Internship Project
Report

Internship in the field of Computational Linguistics / Linguistics

submitted by

Sevde Ceylan
B.A. Computational Linguistics
University of Tübingen

VICO Research & Consulting GmbH

Friedrich-List-Strasse 46,
70771 Leinfelden-Echterdingen

Internship duration : 01/2018 - 03/2018

Contents

Chapter 1

VICO Research & Consulting GmbH

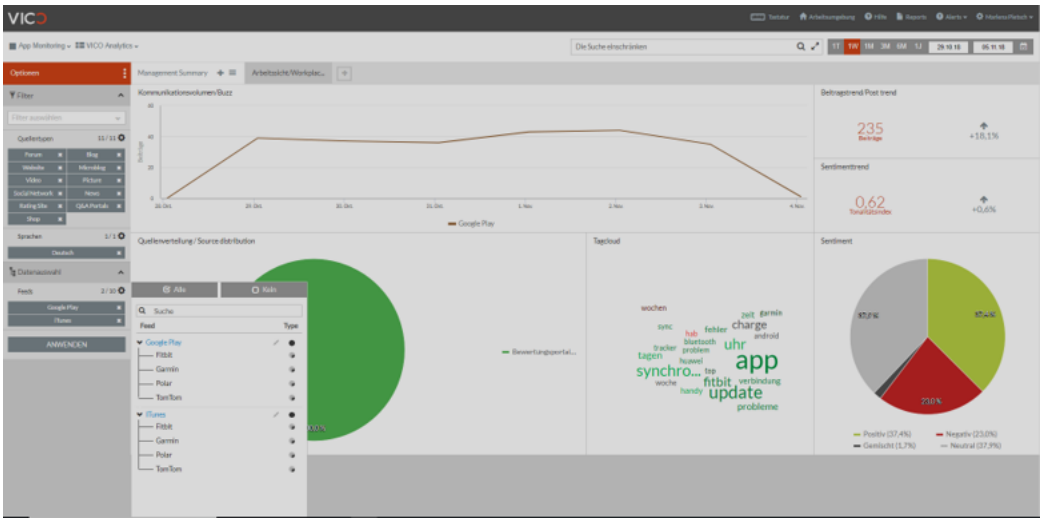
While searching for an opportunity to develop my Computational Linguistics's skills, I found the company 'VICO Research & Consulting GmbH'.
VICO is a company that deals with social media monitoring systems. As service providers for brands or companies, they work on data analysis to give the producer an idea whether its product is consumer-captivating or not. Given that consumers usually share their experiences with a specific product via social media, blog-posts, comments or recessions, data can be collected. This collected data can be analyzed to get information about the user's impression of the product. It's a reflection that the producer gets and that may move him to make changes on the product or to get an approval that everything went right in a best-case scenario.

1.1 About the company

In this small employment of 95 people working, there are several working fields consisting of little teams. Every team has its own importance given that a little work can influence a lot in a small employment. Not only the marketing, influence and IT team were relevant, but also the research team where I had been involved.
The research team consisted of 10 people. Two of them were also Computational Linguists and my supervisors.

1.2 Social Media Monitoring

The daily usage of Social Media becomes more and more a casualness. This is the best resource for researcher in the internet to collect data and make use of the open source media. Especially service providers as VICO, whose slogan is "USE DATA SMARTLY", benefits of this fact as smartly as possible.
VICO developed an own Social Media tool to monitor the gained data. It is the most used program of all teams as it merges all working fields. Image 1 shows a screenshot of this tool consisting of different functions. It is not only helpful to see the most occurring words in a bag of words chart, but also other clustering like time frequency of the posts or sentiment analyses are modelled.



[Image 1] *Screenshot of the VICO tool*

Chapter 2

The Internship

2.1 Gaining data via Query Modelling

For the most part of my internship I modelled keywords semantically with SQL. During the job interview, my supervisors first question was to explain semantic keyword modelling in my own words. With a broad knowledge of Computational Linguistics, this task was easy to solve, here is the answer:
A research for the new shoes of Puma on Google can lead to a lexicon page about the animal puma. As this research wouldn't be successful, such a situation needs to be avoided. But how to avoid ambiguities when searching in the internet ? Exactly this problem was the biggest part of my work - gaining disambiguate data via query modelling with SQL.

2.1.1 Query Modelling

A single person can just click through the Google pages till he gets to his aim. But an automatically modelled tool wouldn't have the capacity to do such an individual search. So such a tool needs a sort of guidance when going through the pages to eliminate irrelevant content. This guidance can be realised by SQL modelling. With the help of modelled queries, useful data can be gained and a data analysis can be started.

In the following I would like to give the details about the rules of the SQL queries:
The program I used for the queries was *NotePad++*. It is a source code editor featuring syntax checking, code folding, scripting etc. [1] Since I already had been using *NotePad++* in the context of my studies, I had no difficulties

[1] Wikipedia, Notepad++: `https://en.wikipedia.org/wiki/Notepad%2B%2B`

with this editor. One of several scripting languages to choose is SQL. With a coloring of each written operator and other syntactical elements, I had a great overview during the work with *NotePad++*. The syntactic construct of a SQL query was logical combinations of terms/phrases and Boolean operators.

- Boolean operators:
  AND, OR, AND NOT

- terms:
  single words and their morphological occurrences
  "gesund" OR "gesundes" OR "gesunde" OR "gesundheit"

- Kleene Star:
  searches for the minimal part of a word
  gesund finds "gesund" OR "gesunde" OR "gesundes" OR "gesunder" ...*
  problem: it would also find blogs about "gesundheitsschädlich" which would be irrelevant - in this case we could eliminate words like that with AND NOT

- phrases:
  composition of terms
  "gesund essen" OR "gesund leben"

- proximity search:
  to fix a distance between single terms in phrases to minimize the searched area
  "gesund sport" 10 OR "gesund sportlich" 10 OR "gesund fitness" 15

The more detailed the request the more relevant contributions can be collected. It also happened that before uploading the final query, we tested the actual query by relevance and changed several terms or excluded specific phrases with the given operators. One has to consider all possible occurrences of a single word in different cases and may has to specify the cases accordingly. For example, when there is a modelling on the new sneakers of the brand name 'Puma', it is not sufficient to write 'Puma' as a term which would end up in data about the animal named 'puma' - as already mentioned in examples below. A detailed phrase or proximity search would help here to eliminate irrelevant data.

2.1.2 Feed Structures

The VICO tool is based on a feed structure. This structure consists of a hierarchy divided into main-feeds and sub-feeds. It is the domain where the scripted SQL modelling has to be implemented to - each query has its own structure place. The main feed includes the query about the topic in general. In a case of research on sport shoes, the main feed would consist of a query about the brand name. Afterwards the sub-feed would specify the main-feed by specialising on the product to be analyzed. This assures an overview for the researcher and saves time when writing queries for one brand name with different products.
Another important fact about the structure is that it wasn't only divided by topic, it was also important to consider the query language. The structure architecture had to consider different query languages as the modellings were not only in German, but also in other languages depending on the searched product. I had the opportunity to write queries in English, German, Turkish, French and even in Arabic.

Technically, the structures had to be created manually by generating a macro for each query. After having generated a macro, the queries could be implemented into the belonging structure.

2.2 Data Analysis with the VICO tool

After having modelled SQL queries and implemented them into the feed structure, the second main task for me was to analyze this data.

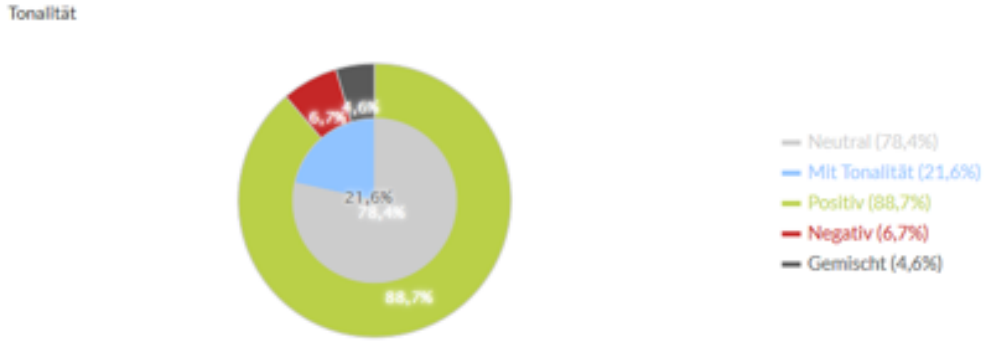
Another part of the VICO tool is the dashboard page - the page for visualizing the gained data (Image1). For a better understanding I can sum up that the SQL modelling and implementation into the feed structure was a background work and the dashboard is the result of this invisible done by researchers. The dashboard can be designed individually, depending on the demand of the clients. The possibilities to demonstrate numeric facts are various. One can visualize the time period showing the amount of data found for each day/month/year, or a bar chart showing the data divided by their origin - social media platforms, such as Facebook, Blogs, Forums, YouTube, Instagram, Twitter, etc. . Of course, there is also a filter that can be used. One can decide which of the feeds to show or on which platforms to search or moreover, specific words such as advertising slogans can be filtered. Another usage of this dashboard is the so called tag-cloud, to be seen in Image

2. It projects the most frequent words. For me the tag-cloud was a sort of guideline to confirm the accuracy of the queries.

[Image 2]*: Tagcloud*

2.2.1 Sentiment Analysis & Opinion Mining

Now we have seen different criteria according to which data is analyzed. One of these criteria concentrates on sentiment analysis and opinion mining. As a result, the content shall be separated by a sentiment which can be positive, negative, mixed or irrelevant. So the task here was to go through the data and make a sort of annotation. This was very time consuming when the data set was big. But knowing that the annotation will succeed in a pie chart showing an emotion analysis of the gained content, such as in Image 3, is motivating.



[Image 3]*: Visualization of a sentiment annotation*

2.2.2 Supervised Machine Learning Algorithm

Supposedly, algorithms for Machine Learning are the most interesting part for Computational Linguists. The time consuming part described below will

one time be performed by an algorithm. VICO researchers implemented an algorithm for the part of Sentiment Analysis Opinion Mining. During my internship the algorithm was going through the collected data-set. It worked in a way that it recognized specific terms indicating one of the sentiments. As for the most implementations of this kind, there is a testing phase as the program has to learn step by step and makes mistakes to be avoided. I was the supervisor for my collected data and found out that the program had mostly problems to detect irony or sarcasm. As a result, I had to correct the false annotations by going through each automatically annotated phrase.

Methods

This algorithm consists of popular formula which I recognized from my study. It was very interesting to get an impression of the practical usage of my theoretical knowledge. Here is a short overview of the basic theories for the Machine Learning Algorithms:

- Naive Bayes
- Maximum Entropy
- Support Vector Machines

Evaluation

My experience with the algorithm was in the way, that I was surprised how qualitative the results were. I did not expect such good results for the Sentiment Analysis when trusting the Machine Learner. But after a comparison of the quality of manually and Machine Learning annotation I could see clearly the deficits of the algorithm. Confronting these two measures gave us the accuracy percentage for manually annotation of 100% and for machine learning of 85%. So the algorithm had still a high quality but not as perfect as the manual one. The reason for this were mostly sarcastic or ironic data. The manually annotated data has the best quality because human can differentiate between sarcasm or irony.
My intuition is that this is a general problem of Computational Linguistics and hard to solve. Under this circumstances the algorithm worked really good and was mostly reliable.

Chapter 3

Conclusion and Acknowledgements

Although I think that monitoring content from Social Media is a very good idea for researchers, there is still need for improvements. In a world full of technologies the queries have to be modelled manually. I had to brain storm about possible word occurrences even though there are programs for that. Afterwards I found out that a Masters student in Computational Linguistics was writing her thesis about exactly this topic. She implemented an algorithm whose output was all possible syntactic cases of an input word. This insight confirmed me that our knowledge from studies can be relevant for the improvisation of the tool used by all employees of a company.

So all in all, this internship was the best opportunity not only to affirm myself that this way of career is the right decision, but also to exchange with other people with the same background of working field. I could collect very useful experiences to develop myself as a Computational Linguist. The fact that my ideas and creativity was also a benefit for the employment and they could make use of it gave me a confirmation of my impressions.

Nevertheless, I also made experiences that were not only positive. Given that VICO is a young employment with an average age of about 28 years, sometimes the seriousness of the work got lost. On the one hand, one can profit a lot from young aged workers, but on the other hand, the term of responsibility was in my opinion too much in shadow. It was not really clear which person was the decider. This kind of work may be also effective but my opinion about this sort of casual hierarchy is different. Anyhow, I tried to utilize this matey ambiance and made really good connections to my colleagues.

Chapter 4

Discussion

In my presentation of the internship, I was asked about the legality of Social Media Monitoring. I have to admit that I did not get into contact with the legal department and therefore could not answer the question properly. But during my internship, I got more awareness about the usage and abuse of the social web. People are strictly trying to share every part of their life what makes it easier to collect data. There could be even found data about dealing with drugs in relation to an analyse of 'Deutsche Bahn'. The accessibility to information about people and their experiences or opinions gets easier. Whether this is positive or negative is hard to answer. On the one hand, sharing experiences or opinions is a way of a collective acting. It offers an exchange of people from all around the world. But on the other hand, it limits the privacy. So the question is not the legacy of social media monitoring. It is more about consciousness of the users and their voluntarily expose.

References

[1] All Images are screenshots from the VICO tool

[2] All numerical facts are published information from VICO