there are very few features. This can lead to poorly performing algorithms, but we plan to extend the dataset with extracted features. |
| Resource constraint | The people (our team), equipment – computers, Internet, available data etc., knowledge which are required to successfully finish the challenge. |
|  |  |
|  |  |

* Risks

Before starting to work on the Deep Learning challenge more in-depth, certain project risks are acknowledged. This is needed because in its essence a risk is an uncertain event or condition that, if it occurs, influences at least one project objective. In the table the risks are outlined, the effect that they can cause in terms of damage to the final product and development, the probability of such risk to happen and the mitigation method or how to resolve such risks. Effect and probability will be measured as Low, Medium, and High.

|  |  |  |  |
| --- | --- | --- | --- |
| Risk Table | | | |
| Risk | **Effect** | **Probability** | **Mitigation** |
| Bad quality of the dataset which causes troubles with the prediction models | High | Medium | Explore and try to improve the dataset by adding more data to it. |
| None of the models achieve the desired accuracy | High | Medium | Re-evaluate the models and spend more time on training data, try different machine learning approaches. |
| The dataset is not cleared and prepared properly | Medium | Medium | Iterate the process, research on different techniques to prepare the dataset in the most optimal way. |
| Lack of required knowledge of student. | Low | Medium | Perform research activities online. Additionally, ask the other team member for help. |
| Behind schedule | Medium | Low | The communication between the two students happens frequently and the tutors are updated about the work done weekly. Keeping track of the progress in the report document. |

1. **Data Understanding (+EDA)**

In week 13 we got familiar the data provided by Kaggle. There are 7614 records in train dataset and 3264 in test dataset. The total number of fields in the train dataset are five – “id”, “Keyword” – serves as category of the tweet, “Location” – the country or city that is mentioned in the tweet, “Text” – the tweet, and “Target” – defines if a tweet is real or fake disaster by the values of 1(real disaster) and 0 (fake disaster). This last column is not present in the test dataset.

Each of us created a Jupyter notebook focused on exploring the data and visualizing it. In this way we were able to understand it. For example, how many missing values and from which columns of the of both the train and test datasets were displayed. The number of fake vs real tweets were also shown. A word cloud with the most common words from the tweets were generated as well.

Since there were null values in the “Keyword” column as well as in the “Location” we checked if that is common for defining whether the tweet is disaster or not. The total null values in “Keyword” were 61. We plotted the null values only where the tweet is marked as fake disaster and the number of values decreased to 19. Therefore, it can be said that the tweet is not defined as a fake one based on whether it has a keyword. Moreover, we decided that we can fill in manually the missing values for “Keyword”. It was not difficult to decide what the keyword should be since most of them were about real disaster tweets.

However, regarding the “Location” there were in total 2533 null values which 1458 are for fake disaster tweets. From here comes the thought that when a tweet is about a real disaster most of the times it is mentioned where this disaster has happened. Still the null values are too many and will be difficult to aim the model.

1. **Data Preparation (+Text Cleaning, Tokenization, Lemmatization)**

For two weeks (14 and 15), we are cleaning the dataset and in this way we hope to prepare a better train dataset to be used for our future models. We have done this individually and later we will compare results and give suggestions to each other what can be added as a cleaning step.

The first actions taken was to clean the tweets by making all words lowercase, remove hyperlinks, hashtags, mentions, emojis, punctuation signs and all kinds of symbols and numbers which will not be useful later. This removal was conducted with the help of regular expressions. Moreover, in the “Keyword” column there were a lot of records with “%20” between two words which stands for space. We replaced it with underscore for better interpretation and visualization.

As part of the data preparation, we extracted the URLs and numbers used in tweets in separate columns. In this way we can use them as additional features because they might be a sign for better predictions of the model. For instance, most of the tweets which have URLs or numbers are marked as disaster.

Furthermore, we have removed any stopwords. In our case these are the English words which do not add much value to the meaning of a sentence. They can safely be ignored without sacrificing the meaning of the sentence. The following code: “stopwords.words('english')” removes all stopwords in the English language such as “the, a/an, our, we,”, etc.

Thanks to the word cloud visualization, we have noticed that specifically in our dataset the words “u”, “im”, “c” and “amp” are one of the most common ones. However, they most likely will not bring value to the training of the models, so we decided to remove them as well.

Next, tokenization was applied. The purpose of the tokenization is to break the raw text into small chunks (words called tokens). These tokens help in understanding the context or developing the model for the NLP. The tokenization helps in interpreting the meaning of the text by analyzing the sequence of the words.

Additionally, we have applied the most common pre-processing technique in Natural Language Processing – lemmatization. For this challenge the understanding of the tweet context is important. Thus, we managed to group together the different inflected forms of a word so they can be analyzed as a single item.

After tokenization and lemmatization, we transformed the tokens into sentences.

**Pad Sequencing??? (**[**https://medium.com/geekculture/fake-or-not-twitter-disaster-tweets-f1a6b2311be9**](https://medium.com/geekculture/fake-or-not-twitter-disaster-tweets-f1a6b2311be9)**)**

Asked for feedback (Bartosz)

1. **Modelling**

During week 15, we also started to experiment with different models. Lia applied BERT on the dataset, while Kristina used XGboost.

*Bert*

Recently, BERT is one of the most used models for Natural Language Processing. With the used of Transformer encoder the model can learn the context of a word based on all of its surroundings. This is exactly what we need for this challenge – to define whether a keyword for disaster is used metaphorically (fake disaster) or indicates a real disaster.

The model was applied two times because during the first time the acquired validation accuracy was 80% with training on 5 epochs and a batch size of 10. This result was achieved by having two input layers, two dense layers and a dropout of 20% between them. Then it was decided to do some hyperparameter tuning to see if there the model will improve.

For the second try we increased the number of epochs from 5 to 10. Additionally, we left only the last dense layer for the output, increased the shape to 128 whereas in the first attempt was 64 and set the batch size to 32. The training of the model took longer time (around 7 hours) and was stopped at epoch 8 with validation accuracy of 81%. Interestingly, from epoch 5 to 7 the validation accuracy remained 0,8096.

As next steps for the BERT model can be considered few optimizations in the dataset. Moreover, when the model will be trained again Early Stopping can be applied in case there is no improvement in the accuracy in the range of few epochs.

# Diagram Description automatically generatedApproach

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.

## 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

## 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.

## 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.

## 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.

## 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.

## 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.

# 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 |

# 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”.