University of Applied Sciences and Arts Northwestern Switzerland
School of Life Sciences

Development and evaluation of a software system to harvest patient opinions about specific medical implants from social media sources

Yves Noirjean

Supervidsor: Prof. Markus Degen
Expert: Luc Girardin

INTRODUCTION

The increasing use of social media provides access to a wide range of information which was previously hard to collect. For implant manufacturers and other professionals within the field, the experiences and opinions of patients using implants are of great value.

Because patient reports in social media are not meant for automated processing and occur within a bigger set of data, special tools are required to find relevant texts and extract useful information. The software developed is an extensible framework with the aim to approach the task.

CONCEPT

The software collects documents form different social media sources and transforms them into a unified data model. The analysis consists of document filtering, pre-processing, simple methods like TF*IDF, inclusion of a sentiment analysis tool and statistical inspection.

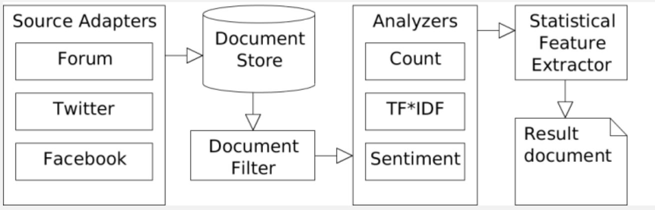

Figure 1: System overview

Sources were manually chosen, and then continuously harvested, converted to a unified data model and written to a document store.

For the Twitter source, a Naive Bayes classifier was trained with 1196 tweets. The tweets were manually tagged as wanted or unwanted where the wanted tweets contained personal experiences of a patient with an implant. After filtering with the classifier, the texts were grouped (e.g. by implant model or manufacturer) and analyzed. The remaining texts were used to form a control group.

In order to trace significant events, time frame processing was implemented. Documents were split into groups based on their creation date and the groups were processed individually.

RESULTS

Results include evaluation of different sources and processing results.

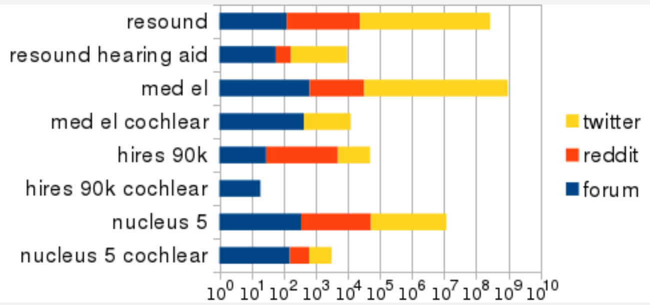

Figure 2: Hit quantity in source systems

Specific products can be very difficult to automatically find. In domain independent sources, results can be refined by adding domain specific key terms, as figure 2 shows. The change in hit quantity is lower in the domain specific forum data.

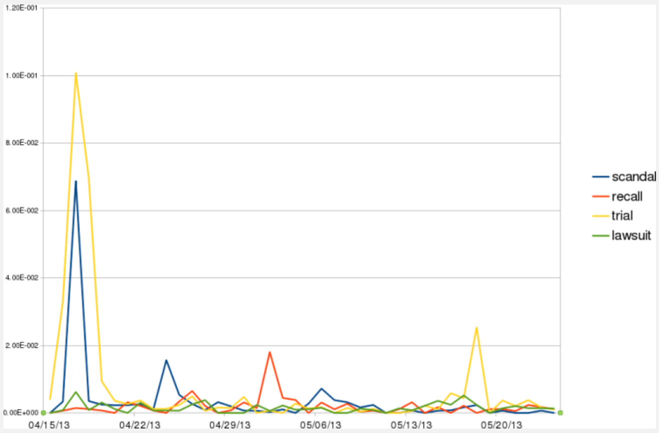

Figure 3: TF*IDF time response in Twitter data

In figure 3, the TF*IDF response to a set of key terms in Tweets is shown. Manual analysis of relevant documents reveal two independent events related to law suits against people associated implant companies. Strong trends in news are visible with little effort.

CONCLUSION

Results show some of the difficulties that occur when processing free form texts. Examples show that, in order to retrieve meaningful results, manual steps are necessary. Strong trends in Twitter data can be seen with relatively little effort.