Analysis of Discussion Forum Data as a Basis for Mentoring Support

Jakub Kuzilek[1,2], Milos Kravcik[1], and Rupali Sinha[1]

[1] German Research Center for Artificial Intelligence, Alt-Moabit 91c, Berlin, Germany
[2] Humboldt University of Berlin, Unter den Linden 6, Berlin, Germany

Abstract. Supporting mentoring processes in higher education is a relevant and challenging aim. Meta-cognitive, emotional and motivational aspects play a crucial role here. Big data can help to recognize the affects of mentees, to react accordingly and to make the mentoring support scalable. In our study, we processed data from university discussion forums utilizing text and sentiment analysis. The results suggest that this approach can raise mentors' awareness of the activities in discussion forums, but limitations need to be considered. Evaluations with real users can help to develop these approaches further.

Keywords: Mentoring · Text analysis · Sentiment analysis.

1 Introduction

Good learning should be individualized and personalized. This goal was already addressed with Intelligent Tutoring Systems, Adaptive and Personalized Learning Environments. These systems mainly aim at the cognitive aspects of the learning process. Intelligent Mentoring Systems (IMS) are going one step further by including metacognitive, emotional and motivational elements in the learning process.

This leads to the following question: What should concepts for designing learning and teaching look like to make the quality of individual mentoring scalable for the acquisition of target competences? Compared to coaching and tutoring, the mentoring process is more spontaneous, more holistic, based on the needs and interests of the mentee and focusing on psychological support. The relationship is more complex, two-way and based on emotions [7]. Here we deal with the question how to help mentors recognize relevant and urgent contributions in discussion forums by means of text and sentiment analysis techniques.

In the following we first very briefly mention selected related work. Then we introduce the analyzed data and the methods applied. In the main part the results are presented and discussed. Finally we summarize the paper and outline next steps.

2 Related Work

Sentiment analysis (SA) aims to analyse people's opinions and emotions from written language. It is widely studied in data mining, Web mining, and text mining to better understand human behaviours [9]. It is usually essential to consider the context of the text and the user preferences [5]. User emotions and intents when contributing to discussion forums can help to elicit their goals [4].

There is a lack of opinion mining systems in non-English languages. Moreover, cross-domain SA is still a significant challenge, including issues like the difference in sentiment vocabularies across different domains and an objective assignment of a strength marker to each sentiment word [6].

3 Data

For this research the data from OPAL discussion forums at the Technical University of Dresden between the years 2005 and 2009 have been employed. The dataset contains 16,614 messages from 123 forums exchanged between 1490 users (students and teachers). Each forum, message and user have a unique identifier.

The data is in anonymised form. Messages contain the plain text with the HTML tags and contain a collection of these emoticons: angel, blushing, confused, cool, devil, grin, kiss, ohoh, sad, smile, tongue, ugly and wink.

The analysis focused on the data from 5 forums containing the highest number of messages. Tab. 1 shows the statistics of the selected forums.

Table 1. Overview of selected OPAL discussion forums.

Forum identifier	Number of users	Number of messages
447053831	80	1085
1012498434	149	938
220528647	28	884
320634883	29	756
436011008	98	697

4 Methods

To uncover information in the data, we applied text mining methods on the selected messages. In the following, the data preprocessing is explained and then each method is introduced.

4.1 Text Preprocessing

The text corpus was preprocessed in the following way:

1. Extraction of emoticons: At the beginning, all emoticons presented in the text as HTML tags "IMG" with class "emoji" were extracted.
2. Removal of HTML formatting: All messages were stripped from the HTML tags to get the clean text messages.
3. Tokenization: All messages were divided into separate words (tokens), keeping the information to which message each word belongs. The *unnest_tokens* algorithm from tidytext R[1] package [8] was used.
4. Stop words removal: From the tokenized corpus German stop words were removed using stop words dictionary[2].
5. Stemming: The remaining words were stemmed, meaning they were reduced to their root form. For example, the words "Abschlusses" and "Abschlüssen" will be reduced to the root form "abschluss". For the stemming, we used Snowball library [2].
6. Removal of tokens with length less than 4: All tokens with the low number of characters representing shortcuts or abbreviations were removed.

The preprocessed data contains 291,151 tokens in the root form. Each word can be mapped back to the original message and user, who created the message.

4.2 Word Frequencies and Document Frequencies

The analysis of word frequencies is the most common way to approach text corpus. The purpose is to uncover the most common words reflecting the text content. At first, the word counts for each forum were analysed by merely counting the number of word occurrences. The analysis showed the most common words in each forum.

To quantify what are the discussion forums about the term frequency - inverse document frequency (tfidf) measure was used. It measures how each word is important to the forum in the collection. The tfidf of word i in the document j is product of two measures: $tfidf_{i,j} = tf_{i,j} * idf_i$ where term frequency $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ is number of word occurrences ($n_{i,j}$) divided by document length ($\sum_k n_{k,j}$) and inverse document frequency $idf_i = log\frac{|D|}{|j:t_i \in D_j|}$ is the logarithm of number of documents ($|D|$) divided by number of documents in which the word is presented ($|j : t_i \in D_j|$).

4.3 Sentiment Analysis

For SA, the SentimentWortschatz sentiment lexicon [3] has been used. It contains approximately 34.000 German words annotated by sentiment value ranging from

[1] https://cran.r-project.org/
[2] https://github.com/stopwords-iso

-1 to 1, representing both negative and positive sentiment. The sentiment of a message is calculated as a sum of the sentiments of the individual message words. Three kinds of sentiment analysis were performed, which will be described in the following sections.

Sentiment Trajectory The sentiment of the messages in chronological order was visualised. The information can be interpreted as sentiment trajectory during the whole forum lifetime. This visual interpretation can uncover the general sentiment trend as well as outliers from the overall sentiment. Outliers are messages "too" positive or negative compared to the others.

Sentiment Wordcloud The Wordcloud visualisation showing the most common words can be used in combination with the sentiment. The "cloud" is divided into two halves. One half of the cloud represents words with a positive attitude, and the second half those with the negative. The size of halves, in this case, is irrelevant. What is important are the terms themselves. They represent the most common negative and positive words in the text.

Correlation of Sentiment and Emojis The last analysis answers the question of whether the emoticons used within the messages somehow correspond to the sentiment of the message. We assigned the sentiment values to the emojis (sentiment value is in brackets): angel (0), blushing (0.4), confused (-0.2), cool (0.8), devil (0), grin (0.8), kiss (0.4), ohoh (-0.8), sad (-0.8), smile (0.4), tongue (0.6), ugly (-0.8) and wink (0.4). Then the emojis and corresponding text sentiment were compared using Pearson's product-moment correlation test [1].

5 Results and Discussion

The previously presented methods have been applied to the data, and the corresponding results are presented within this section.

5.1 Word Count

Fig. 1 presents the results of word count analysis for top 5 forums. Every chart represents the top words used in the discussion forum. We can observe that one of the most used words is "aufgab", which is the root form of the word "Aufgabe", representing the assignment within the course. Other terms such as "frag", "klausur" or "dank" are also standard within the selected discussion forums. Thus one can assume that most of the message content are questions about course assignments and exams. The content is not surprising since that is why forums exist in many educational settings.

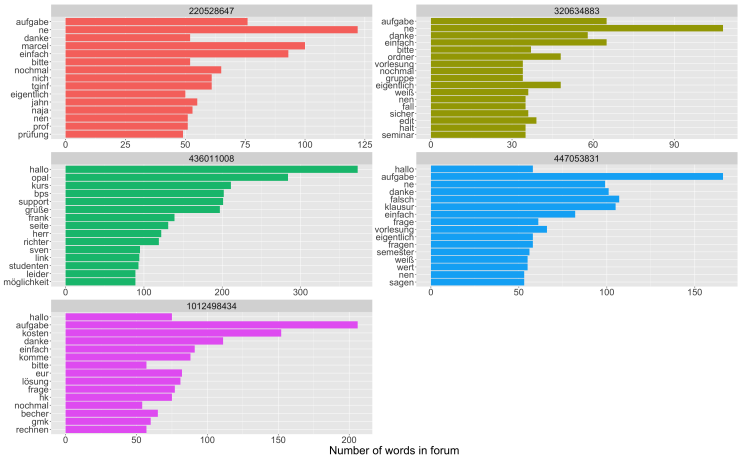

Fig. 1. Word count for the most common words in each discussion forum.

5.2 Term Frequency - Inverse Document Frequency

Fig. 2 shows the most representative words for each forum. One can observe that forums 447053831, 220528647 and 320634883 discuss mathematical issues in their courses. There are words like "hilbert", "algebra", "logit", which are representatives of the mathematical terms. The other two forums cover topics in economics containing the terms like "frank", "gmbh" and "gemeinkost". Based on the analysis of word count and tfidf, one can assume that the forums focus on questions related to the course assessments.

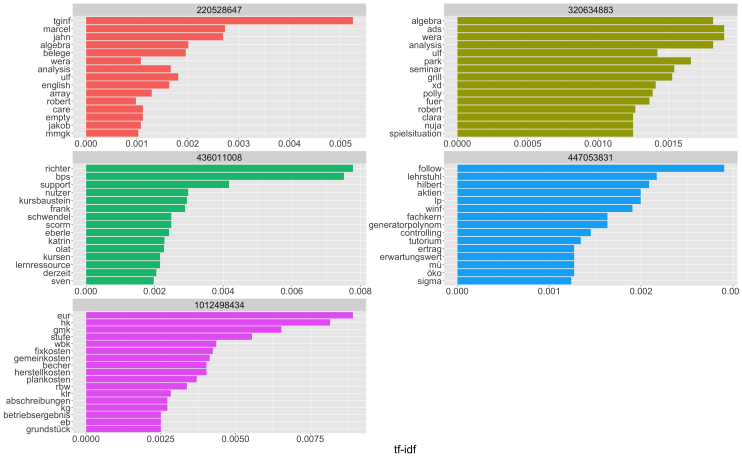

Fig. 2. The analysis of tfidf measure for each discussion forum.

5.3 Sentiment

Our analysis focused on the sentiment within the discussion forums. Fig. 3 shows the sentiment trajectory of each forum. One can observe that most discussion forums tend to be slightly negative, and there are several negative outlier values. For example, forum 447053831 has multiple negative peaks, suggesting that these messages may be worth analysing and answering by a mentor. One can also observe that the trajectories have lowering sentiment values over time, which suggests that in the end, the messages were more urgent. The forum educational focus can explain the overall negativity of the messages. One can expect that the huge portion of words used in communication between students and their mentors will be of neutral sentiment. Still, if the student has difficulty, the sentiment tends to be more negative.

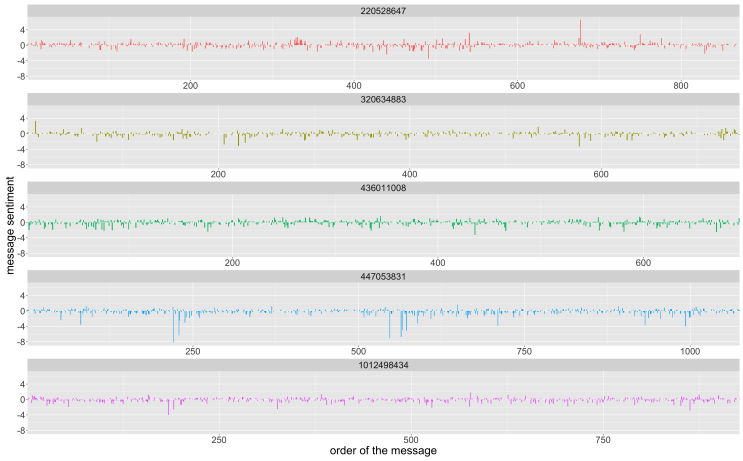

Fig. 3. Sentiment trajectory for each of the selected discussion forums.

Another way to analyse sentiment is sentiment wordcloud, as shown in Fig. 4. One can observe important positive and negative words. The typical representatives of positive words are: "einfach", "genau" and "verstand". These words reflect the understanding of assignments and student's success. Words representing negative sentiments are for example "falsch", "frag" and "nicht".

Finally, we also compute correlation between emojis and sentiment of the corresponding text. The resulting Pearson's product moment correlation test estimates correlation of 0.04 with p-value 4.5 thus we can conclude, that there is only small correlation between sentiment in the text and emoji used.

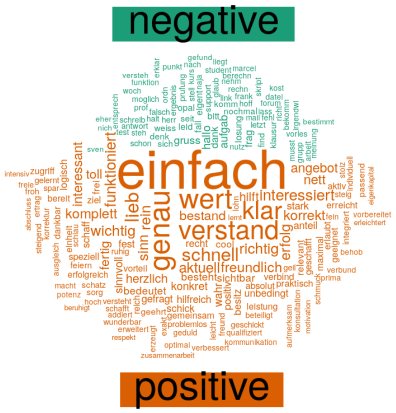

Fig. 4. Sentiment wordcloud for all messages in discussion forums.

6 Conclusions

Mentoring is of crucial importance for university students if provided timely and with good quality. As the capacity of experienced mentors is limited, scalable solutions are needed to make their work efficient. The available technology can analyze the emotions of students from the big data, revealing the learning progress and the need to intervene.

This work in progress deals with the text and sentiment analysis of university discussion forums, which may help mentors to notice critical points where their help is required. As mentioned earlier, there are many challenges in this research area, including a lack of opinion mining systems in non-English languages and cross-domain sentiment analysis. The sentiment vocabularies with their assignments of words are of significant importance.

Using our approach, we made several observations. In one case, the exceptionally high sentiment positivity was caused by the commercial interests of the author. Several other posts in this category were not related to learning, but rather to various celebrations. Apparently, for mentoring purposes, additional pre-processing and fine-tuning would be helpful. On the other side, a post with high sentiment negativity explained privacy and data protection rules. In another post, the indicated negativity was a context issue. E.g. the short text "das Risiko des Verlusts der Paritätsinfos geringer ist" contains three strongly negative words, but it just describes a fact. One critical post deals with the information architecture of an educational web site. So it seems for mentors, especially posts with an increased negative sentiment, are relevant to consider. The sentiment trajectory indicates that the messages after a certain point of time may become more urgent.

Of course, we realize various limitations of this study, in addition to the general challenges mentioned above, regarding the domain and context-dependency.

Also, the tf-idf method does not capture the contextual information, assuming the complete independence among all the words.

Nevertheless, despite these limitations, text and sentiment analysis of discussion forums can undoubtedly help to make the work of mentors more effective and efficient. Even if not every pointed post turns to be urgent, such notifications should be valid in a longer time frame, if proper tools are deployed. To justify their usefulness evaluations with real mentors need to be performed. Relevant functionalities will be integrated in the infrastructure of the tech4comp project (https://tech4comp.de/).

7 Acknowledgments

The project underlying this report is funded by the German Federal Ministry of Education and Research under the funding code 16DHB2102. The presented work was partially inspired by discussions with Cathleen Stuetzer and Ralf Klamma. Responsibility for the content of this publication lies with the authors.

References

1. Best, D.J., Roberts, D.E.: Algorithm as 89: The upper tail probabilities of spearman's rho. Journal of the Royal Statistical Society. Series C (Applied Statistics) **24**(3), 377–379 (1975), http://www.jstor.org/stable/2347111
2. Bouchet-Valat, M.: SnowballC: Snowball Stemmers Based on the C 'libstemmer' UTF-8 Library (2019), https://CRAN.R-project.org/package=SnowballC, r package version 0.6.0
3. Goldhahn, D., Eckart, T., Quasthoff, U.: Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). pp. 759–765. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf
4. Krenge, J., Petrushyna, Z., Kravcik, M., Klamma, R.: Identification of learning goals in forum-based communities (07 2011). https://doi.org/10.1109/ICALT.2011.95
5. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal **5**(4), 1093 – 1113 (2014). https://doi.org/https://doi.org/10.1016/j.asej.2014.04.011, http://www.sciencedirect.com/science/article/pii/S2090447914000550
6. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems **89**, 14 – 46 (2015). https://doi.org/https://doi.org/10.1016/j.knosys.2015.06.015, http://www.sciencedirect.com/science/article/pii/S0950705115002336
7. Risquez, A., Sanchez-Garcia, M.: The jury is still out: Psychoemotional support in peer e-mentoring for transition to university. The Internet and Higher Education **15**(3), 213–221 (2012)
8. Silge, J., Robinson, D.: tidytext: Text mining and analysis using tidy data principles in r. JOSS **1**(3) (2016). https://doi.org/10.21105/joss.00037, http://dx.doi.org/10.21105/joss.00037
9. Zhang, L., Liu, B.: Sentiment Analysis and Opinion Mining, pp. 1152–1161. Springer US, Boston, MA (2017)