PVA

Mats Fockaert
R0695565
3 ITF BI
Thomas More, Geel
2020

1 INDEX

Contents

2 INTRODUCTION TO THE PROJECT

2.1 Description of the project

Company 'Maat van het Klimaat' wishes to get a more in-depth view around the perception of climate change. By analyzing the correlation between comments on social media and weather history/updates we will try to achieve those wishes.

They also want a way to analyze what people are saying about their services and/or products. An overview of the analysis of the comments will be made and displayed in a dashboard-like manner.

2.2 Research goal

2.2.1 Preface

Online there are thousands of messages being generated every second. Almost all of them have a sentimental value. Some of them are about climate change. This pool of messages could give a better insight in how people think and how people react to climate change during certain weather events.

The goal of this project is to:

- create an API that can interact with the company's in-use system to analyze the sentiment of messages posted about their firm;
- create an overview of all the comments that shows important insights;
- create an analysis between comments on social media and weather reports that may provide a helpful understanding in their possible correlation.

A lot of data will be needed during this project, therefore, research around the data sources will be done with a short list of pros and cons.

2.2.2 API

For this part of the project, the first steps will be made locally. A testing environment will be set up. In the testing environment Twitter data and Reddit data will be used as final testing data, training data will have to be researched. Options for training data are: IMDB data set with sentimental values, Stanford sentiment treebank, 'Bag of Words Meets Bags of Popcorn', others still to be researched.

When the training set is ready, different algorithms will be created and tested to find the most accurate way of doing sentiment analysis.

The most accurate model will be chosen to validate the Twitter data.

The final step is to create an easy-to-use API that works together with the firm's system.

2.2.2.1 Sentimental research's insights

During the steps of the creation of the API, first insights can be made about the testing data. The backbone of those conclusions will then be used to get an accurate depiction of negative and positive tweets about the firm, what words are mostly used, etc.

2.2.2.2 Weather analysis

To perform weather analysis, we need historic data. Research into sources that can provide this data will be the first step.

Finding historic weather data is not free. Therefore, the research is not only confined to historic data, but also to day-to-day data that then will be stored to create historic data.

2.2.3 Database

A database is necessary to save all the data generated by the forecast, social-media API and other data that may be needed during the project.

Because we will be working with incoming and changing data, we will be working with a GraphDB like MongoDB. GraphDB's are a type of databases that handle change in their architecture well.

Eventually we decided to work with S3. The database system of AWS because we will be using Lambda call and S3 works perfectly with this.

3 BUSINESS CASE

This project creates great opportunities and profits for the business. In this chapter they will be discussed.

3.1 Broader view of customers

By doing sentimental analysis, we can classify customers into different groups and learn more about what they want and how they want it. This is done in a very efficient and profound way.

3.2 Insights in climate change awareness

Since the goal of the company is to spread awareness of climate change, we need to know how people think about climate change during certain events in time. Hence, insights into the awareness will bring opportunities to the business. Opportunities to react to certain weather events or others and knowing what the reaction will be, quite accurately.

3.3 More profit

Combining both aspects, previously mentioned, the firm has efficient tools into finding what the customer wants and when it wants it, making it easier and needing less advertising-tools to create profit.

4 PHASING OF PROJECT

Phase 1: Researching data sources

Since we are working with learning algorithms, we need a lot of data to get an accurate overall picture.

Phase 2: Functional and non-functional requirements

After obtaining the necessary information, functional requirements that need to be satisfied for the project, can be made. The requirements will be grouped in usecasediagrams. Separate usecases will provide a better insight in how the API will be used.

The non-functional requirements like performance and accuracy will also be looked in to.

Phase 3: Researching different technologies

When a clear image of the requirements is made, the research in the needed technologies can start. Certain technologies that will be used: Anaconda, Python (with: TweePy, sklearn, textblob, plotly, pipeline from sci-kit), Github and as coding platform, VisualStudio.

How the API will be set-up, the dashboards will be made, the weather will be displayed needs to be researched.

Phase 4: Creating prototypes

A view of all the small representations of all the separate functionalities. This will not be an API but will show all the different aspects of the project and how it will work when it is finished.

Phase 5: Formulating solutions and recommendations

With the accumulated information from previous phases we can formulate an easy-to-use and efficient solution. Furthermore, information on how to use the solution most-efficiently can be advised.

5 RISK ANALYSIS OF PROJECT

With research carried out in the previous chapters, we have identified threats that can be brought by this project and found possible solutions for this.

5.1 Privacy

Problem: the collection of Twitter data and other personal data can result in infringement of the GDPR laws.

Possible solutions:

- splitting the sensitive data from the pool of data that will be in use;
- hashing the sensitive data.

5.2 Budget

Problem: weather stations that can generate weather data are expensive and hard to set-up.

Possible solutions:

- use online historic weather data (cheaper but not free);
- creating own historic data from free-to-use API's.

Most software that will be used will be Open Source Software.

5.3 Ease-of-use

Problem: the implementation needs to be seamless and flexible for own data.

Possible solutions:

- as suggested an API will be created, solving a lot of the problems;
- modelling the testing data with the structure of the firm in mind to create a seamless fit.

6 CONCEPTUAL PHASE

6.1 How analysis will be performed

Using vast amounts of data – originating from social media platforms, weather reports, news articles (possibly), climate data – to first, search for anomalies. Secondly, search for a pattern in the data that can be described or repeated in future events. Thirdly, using known algorithms to determine the emotion within a comment on a social media platform. Finally, with an overall view of all performed analysis, a conclusion about the data can be made.

6.1.1 Functional and non-functional requirements

Functional requirements are the technical requirements that a project needs to satisfy and what a project needs to be able to do.

The non-functional requirements are non-technological items, e.g. price and efficiency.

6.1.1.1 Functional requirements

API request: The user needs to be able to send an API request from its system.
Validate credentials: The API needs to validate the credentials send by the user for security reasons.
Sending information: The API retrieves the requested information and sends it to the user.
Weather dashboard page: A page where all the insights around the links between social media and the weather will be displayed.
Comment dashboard page: A page where all the insights around the comments about the firm will be displayed.

6.1.1.2 Non-functional requirements

API needs to be multi-platform: The data that needs to be accessed needs to work on all platforms.
High accuracy of A.I. models: The firm clearly stated they want an accuracy of over 85% in the A.I. models.

6.2 Data sources

6.2.1 Social media platforms

The objective of this project is finding out how humans react to climate change and how this may affect our business. For this we need data containing human reactions about climate change.
A perfect platform for this is Twitter. Twitter allows its users to set a topic in their comment. This makes it easier to find topics and comments about climate change.
Other platforms are Facebook, LinkedIn and Instagram.

6.2.2 Weather forecasts

There are a lot of open API's to collect data about the weather. Unfortunately, they do not contain free historic data. We will need to build our own database to get an accurate picture of our data.

Other options are climate figures that NASA and other instances collect. They take monthly and/or yearly averages of temperature, emission, sea level, glacier level and precipitation. This can give an accurate depiction of when a drastic change in the climate has happened and what comments then followed on social media.

6.3 Data modelling

6.3.1 Sentimental analysis

When we use lexicons like VADER or TextBlob, we don't need any extra data. No data modelling for these are required.

When we want to train our own model, modelling is required.
The modelling will be done on IMDB's review database. It is easy to use since the database only has two columns.

The chosen algorithm will depend on how much accuracy it produces. During this phase, multiple algorithms are training to find the most accurate one.

When the best algorithm is found and performant, we will load Twitter and Reddit data to make real-life tests.

6.3.2 Correlation social media and weather

Scouring the web to search for historic weather data and creating an application that will collect daily weather data to create our own historic dataset over time.

When the pool is big enough, using deep learning (probably machine learning), relations can be found in the data set of weather updates and when we combine the sentimental analysis of tweets regarding climate change, interesting insights may be made.

6.3.3 Data storage

The project is based in an AWS environment and thus the technologies included will be mostly considered.

Technologies

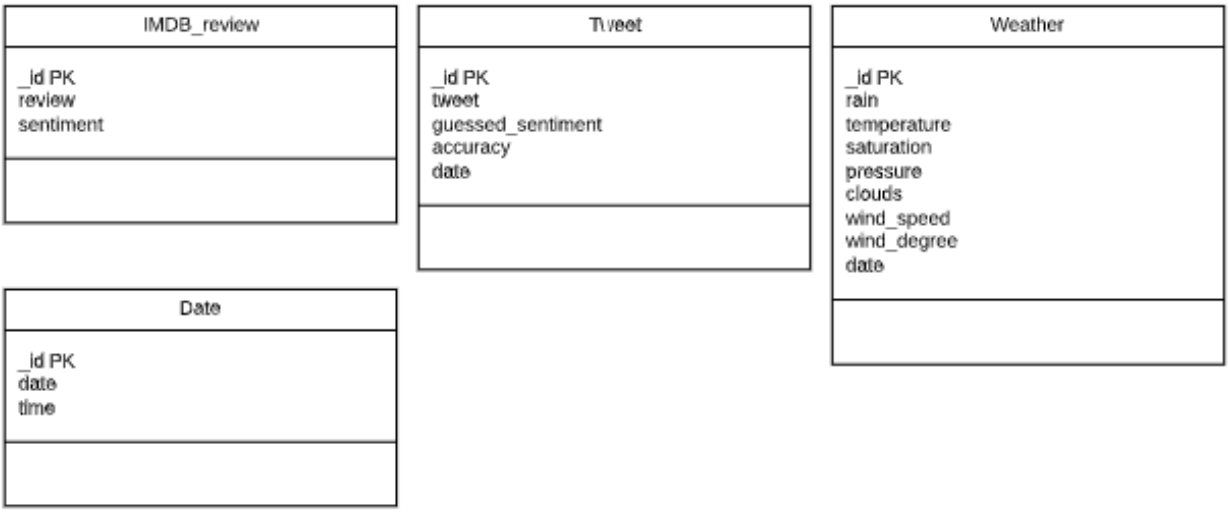
Working with changing data and maybe even changing data models, the choice to use non-relational databases is standard. Previous experiences in MongoDB make this technology stand out.

AWS offers S3. This is an online filesystem that works with folder-like storage. It does not require a strict data model, but it is advised.

More research will happen to decide what the best option will be

Data model

A possible model could be:



A possible key for S3 could be:
Weather-data-a/Raw/2012/05/04/model.file

6.4 Algorithms and models

6.4.1 Sentiment analysis

How does the sentiment analysis work?

There are existing methods to do sentiment analysis. One of them is called 'TextBlob'. It uses a lexicon to calculate if a word is positive or negative. It uses three structures for each word: polarity (negative, positive or neutral), subjectivity and intensity (does it modify next word?). When the word 'very' is analyzed, polarity and subjectivity will be ignored because it is an intensifying word used to modify the following word.

Complete calculations for a phrase with negations:
Phrase: 'not very great':

- 'very': intensity = 1.3
- 'great': polarity = 0.8, subjectivity = 0.75

The rule for negation: Negation multiplies the polarity by -0.5 and doesn't affect subjectivity.
Polarity: -0.307 (- 0.5 * 1/1.3 * 0.8)
Subjectivity: 0.577 (1/1.3 * 0.75)

This is a perfect example of how a computer knows what sentiment a message will try to convey.

Other ways are by using different A.I. techniques (discussed in realization document).

6.4.2 Weather model

Based on the data that will be found, different analysis can be performed. The main things we want to find is what type the weather is.

This way we can make analytical comparisons between what the weather type is and the sentiment of people.

6.4.3 Relation between weather model and social media

The comments will be first filtered so that we only look at the comments about climate, weather phenomena and climate change. The amount of comments, sentiment of the comments, words used … will be analyzed.

Then we can combine the weather phenomena database and the sentiment database to find relations and other interesting insights.

6.5 API

All the previous steps will now be fitted in a RESTful API. This makes it easier for the client to access our services.

Flask is the foundation of the API. The API will hold pickled models that the client can use for researching their own sentimental or weather data.

6.6 Cloud and Cyber Security

AWS will do the heavy lifting for this project. Configuring simple parameters for what is allowed and what our performance should achieve will be set. Tools like:

- AWS Cloudwatch:
  a monitoring service for resources and apps. Collecting metrics, monitor logs, alarms, reacting to changes automatically (with lambda...);
- IAM:
  Identity and Access Management, create and manage AWS users and groups, and use permissions to allow and deny their access to AWS resources;
- Amazon Inspector:
  Automated tool that makes security assessments and produces a report with prioritized steps for remediation;
- AWS Shield:
  managed DDoS protection service that safeguards applications running on AWS;
- elastic load balancers:
  trigger to notify high-latency, overflown, unhealthy EC2 instances;
- elastic IP addresses:
  greater fault-tolerance, mask failures, continues to access app if an instance fails;
- route 53:
  authoritative DNS-service. highest level of availability;
- auto scaling:
  terminate or launch instances, assist with adjusting capacity automatically,

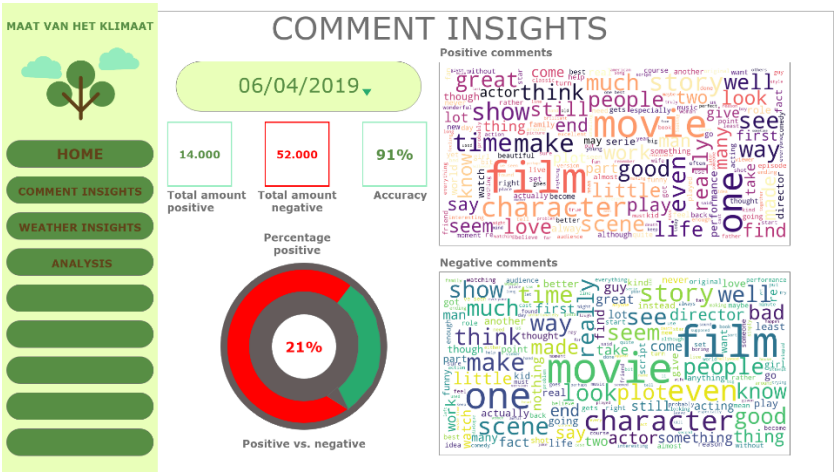
will be used to maintain a surefire and safe way to deliver content to the customer.

6.7 Proof of concept

As a way of displaying the data, three screens have been made. They are a suggestion of how the client may want to display the generated insights and data.

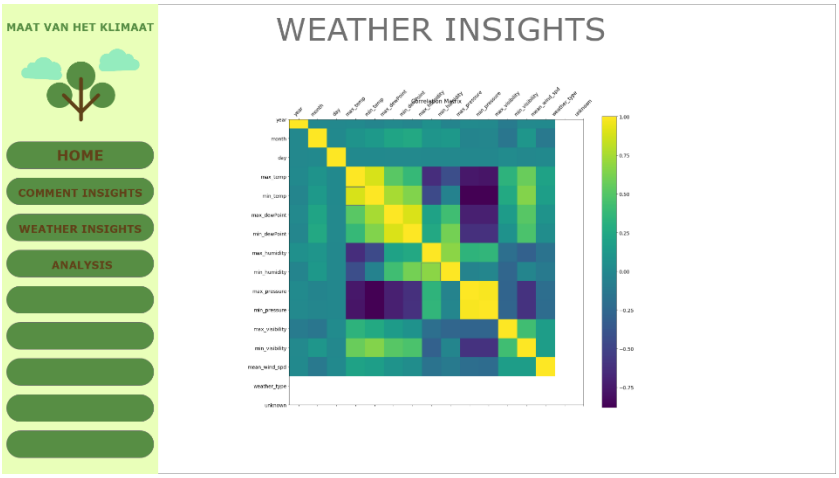
a. Screen 1: Comment Insights

Here we see data generated on a specific day (or month). We can analyze the difference between positive and negative comments on the client's site. We can also see with how much accuracy the model is operating on.

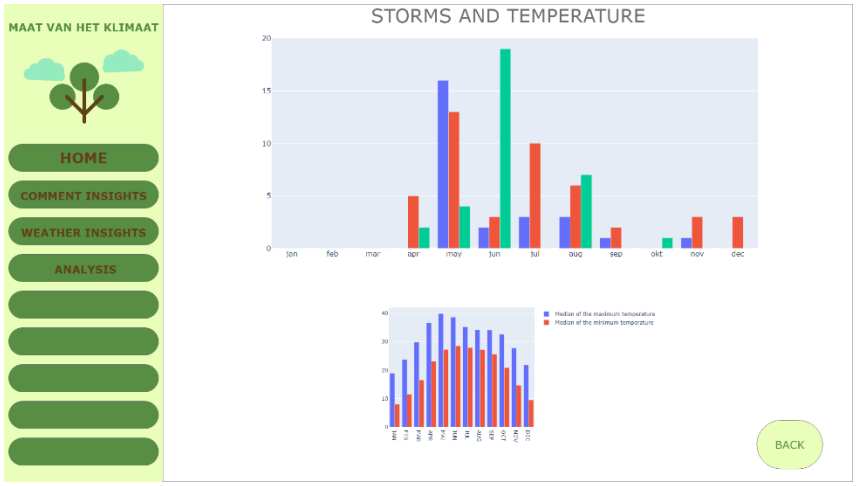


b. Screen 2: Weather Insights

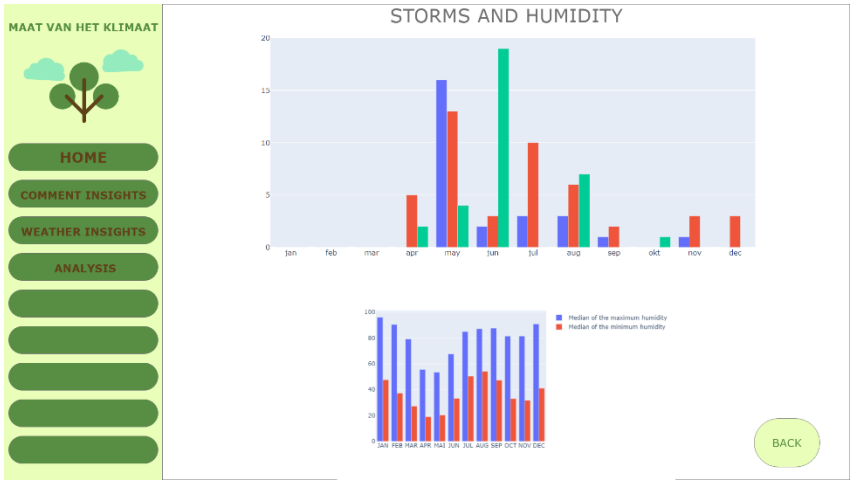
A correlation matrix has been generated. This shows which climate effects affect each other most. When clicking on a cell to view the correlation a new screen will be opened.



i. Screen 2.1: Insights in the correlation between temperature and storms



ii. Screen 2.2: Insights in the correlation between humidity and storms



c. Screen 3: Comment-Weather Analysis

On this dashboard an analysis between comments and weather can be made. Historic data can be used to generate a baseline of weather phenomena and when they change, anomalies can be detected, and comments may be caught. With this information, predictions can be made on how the weather will be and how the comments will follow this weather.