Using sentiment analysis on social media data regarding environmental issues

Marija Nedjalkova

February 16, 2017

Description

Using social media as means of communication with customers and general population has long been proven to be an effective marketing tool as well as a way of receiving feedback about the business in a variety of forms. Social media platforms provide easy to use tools that encourage this dialogue and unite big number of users from diverse backgrounds, thus allowing to form rich data.

In a lot of cases, the feedback flow towards different businesses is huge and it seems impossible to keep track of it manually. Companies need to be able to process the data, make it as useful as possible, summarize it, and respond to the users, sometimes individually.

Natural Language Processing is an evolving area of Artificial Intelligence, and the existing techniques such as sentiment analysis allow to extract the general 'mood' of a text. This can, for example, prove useful when trying to summarize a large amount of data into a numerical grade which can then be analyzed in a variety of ways.

More known Data Mining tools allow to train models and later classify new data, for example, based on its topic or so called aspect.

This project will attempt to help establish a useful dialogue between the Grangemouth refinery in Scotland and the general population in the vicinity of the refinery through Twitter. The known data mining and natural language processing techniques will be used to train a model which should recognize tweets that talk about the refinery (i.e. mention the refinery as a feature) and detect their sentiment. If the sentiment is negative, they should be forwarded to someone in the company. The purpose of this tool is not only to establish a connection with the population for the sake of publicity but also to create another channel that could potentially warn the company about possible leaks or other accidents.

Objectives

The objectives for this project have been divided into two groups: primary and secondary objectives. The tertiary objectives are here in case there is enough time and are not of big importance. The primary objectives are the following:

- Extract the existing data from Twitter.

- Study the existing text classification techniques, compare them and identify those that could be used for the current project.

- Study the existing sentiment analysis techniques, compare them and identify those that could be used for the current project.

- Create and train a model which should recognize negative tweets about Grangemouth.

The secondary objectives are the following:

- Create a mechanism to send the selected tweets to the company as notifications.

- Train another version of a model.

- Analyze and compare the models and their effectiveness.

The tertiary objectives are the following:

- Create a live tool which will receive a stream of tweets and classify them 'on the go'.

- Look into existing studies of detecting irony and sarcasm.

Ethics

Ethical self-assessment has been filled separately.

Resources

Twitter data mentioning Grangemouth will be necessary for this project. Apart from that, it seems that everything that is needed for this project such as NLTK is either installed on the lab computers or can be found online easily.