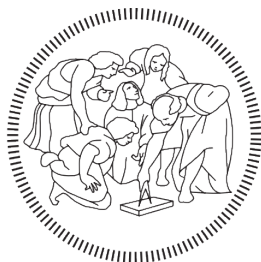


POLITECNICO DI MILANO  
Scuola di Ingegneria Industriale e dell'Informazione  
Corso di Laurea Magistrale in Ingegneria Informatica  
Dipartimento di Elettronica, Informazione e Bioingegneria



**POLITECNICO**  
**MILANO 1863**

A framework for comparing open source sentiment analysis  
APIs

Relatore: Prof.ssa Letizia TANCA  
Correlatori: Prof.ssa Maristella MATERA  
Prof. Riccardo MEDANA

Tesi di laurea di:  
Milica JOVANOVIĆ Matr. 835953  
Mirjam ŠKARICA Matr. 836505

Academic Year 2015–2016



*Some nice inspirational and aspirational quote. Some nice inspirational and aspirational quote.*

*Someone*



# Summary

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.



# Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.





# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Structure . . . . .	1
<b>2</b>	<b>State of the art</b>	<b>3</b>
2.1	Need for sentiment analysis . . . . .	3
2.2	Application of sentiment analysis in various companies and non-profit organizations . . . . .	4
2.3	Most used tools for sentiment analysis . . . . .	5
<b>3</b>	<b>Sentiment analysis workflow</b>	<b>9</b>
3.1	Sentiment prediction workflow . . . . .	10
3.2	Determining real sentiment workflow . . . . .	14
3.3	Evaluation workflow . . . . .	14
<b>4</b>	<b>Framework</b>	<b>15</b>
4.1	Design . . . . .	15
4.2	Implementation . . . . .	15
4.3	User interface . . . . .	15
<b>5</b>	<b>Results</b>	<b>17</b>
<b>6</b>	<b>Conclusion</b>	<b>19</b>
<b>7</b>	<b>Future work</b>	<b>21</b>
7.1	Connecting two mining approaches – Applying clustering in sentiment analysis . . . . .	21
7.2	Spam detection . . . . .	22
7.3	Out of the black box – Training a model . . . . .	23
	<b>Bibliography</b>	<b>23</b>



# List of Figures

3.1	Sentiment analysis workflow . . . . .	9
3.2	Sentiment prediction workflow . . . . .	10
3.3	Determine real sentiment workflow . . . . .	14
4.1	Test caption . . . . .	16



# Chapter 1

## Introduction

For start let us define what is sentiment analysis and why could we potentially depend on it. Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral [Oxford dictionary definition]. How could we in future depend on it? With growth of applying marketing through social media gives us opportunity to seize information about products spread around the media. By analyzing customer's input about a certain campaign, product or brand's strategy a company could predict future trends, decrease costs and increase profit.

Within the described context, this project aims to give statistical comparison of open source APIs used to determine sentiment on small dataset consisting of Facebook posts and related comments. The main focus will be on showing results of the analysis, as well as which is the most efficient sentiment analysis API.

### 1.1 Structure

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

- In the chapter 2 we will present the reasons for doing sentiment analysis within a company or a non-profit organization as well as today's successful solutions and their consequences. We will mention about most used commercial solutions and their strengths and weaknesses. Afterwards we will list some of most

known open source libraries.

- In the chapter 3 is dedicated to describing the workflow of our project. Describing all the steps of the process from collecting the data, analyzing it with different APIs, determining real sentiment, and finally calculating the APIs accuracy.
- In the chapter 4 we will give an overview of the Framework we have built with all its components. Describing the use of Django REST Framework, libraries we have included, etc. NOT FINISHED
- In the chapter 5 is the part where the project results are supported with interpretations. We will present all the statistical results we have obtained after comparing different APIs
- In the chapter 6 will be finalize the purpose of the project and conclude our findings.
- In the chapter 7 we will mention about future improvements such as doing clustering in sentiment analysis, finding a better spam detection solution and eventually training a model that would accommodate our domain problem.

## Chapter 2

# State of the art

This chapter describes state of the art of sentiment analysis in social media. Chapter consists of three sections, each of them trying to bring closer the need for sentiment analysis in current market:

1. Need for sentiment analysis
2. Application of sentiment analysis in various companies and non-profit organizations
3. Most used tools for sentiment analysis

### 2.1 Need for sentiment analysis

With growth of people's interaction and company's advertisements through social media, we have come to the point of realizing that people sharing opinions could help us "predict" stock market and as well follow current trends by guiding the market according to the customers input. Customers nowadays have endless ways to interact with brands which could help increasing brand's awareness, but if not properly analyzed could also lead to obtaining not quite accurate view of customer's satisfaction. The idea of analyzing customer opinion has driven companies to search for an automated way of understanding content that are customers sharing online. The main network for spreading opinions is social media. Almost every tweet, comment, re-share or review gives an information that could guide a company towards better planning, optimizing production and better stock managing. Reason for finding an automated way of analyzing customer's opinion comes from a problem of big data being generated each day which makes impractical to do manually analysis of each user input. Leaving the big data problem aside, brings us to another issue; being able to beat natural language processing challenge. Reason for making the task harder is that user input might be informal, "slang like" content with emojis

, hashtags, full with sarcastic sentences which would lead to unreliable results of analysis.

### 2.2 Application of sentiment analysis in various companies and non-profit organizations

Customers engagement through social media can be a valuable asset for companies to understand level of acceptability of their products. By finding a way to analyze raw text data and catch the key context from it could potentially result in decrease of stock planning cost and in increase of profit. Over the years many companies and nonprofit organizations have started applying some kind of data analysis for this purpose. One of current expanding areas of data analysis is the sentiment analysis, this emerging technique has been applied in various spheres. We will give an insight to few representative examples of successful solutions and their consequences. Cathay Pacific as one of leading airline companies started using one of the commercial solutions for sentiment analysis. They have used Brandwatch platform in order to monitor how their campaign hashtag has been mentioned across social media. Specifically, in which connotation the hashtag has been used, what people talk about when putting it in their posts. Aside the sentiment sensing, they are using the platform to identify trends, such as what are people talking about when traveling and where are they traveling to. This kind of data can help them create new ideas for future developments. Nonprofit organizations like American Cancer Society are using sentiment analysis in order to obtain feedback on organized fund raising event. The difference between this kind of sentiment analysis and the one used by Cathay Pacific is that for ACS a model needs to be trained because words used have another connotation for them. For example, words such as “kill” and “cancer” in case of ACM should have positive sentiment. Sensing on social media has shown that ACM had spikes in user engagement every time there was a big fund-raising event. In order to raise more funds, by data analysis they have realized that doing more announcements before, during and after the event could help them to raise people’s awareness about the problem they are fighting against. The ACS also used sentiment analysis to find out what it should be tweeting and posting. For instance, October is Breast Cancer Awareness month and pink ribbons are everywhere. Some NFL teams even wore pink shoes and gloves to raise awareness.



## 2.3 Most used tools for sentiment analysis

### Commercial solutions

As every commercial product, basic goal is user satisfaction. Commercial solutions provide user with rich customizable, easy to use interfaces for a not so fair price. By paying for the service users, usually medium to large scale companies, receive a platform which contains algorithms for data analysis used as a black box and detailed colorful visualization tools for representing results of the analysis. One of important issues that users wouldn't deal with, as they would if building their own solutions, is that such platforms usually come with needed infrastructure to support such data intense analysis. Here we will mention few most widely used commercial tools.

#### *Google Analytics*

Google Analytics helps you know your audience, find your best content, and optimize ad inventory. Providing you with real-time reports of what is happening on your site right now so you can make adjustments fast. Engagement metrics help you see what is working, while integrations with Google and publisher tools like AdSense, DoubleClick AdExchange, and DoubleClick for Publishers (Analytics 360 only) make it easy to package and sell your ad inventory. Google has developed a solution which enables the user to gather data, preprocess it, and train a model using Google Prediction API like a black box.

#### *Sales Force Marketing Cloud solution - Radian6*

Most certainly that human sentiment analysis is the most accurate method even if you think how much human differ in their interpretation. Radian6 has introduced an automated sentiment analysis tool which has flexibility to allow users to change the perspective of analysis. If you do real sentiment evaluation manually, you will obtain more accurate results than any other automated tool could give you. Given a simple example, if a user compares different beverage brands, most likely he would rate better the beverage he prefers based on the prevailing taste of it. Radian6 solution will enable the user to do deeper analysis into specific topics via different types of ad hoc analysis Radian6 has given various solutions to fill the gap between marketing and customer satisfaction by using social insights to drive marketing campaigns. By listening, engaging and analyzing data on social media, users are able to create sales plans which could lead to better stock planning.

#### *Brandwatch*

Brandwatch Analytics is a web-based platform with monthly subscription basis with different range of packages meeting needs of various company sizes. They search and store data based on users queries on the market. Quite accurately guarantees spam free and duplicate free data. With the gathered data they assure you of optimizing marketing in social media. The platform offers various customizations that could

accommodate to the needs of the user. By acquiring data every day and providing users with tools to analyze and visualize them, they have convinced a lot of famous brands that Brandwatch is a good tool to help them make data-driven market decisions such as Cisco, British Airways and Dell. Good thing about Brandwatch as a commercial solution is that it provides coverage of various data sources, independent of language barrier or data quality. Besides of the coverage advantage, it provides stable analytic tools, as well as visualization tools. It is mostly used by large companies that could afford the platform.

### Open source solutions

Main benefits of adopting an open source solution are lower costs, in this case using an open source library is free, as well as trend of keeping an open source solution always available because it is usually maintained by a community. For a commercial solution, it could happen that a vendor shuts down his business and with it taking its software out of market. Another major advantages of using an open source solution is that often there is a collaboration between libraries and as well as variety of available solutions. Open source solutions are not bind to changes and updates to releases; instead they can be developed collaboratively when functionality is needed. Of course using an open source library can have its down sides, a library or an API call can be limited by number of usages by day or by accepted amount of data it can receive. Thus implicates that these kind of solutions are not setting up an infrastructure that could handle data-intensive analysis and are usually used for educational purposes.

#### *Natural Language Toolkit*

Natural Language Toolkit is an open source platform for building programs which work with textual data. It is equipped with various libraries for text processing which provide tokenization, tagging, stemming and handy wrappers around NLP libraries. NLTK has a very detailed documentation which can guide a developer in building an application that suits his needs. Highly recommended for people that feel free working in Python.

#### *Stanford's CoreNLP*

Provide set of tools written in Java for purpose of natural language processing. Initially was built to work only with English language, but latest releases support languages such as Arabic, Chinese, French, German and Spanish. It is an integrated framework easy to use for language manipulation on raw inputted text. The result it gives after initial analysis is a good starting point for building application with domain-specific problems. Besides low level natural language processing, it contains as well some traces of deep learning algorithms.

#### *Text-Processing*

The text-processing is an open source API which returns simple JSON over HTTP web service for text mining and natural language processing. It is an API which supports speech tagging, chunking, sentiment analysis, phrase extraction and named entity recognition. As an open source solution it has its limitations, such as 1000 calls per day per IP. To get the sentiment of a text, users should do an HTTP request with form encoded data containing text to analyze. As a response, users will receive a JSON object with a label marking the sentiment (can be pos as positive, neg as negative or neutral) and a probability for each label.



## Chapter 3

# Sentiment analysis workflow

This chapter describes the workflow used to analyze the sentiment of social media comments and their corresponding posts. In order to outline the workflow, a top down approach was taken where each subsequent section provides an ever more detailed insight into a particular step of the workflow. The big picture is shown in Figure 3.1 and consists of four parts:

1. Obtaining data
2. Sentiment prediction using an API
3. Determining real sentiment of data
4. Evaluation of that API's performance

First part is the simplest one and as such doesn't merit a more detailed recounting other than mentioning that we were provided with a small sample dataset which, most relevantly, contained about 6000 comments.

In the sections that follow, each of the three remaining parts are broken down into conceptual steps describing the methodology used whilst not cluttering it with too many implementation details. Additionally, it is interesting to note that the first and third steps are done only once. This means that, for each new API we want to use, the workflow for sentiment analysis effectively consists of only steps 2 and 4, namely sentiment prediction and performance evaluation.

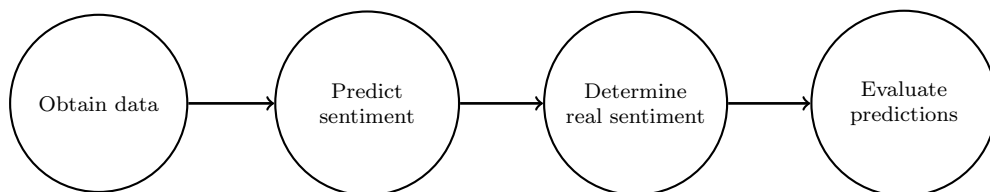


Figure 3.1: Sentiment analysis workflow

### 3.1 Sentiment prediction workflow

Let's assume we have access to an API for sentiment prediction. And by having access we mean being able to programmatically call the API with a text payload and have it return a prediction in some data format. The end goal is to analyze sentiment of all the comments in our sample dataset and aggregate the obtained data on a per post basis in order to infer whether it was positively or negatively received, or even if it had no emotional impact whatsoever. And we want this to be done automatically, practically with a push of a proverbial button. By automatizing the process, it is easy to see how it can derive value for possible future ventures that extend far beyond our modest 6000 comment database.

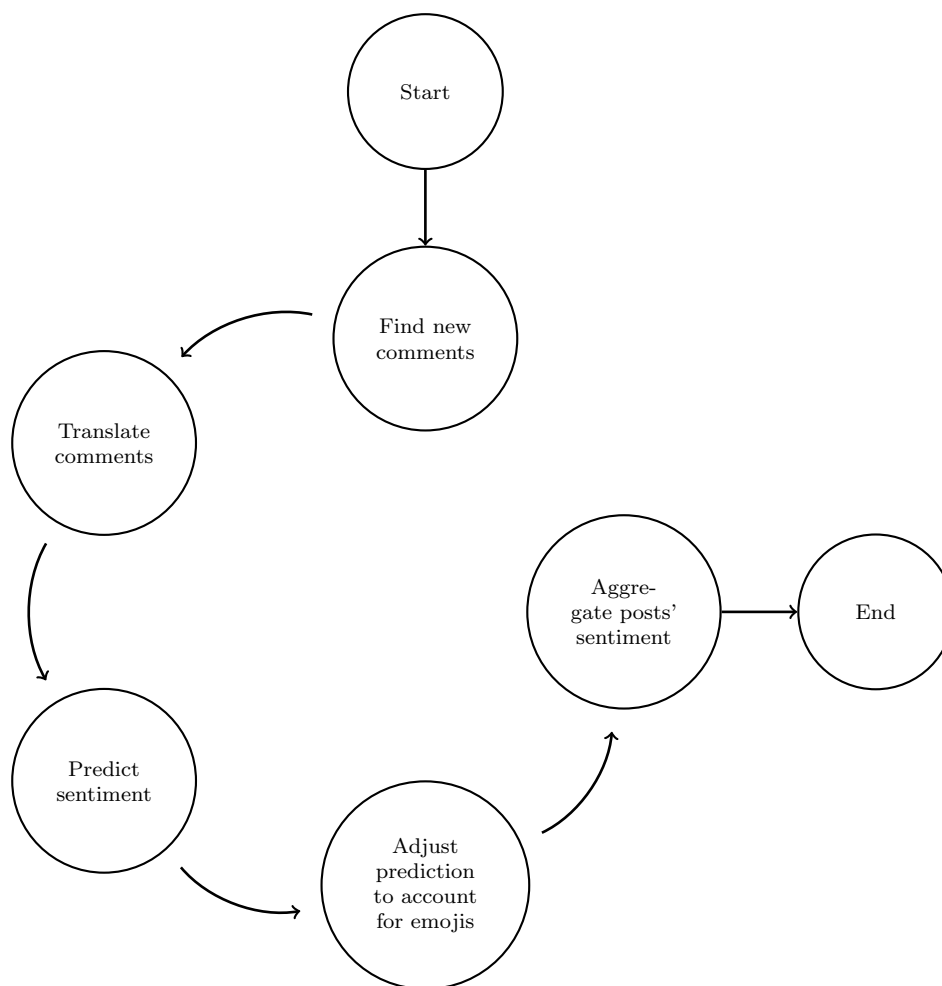


Figure 3.2: Sentiment prediction workflow

Figure 3.2 shows the main concepts that build up the workflow of our sentiment analysis. Since the term *workflow* can be a bit ambiguous, let us clarify exactly what we mean by it. In our case it is simply a python script named named *au-*

*tomated\_sentiment\_analysis.py* that can be run manually, or scheduled to run on a server at desired times/intervals. Sections that follow will explain each step in more detail and will also provide motivation for some, perhaps not so obvious, choices.

### Find new comments

This part is quite straight forward. Once run, the script scans the database looking for comments that don't have a sentiment record attached to it and inserts one. The inserted rows' sentiment columns default to a json shown in Listing 3.1. The reason for this particular choice of json and for using the json format in the first place is discussed at length in Section 4.1. Also, notice the use of the plural form-sentiment columns. This way we are able to store sentiment predictions from each API we planned on using in their own columns.

```
{
  "sentiment_label": "",
  "sentiment_stats": {
    "positive": 0,
    "negative": 0
    "neutral" : 0
  }
}
```

Listing 3.1: Default sentiment json

### Translate comments

To reiterate, our dataset consists of real comments to posts published by actual fashion brands. Since fashion truly is a global industry, the posted comments are in a myriad of different languages. In our case the number of different languages is somewhere north of 70. This provided us with a challenge because most sentiment analysis related APIs handle (well) only content written in English. And the very few that offer support for other languages do so just for a handful of them. This is especially true for the open source variety of APIs that were used for the purposes of this thesis.

Even though the rationale for using comments' English translations seems to hold, we wanted numbers to back up our claims. In other words, we wanted to quantify just how much worse the APIs would perform if we fed them comments in their original language as opposed to English. So for two out of four APIs used, we analyzed both, the content in original language and the English language. The results are examined in Chapter 5, but in short, they are in accordance to what we expected.

This brings us to another caveat. We've just coupled the quality of sentiment predictions with the quality of the translations. After all, the prediction can only be as good as the translation. Since we were trying to evaluate performance across multiple open source APIs, we wanted the best translations possible to try to mitigate this problem. Hence we opted for what we felt was the current industry standard, Google's Translate API<sup>1</sup>. It is worth noting that this is the only step we hadn't taken the open source option but used a free trial period instead to do a one-off translation of our entire dataset.

## Predict sentiment

For each unanalyzed comment we call a specific API requesting a sentiment prediction of the comment's translated content<sup>2</sup>. If no API is specified the script sequentially makes requests to all defined. Since each API's response is in a slightly different format, the response is parsed to adhere to the json definition shown in Listing 3.1. After which, the API's sentiment column for that particular comment is updated with the received (and parsed) values.

## Adjust prediction to account for emojis

In this day and age everybody uses emojis and emoticons, and a lot of it. To disambiguate the two terms, here are the definitions offered by the Oxford dictionary:

**emoji** / ɪ'məʊdʒi /


*origin* (1990s) Japanese, from e=picture + moji=letter, character

*noun* a small digital image or icon used to express an idea or emotion

**emoticon** / ɪ'məʊtɪkən /

*origin* (1990s) blend of words emotion + icon

*noun* a representation of a facial expression such as a smile or frown, formed by various combinations of keyboard characters and used to convey the writer's feelings or intended tone

To put it simpler, the difference is between symbols  and <3. The former being an emoji and the latter being an emoticon. But we digress, the point was to emphasize the very emotional nature and motivation behind using these symbols in a text, comment or post. Having an emoji or an emoticon mixed with text can drastically change our perception of the sentiment behind it. Take these three simple comments:

---

<sup>1</sup><https://cloud.google.com/translate/v2/translating-text-with-rest>

<sup>2</sup>As mentioned in the previous section, there are two APIs for which we requested sentiment predictions in both, their original language and the English translation



```
I read that book
I read that book <3
I read that book ❤️
```

Unless we happen to know the person that wrote the the first comment, its content in plain text doesn't really codify enough information for us to make a judgment call weather or not this person liked or disliked that book. On the other hand, the other two comments are quite unambiguously positive. That one little symbol made all the difference in how we perceive the text that preceded it. Unfortunately, all APIs that we tested would ignore these descriptive symbols, so we decided to write up a very simple algorithm based on the *Emoji Sentiment Ranking*<sup>3</sup> which came to be as a part of the Sentiment of emojis study[?]. The algorithm will be described in more detail in Section 4.2. But in short, the algorithm tweaks the sentiment of comments which contain emojis or emoticons. Then it stores the recalculated result in a separate database table so it doesn't clobber the original data. This allows us to both fine tune our algorithm and to compare the predictions that took the sentimental value of emojis into account to those that didn't.

### Aggregate posts' sentiment

Everything leading up to and including this point was done automatically by running the *automated\_sentiment\_analysis.py* script. Finally, all that is left for the script to do is to aggregate the sentiment data for each post. This boils down to counting how many sentimentally negative, neutral or positive comments does a post have. The results of this data aggregation are stored in a json format as shown in Listing 3.2. Perhaps the most informative field there is the *sentiment\_label*. It is essentially one API's appraisal of how well (or badly) had the public received a published post. Of course, this aggregation is done for each post and API separately. So, for example, according to one API a post might have been overall positively received, while data coming from another API might yield a different conclusion. Sections 3.2 and 3.3 lay out workflows for assessing API's reliability.

```
{
  "sentiment_label": "positive",
  "sentiment_stats": {
    "positive": 38,
    "negative": 2,
    "neutral": 9,
    "total": 49
  }
}
```

Listing 3.2: Example of a post sentiment json

---

<sup>3</sup>[http://kt.ijs.si/data/Emoji\\_sentiment\\_ranking](http://kt.ijs.si/data/Emoji_sentiment_ranking)

## 3.2 Determining real sentiment workflow

In order to answer the question weather or not the obtained sentiment predictions are any good and to determine if any one API outperforms all others- we need sentiment data that we hold true and we need it for each comment. That way we have a real (true) sentiment record to compare against. Since the only state of the art sentiment analyzing machines at our disposal were the two humans writing this thesis, we decided to read all the comments one by one and input our sentiment predictions by hand. Thus, from this point on, when ever we refer to *real sentiment* we mean our own judgment of the sentiment behind the comment. To make this manual process a bit easier for ourselves, we've made it possible to input or modify sentiment for each comment in multiple ways. It can be done via the command line, e.g by doing a curl call to the framework's REST API or via its graphical user interface. But easiest and most efficient way is to run the *update\_real\_sentiment.py* script. The script allows you to specify a range of comments which you want to analyze using command line arguments. The script then sequentially fetches specified comments, prints out their ids, content and English translations and asks for 3 pieces of information as shown in Figure 3.3. It requests a sentiment prediction to be input, weather or not one assesses this comment to be spam and if there was a mention of another user in the comment in question. We were interested to have the two last pieces of intelligence mainly out of curiosity to see how API's would have performed if the dataset was clean from these types of comments, however, they are also a good basis for future extensions of our work.

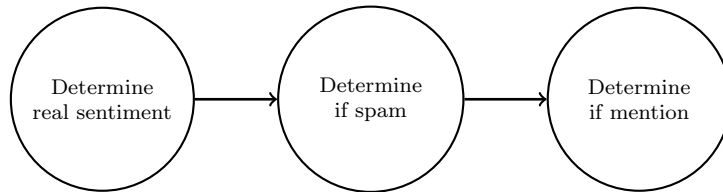


Figure 3.3: Determine real sentiment workflow

## 3.3 Evaluation workflow

Now that we have real sentiment of each comment as well as sentiment predictions, we can evaluate performance of each individual API. Running the *evaluate\_api\_performance.py* will calculate accuracy, precision and recall for each API unless a specific one is specified as an argument.

## Chapter 4

# Framework

### 4.1 Design

Why json when it violated the 1NN rule? already in mysql, will eventually support json, and easily movable to nosql db, or even elastic search.

### 4.2 Implementation

Emoji analysis describe the simple alg <https://github.com/mirjamsk/sentiment-analysis/wiki/Emoji-analysis>

### 4.3 User interface

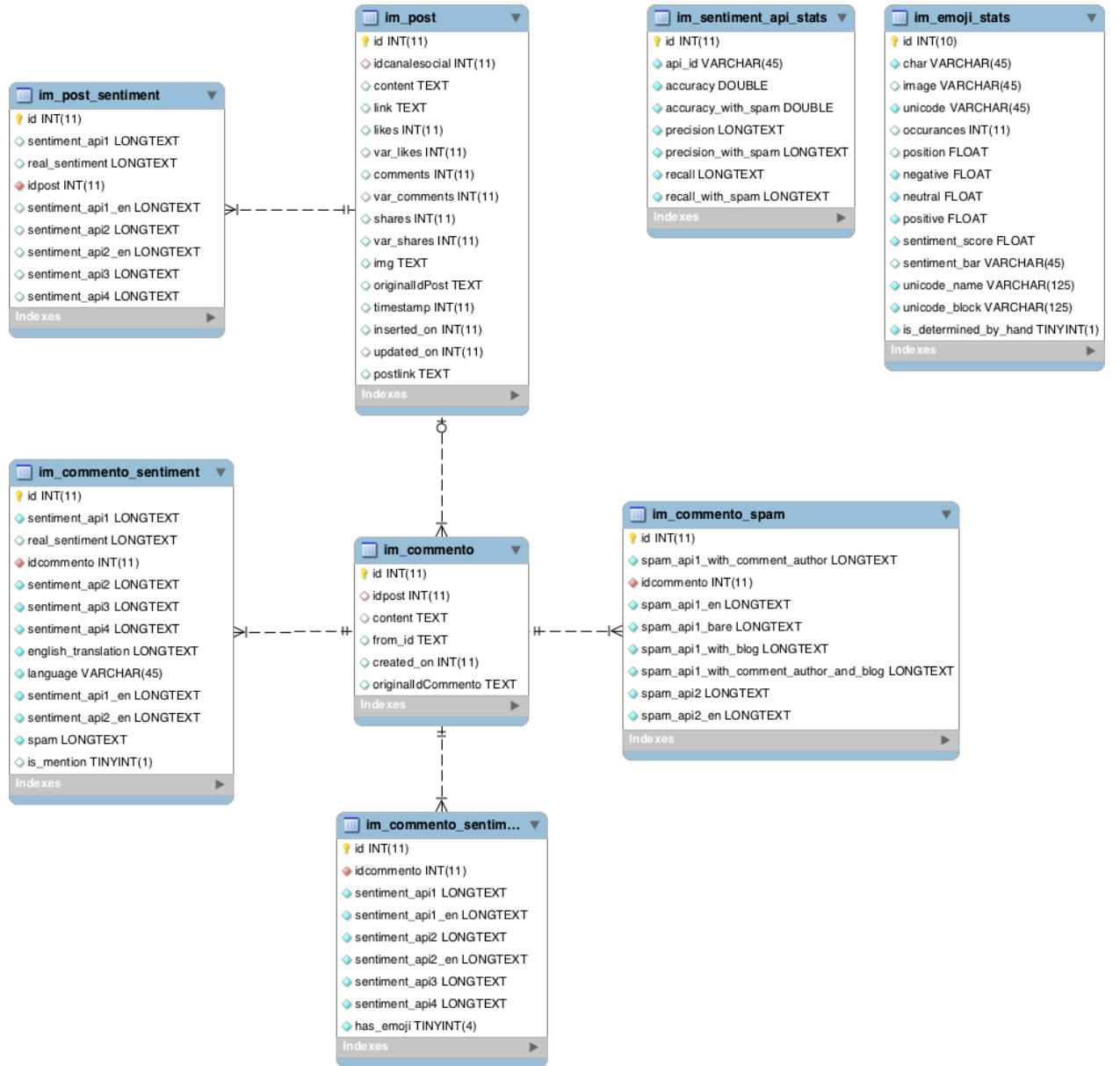


Figure 4.1: TTest caption

## Chapter 5

# Results

For the sake of completeness, below are the definitions used to calculate those 3 metrics. Accuracy is the simplest of all metrics as it is just the fraction of correctly classified comment sentiments.

$$Accuracy = \frac{\text{number of correct prediction}}{\text{total number of comments}}$$

Precision and recall, on the other hand, are a bit more complex to understand and are calculated separately for each sentiment label  $\in \{positive, negative, neutral\}$ . They use concepts such as *True Positive*, *False Positive* and *False Negative*. In the context of our framework, these concepts are calculated as:

**True Positive** correctly predicted labels correctly identified like pairs real sentiment, predicted sentiment positive, positive, negative negative

**One** first item

**Two** second item

**Three** third item

**False Positive** number of times it didn't predict the label when it should have incorrectly identified negative, neutral

**False negative** number of times it predicted the label when it shouldn't have

Recall is the proportion of positive, negative, neutral comments were actually predicted as positive, negative, neutral. In other words, out of all the positive, negative, neutral examples, what fraction did the classifier pick up?

$$Recall = \frac{TP}{TP + FN}$$

Precision is the proportion of labels that were predicted as positive, negative, neutral actually are positive, negative, neutral. In other words, out of all the examples

the classifier labeled as positive,negative, neutral, what fraction were correct?

$$Precision = \frac{TP}{TP + FP}$$

## Chapter 6

# Conclusion

In this final chapter we will summarize all the reasons for involving in such project, as well as the main contributions of the same. After listing the results we will address possible future improvements.

Nowadays with emerging markets and information flow it has become a necessity to try to predict future trends. Thus, companies are processing information luring through Internet with hope they will make a right choice. Big role in company's marketing strategy are social media channels, such as Facebook, Twitter or Instagram. Recognizing the potential use of customers input on the Web, companies have started gathering data related to their online advertisements. Logically, next step was to find a proper way of processing the data in order to discover certain correlations that could guide their production planning. One of recent methods for doing so is called sentiment analysis.

Our report consists of describing current trends in the field of sentiment analysis and how it is applied in business. Showing the reasons for using such method has brought us to idea of investigating about available open source solutions. We have tried to make a comparison with some of the most known sentiment analysis APIs on a given dataset which consists of Facebook comments related to a certain post about fashion industry products. The project itself consists of building a framework representing in a user-friendly way data that has been provided to us with obtained sentiment results of different APIs. We have made a comparison of each API on the data as it is and on data translated to English language. Our statistical results have shown that APIs in general perform better when doing analysis of comments translated in English. Another point where we have seen a potential improvement in our analysis was taking into consideration emojis or emoticons. Currently, emojis have been one of the easiest and most used way of communication. People have seen them as a fast, expressive enough version of typed text. Taking this into consideration we have investigated about finding a proper way to use power of emojis to improve our results. Most obvious solution was building a hash table of emoticons

and their related English translations (for example :) equals happy). Using this hash map to translate the emoticons into their word equivalence we have manage to improve accuracy of APIs.



## Chapter 7

# Future work

### 7.1 Connecting two mining approaches – Applying clustering in sentiment analysis

One of newer ideas in area of sentiment analysis is using clustering algorithms in order to obtain better sentiment analysis results. Let's start with explaining what is clustering. Clustering is a method of splitting datasets into subsets of similar items based on the content of the items. In our case this would be splitting different post sentences in the same "basket" depending on the content, which would result in groups of sentences talking about a similar product feature (product aspect). Clustering is an unsupervised learning method, which means that items are split in separate groups only based on similarity value calculated by its features (in this case the content of the posts). Difference from classification method, where a model will be trained based on "past data", clustering method is based mostly on choosing an appropriate similarity measure. Other difference is that as output of classification your dataset is labeled with a class attribute, and in case of clustering, output is subsets of items.

Applying clustering based sentiment analysis, we might obtain high accuracy results. The process contains few steps:

1. Data gathering
2. Data cleaning
3. Computing the Term Frequency and Inverse Document Frequency
4. Applying K-means clustering algorithm
5. Sentiment Analysis Engine

After the data is gathered, it would be ideally to find an automatic data cleaning method that will remove all the outliers from the dataset. When obtaining relatively

noisy free data we should perform TF-IDF in order to determine "keywords" in the content that could possibly represent a feature of an analyzed product. TF stands for Term Frequency, represents how many times a term occurs in a document. IDF stands for Inverse Document Frequency, and represents how common is a term in all documents. After determining the "keywords", we can have an idea of how many clusters we should expect in our dataset. This could be an input to the K-means algorithm, thus K-means chooses  $k$  random points as initial centroids and assigns all other points to the nearest centroid. Next step of the algorithm is re-centering the current centering. The process repeats until the next iteration produces same result as the previous iteration. Output of clustering method is set of clusters, where each cluster contains similar sentences.

Output of the clustering step represents an input to the sentiment analysis engine. Each cluster is inserted in the engine to determine sentiment of people on a particular feature of the product. Applied to our problem domain, the steps would go like so: On each comment related to post, we would do data preprocessing, removing noisy data from the comment. Afterwards determining keywords in the comment related to the post. For example comment describing customers experience after using the product, such as a customer commenting on the fabric, quality, and color of the blouse. As input in K-mean we would have three aspects of product, in our case  $k$  would be 3. K-means would group all comment sentences related to each of the product features (quality, fabric and color). Afterwards, we apply the sentiment analysis on each cluster and obtain a sentiment value. This way we could calculate the overall sentiment value joining the cluster sentiment values, thus obtaining higher accuracy results.

Downsides of this approach:

1. Applying method on large scale data
2. Eliminating noisy data from social media content should usually involve human interaction
3. Weaknesses of K-means method

## 7.2 Spam detection

As one of the future improvements could be finding a better method for spam detection in order to filter out noisy comments that could give us less accurate sentiment estimation. How could we differ spam from comments that are related to a post? Spam is an unwanted content appearing in the stream of comments. By unwanted content we assume content not related to the post or any other comments of the post. For example, URLs leading to third party web pages used as advertisements or a person that is tagged without any other information about the post. It

is in our favor to try to remove such content from the analysis set which could lead to better sentiment estimation.

We can look at spam detection as one of data preprocessing techniques. Currently we have used Akismet for spam detection; it is a web service for recognizing spam comments. To be able to estimate effectiveness of Akismet service, we have determined manually if a comment is a spam or not. Having a small comment dataset it was not time consuming to manually analyze the comments. After the analysis we have calculated the accuracy of chosen spam detection method. Unfortunately, the results have shown that on our dataset Akismet was not so effective.

### **7.3 Out of the black box – Training a model**

Our project has been based on testing different APIs for sentiment analysis and determining which of the used APIs has given the best results. Using APIs as they are could be imagined as using a tool without knowing what is actually going on inside it. This way we could not tune the algorithm to suit our problem domain. By going out of the box, we could build a model that would provide us with better sentiment analysis results.

Training a model can be seen as producing a function that applied on future data could give sentiment “class label”, in our case positive, negative or neutral value, with higher accuracy than the used APIs. The process of building a model consists of training a model on training set, validating the model and afterwards testing in on the test set.