Internship Review

Author

Valentin Deyringer

valentin.deyringer@student.uni-tuebingen.de

Supervisor

Verena Henrich

Company

VICO Research & Consulting

www.vico-research.com

Supervisors

Dr. Stefan Ploch

Dr. Eduardo Torres Schumann

Person of Contact

Nadine Nobis

nadine.nobis@vico-research.com

April 30, 2014

1 The Company

VICO Research & Consulting is one of Germanys leading social media agencies. The Company was founded in 2005 and employs a staff of about 40 at its location in Leinfelden-Echterdingen. The company also offers a good variety of interesting internships for undergraduate and graduate students.

Today VICO is a competent contact partner for all sorts of questions regarding social media marketing.

1.1 Products and Services

The communication on the different channels of the social web represent a vast amount of information which can be of relevance for companies. But this information is not easily accessible. To extract the relevant data, VICO offers a social media monitoring tool and connected services to transform this data into applicable knowledge. These services include market investigation and consulting customers in their social media concepts and measures.

1.1.1 Monitoring Tool

The core product of VICO is a web application. This tool is developed to capture posts from different sources of the social web and automatically rate these posts in terms of different insights they may provide.

There is a huge base of social media sources supplied by default and additional sources may be connected on request. Platforms like the social network Facebook or the microblog community Twitter offer their own APIs to obtain relevant posts. Data from other sources are received from data suppliers or the sources are parsed with VICO's own tools.

The captured posts are sorted in so called feeds resembling different topics. The posts captured for these feeds may then be subdivided in subfeeds, subsubfeeds and so on. To capture relevant posts for each of the feeds, the underlying queries have to be modelled appropriately (see section 2.1.1).

Once the tool is set up according to the customer's specification, the posts of the different feeds are displayed in the overview (see figure 1). Additional information about the communication history, sentiment and much more is shown. An important feature is the tagcloud shown in the centre of the page. This tagcloud can be set to show the most frequent nouns or adjectives in the posts. The search can also be refined with various settings including source type, country of origin or data period.

The shown information can be exported from the tool into excel charts or jpg images and RSS-feeds or email reports may be created.

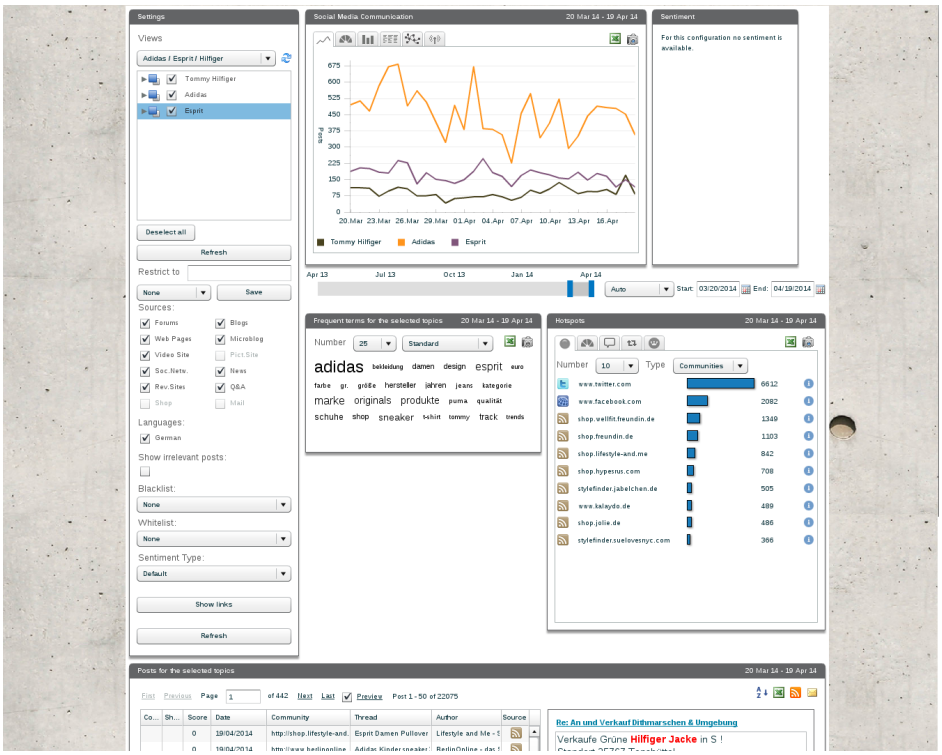

Figure 1: Overview page of the tool showing a monitoring set up for a MA thesis analysing the brands *Adidas*, *Esprit* and *Tommy Hilfiger*
On the left side feeds can be selected and settings for source, language and more can be made. On the top in the middle, the communication development is shown. Left to the communication development the sentiment is displayed if available. In the middle the tag-cloud is shown. On the right the sources with the most communication is shown. On the bottom of the page a list of the found posts is shown.
This is just a small extract of the tools functions.

To assign manual sentiment or categories to the posts, the dashboard offers the coding kit. Here the user can view posts according to certain settings and evaluate their sentiment (see section 2.1.2). The posts can be sorted into existing categories and new categories can be created. It is also possible to assign sentiment values to one post for each of the categories the post belongs to.

Direct interaction via the companies own social media channel is possible with the engagement function. The user can log in with their account and directly answer or comment on posts on their channel or set up postings. These postings are editable in their layout and timing.

The tool has a lot more features and more functionalities are planned which have not been covered in this short overview.

1.1.2 Reporting and Analysing

To obtain the relevant information from the gathered data in an accessible way, customers may get automatically or manually created reports about their reputation in the social web. A summary of the status quo or continuous reports are possible. The data also may help investigating the market relevant to a customer.

Responsible for the manual work within this task is the analysis department. This department employs qualified experts with many years of experience who are supported by continuously trained internees.

1.1.3 Consulting

Another segment is consulting customers regarding their social media appearance and measures. Customers may get help coming up with or improving their social media strategies.

For example an opinion leader campaign can be launched by first identifying the most important authors of the domain and arranging agreements with these authors about the campaign.

2 The Internship

A requirement of my studies of computational linguistics at the *University of Tübingen* is an internship which comprises at least 160 working hours of programming. For this reason I was working in the computational linguistics department of VICO from October 2013 until March 2014. I have become aware of the company by posters around the *Seminar für Sprachwissenschaft* building of the university.

The motivation for me to work for VICO was the interesting field of work of social media. I also liked the friendly behaviour and team spirit I came to know during the application procedure.

2.1 Tasks and Goals

As an intern in the computational linguistics department I was charged with many different kinds of work. My tasks ranged from customer interaction to writing small scripts simplifying the work with queries. Modelling queries, sentiment annotation, coordination of external staff and working on the programming project *gender detection* have been the four main tasks during my internship and are explained in more detail below.

2.1.1 Modelling Queries

To ensure a good quality of the collected social media posts for different topics specified by customers, the queries for data acquisition have to be modelled appropriately. A good balance between precision (quality of the results) and recall (success capturing posts) has to be established. This is the main task of the computational linguistics department and therefore was also a focus of my work.

Queries for the tool are written in an SQL like syntax. The search terms (phrases and/or words) are connected with the operators "OR" "AND" and "AND NOT". Additionally wildcards (*) resembling any string at the end of a word are possible to search for all words starting with the specified string. Phrases can be extended with a proximity operator to search all contained words in a specific range. Expressions can be parenthesized and become quite complex.

Modelling queries is not only a non-trivial task but also is complicated by phenomena like homographs or internet slang.

For example there might be a feed covering all aspects discussed by users about a certain product. The task is now to model a query for a subfeed concerning problems with that product. Searching a keyword like *"problem"* is not enough. The query needs to contain phrases users would use to describe certain problems like *"broken"* or *"doesn't work"*.

It is also important to resolve ambiguities to increase precision. This is mostly realized by using excludes. A toy example might be to get rid of the ambiguity of the word *ball*. One might exclude the word *dancing* (i.e. *ball AND NOT dancing*) to shut out posts dealing with ballroom dancing.

To get familiar with the topic that is modelled is an important first step when writing queries. A good feeling for language, using internet search engines and testing the query by searching on already indexed posts help developing good queries.

It is important to note that the customer's specifications need to be regarded, since the idea of a good result often differs from person to person.

2.1.2 Sentiment Annotation

The dashboard also has the feature of displaying sentiment of posts, which is based on a model created by a machine learning algorithm. To create such models, a number of manually annotated posts is needed. This manual encoding of sentiment was another task during my internship.

There are four different levels of sentiment which are differentiated, these levels are *neutral*, *positive*, *negative* and *mixed*. It is not always easy to evaluate the sentiment of a post since the perception of sentiment is also based on personal interpretations.

For this task the customer's specifications are really important, since the perception of sentiment is quite varied.

2.1.3 Coordination of External Staff

VICO offers their services not only in German and therefore all tasks have to be accomplished for a number of different languages. Since not all languages are covered by internal capacities, external staff has to be acquired and coordinated. During my internship this was also one of my responsibilities.

Contact with external staff has been a part of my daily routine. This included calling them, writing emails and explanatory documents and looking for new external employees on platforms like *proz.com*.

2.1.4 Programming Project *Gender Detection*

To get better insights into who is talking about their brands and products the additional feature of detecting the gender of a posts author is planed for the monitoring tool. As part of my internship, I was charged with preliminary work on this new feature.

To familiarize myself with the topic I read related literature and examined what techniques are used by VICO's competitors.

Next, I became operational and familiar with the existing databases and VICO's libraries written in Java.

Some of the sources make the gender (if given) publicly accessibly. On Facebook one can easily access the authors gender via Facebook's *graph API*. For other sources it may be the case that the gender is represented with a little icon next to the user name. With this information I implemented an algorithm which built a new database connecting the post's content with the author's name and gender.

Since matching names with a gender seemed to be most promising regarding the expected output, I compiled gender specific lists of names. The result of a short test showed that this method resulted in an accuracy of about 89,9%. (Out of 760 assignments 683 were correct)

Feature	Classified Correctly	Classified Wrong
Token Identity	863	137
Emoticons	603	397
Parts of Speech	500	500
Combined (Token Identity, Emoticons and Parts of Speech)	863	137

Table 1: Outcomes of machine learning algorithm for gender detection by features on a test set of 1000 social media posts

This relatively good result was devalued by the fact that the test set contained many real names of the users. On social media platforms like forums, users certainly tend to use nicknames and matching a name list is not practicable here.

So a next step was developing a machine learning algorithm analysing the post's content and classifying the posts on the basis of this analysis. With the help of the frameworks ClearTK[1] and DKPro[2] I implemented such an algorithm and evaluated the results obtained when taking into account different features of a text (see a small output in table 2.1.4). The algorithm used support vector machines, since this model was discussed as the most effective machine learning algorithm in the literature I read. ([3], [4], [2], [1])

Despite the fact of the outcomes looking good, their validity needs to be questioned, since the test set was relatively small and its domain was limited.

The project is a work in progress and there are many options for further work. The features taken into account for the machine learning algorithm clearly need to be examined better. The lists of names could be improved and adapted to fit language specifics. For example the name *Andrea* is female in German, but male in Italian. The importance of meta data of the posts must not be ignored. Clues in users profile could also contribute a lot towards correctly assigning a gender.

If the algorithm will eventually be implemented as new feature of the tool depends on the quality of its outcomes.

[1]ClearTK is developed by the Center for **C**omputational **L**anguage and **E**ducation **R**esea**r**ch (CLEAR) at the University of Colorado at Boulder. It offers tools to develop statistical NLP applications. (`http://code.google.com/p/cleartk/`)

[2]DKPro is a collection of software components for NLP based on Apache UIMA (`http://www.ukp.tu-darmstadt.de/research/current-projects/dkpro/`)

3 Windup

The six months of my internship have been a great experience. Not only did I get interesting insights into machine learning during my work on the project of gender detection but I also learned a lot about social media, marketing and the daily routine of working in a company. All the colleagues I worked with were very kind and helpful and some even became friends. I felt like a part of the team during the whole internship and I am very happy to continue my work for VICO as a *Werkstudent*.

References

[1] Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[2] Na Cheng, Xiaoling Chen, Rajarathnam Chandramouli, and KP Subbalakshmi. Gender identification from e-mails. In *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*, pages 154–158. IEEE, 2009.

[3] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2011.

[4] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.