* **What data set(s) you are using:**

**1. Eurovision Song Contest participants (website)**

For example: <https://eurovision.tv/event/dusseldorf-2011/participants>, I haven’t found any list yet. So I need some time to create 9 lists of participants for each year)

**2. Eurovision Song Contest winners (website)**

<https://nl.wikipedia.org/wiki/Lijst_van_winnaars_van_het_Eurovisiesongfestival>

**3. Twitter API (for academic research)**

<https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api>

* **What is your** [**unit of observation (Links to an external site.)**](https://en.wikipedia.org/wiki/Unit_of_observation)**? That is, what are the variables a variable of? What does a row in your final data set represent? Date/person/session/etc.? And how many cases (days/people/sessions) do you have (approximately)? You should have at least about a 100 rows, preferably several hundred and the more, the better.**

1. The **unit of observation** is a Tweet which includes a name of the artist from the Eurovision Song Contest

2. The **amount of cases** will be based on the most trending Tweets about the Eurovision Song Contest per year including a name of the participant of that year. I think I can go for a top 1000 of most trending Tweets for each year (1000 x 9 cases in total). The Twitter API with the license I use has a limit of 500,000 Tweets / month per project. So, if it makes sense, I can easily go for 10.000 x 9 cases in total.

* **What you want to predict: Y variable**

The question for this project is: can Tweets predict the winner of the Eurovision Song Contest? Therefore, the Y variable will be the about if **the artist is known as a winner yes/no**. I’ll investigate if this artist is actually the winner of the contest.

* **What variables you think of using as a predictor: X variables. You should have at least 6 variables, preferably at least a dozen, and the more, the better. If you are using a text variable, a single text variable will suffice (more is possible). The texts have to be at least the length of a tweet and you should have several hundred of them.**

**0. Id (unit of observation: one tweet)**

1. **Retweets**

How many times the tweet has been retweeted

1. **Likes**

How many times the tweet has been liked

1. **Comments**

How many comments the tweet has

1. **Followers Twitter account**

How many followers the twitter account of the tweet has

1. **Sentiment**

The sentiment of a tweet based on a sentiment score

1. **Location**

Geolocation of the Tweet (if the tweet includes a geolocation)

1. **Year**

Year of the tweet (2011-2019)

1. **Month**

Month of the tweet (January-May)

1. **Day**

Day of the month (January 1 – May *day of the finals* )

1. **Time of the day (CEST)**

Time of the day (before the time the winner is announced on the day of the finals)

1. **Mentioned artist**

The artist that has been mentioned in the tweet

1. **Winner per year**

The winner of the Eurovision Song Contest for each year

1. **Matches the winner (true/false)**

Is the mentioned artist also the winner? (yes/no)

* **What you see as difficulties for your project**

At this moment I find it difficult to say what I find difficult because this is quit knew stuff for me and I don’t know what to expect exactly and what difficulties I’ll get on my journey. But I can imagine that the Twitter API needs some data cleaning and translations. I’m sometimes too precise, so maybe I’ll be doing too much data cleaning and get stressed because I spent too much time on data cleaning instead of the actual prediction.

References

<https://developer.twitter.com/en/docs/tutorials/how-to-analyze-the-sentiment-of-your-own-tweets>

<https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-recent>