Take out what you can: Quantitative Analysis of the Open Question Results from the National Student Survey

Masha Boosten-Ovtchinnikova , Gerard Schouten, Bartosz Paszkowski, Levi van den Bogaard, *Fontys University of Applied Sciences, The Netherlands*

***Abstract* - The main goal of this study was to investigate if a computational analyses of text data from the National Student Survey (NSS) can add value to the existing, manual analysis. The results showed the computational analysis of the texts from the open questions of the NSS contain information which enriches the results of standard quantitative analysis of the NSS.**

I. Introduction

With over 44,000 students and more than 4,500 employees, Fontys is one of the largest universities of applied science (UAS) in the Netherlands. It offers 85 bachelor programs and 22 master programs in varying forms (full-time, part-time, dual), and across almost all sectors of higher education. Fontys strives to achieve excellence in terms of content, organisation and culture. It is a professional educational organisation with high quality staff and organisation, including state-of-the-art facilities (Fontys Focus 2020).

Fontys participates in the annual National Student Survey (NSS) to increase the quality of student study experience. The NSS is a large-scale survey in which nearly all students in accredited Dutch higher education (both academic universities and UAS's) are invited to give an opinion about their education.

The NSS serves two primary goals: to provide prospective students with comparable information about universities and to gather feedback for universities to improve their educational qualities (website Studiekeuze123, https://www.studiekeuze123.nl/nse).

The NSS consists of over 100 items, asking students to rate their satisfaction on a 5-point Likert scale. In consequence, the nature of the results is mostly quantitative.

However, the survey ends with an open question: *"Do you have any wishes or ideas for the improvement of your education, or do you have any other remarks? Your remarks will be forwarded to your institution. They use this information to improve their education"*. Students can reply to this question in a free-format text box. The answers provided are used for quality assurance. To our knowledge this analysis is conducted manually in a labour-intensive and time-consuming process.

The presumption is that the open answers in the NSS contain a potential richness of information that is currently underused. Due to the amount and diversity of the data, the manual analysis, currently not supported by tooling, makes it hard to recognize patterns in a consistent and structured way.

At the same time an opportunity arises due to new developments in cognitive and data sciences backed by the recent technology. The possibilities for an automated computational text analysis could enhance NSS analysis.

The goal of this study is therefore to test the idea that computationally analysing the NSS open answers using a selection of standard text mining methods will increase the value of these answers for educational quality assurance. It is also expected that human effort and time of analysis will decrease.

II. Natural Language Processing

The computational text analyses are based on Natural Language Processing (NLP). NLP is an area of research and application that explores how computers can be used to understand and interpret both the intent and content of text (Josi, 1991). NLP

researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to enable computer programs to perform the desired tasks. The foundations of NLP lie in computer and information sciences, linguistics, mathematics, electrical engineering, artificial intelligence and robotics, and psychology (Chowdhury, 2003).

Applications of NLP include machine translation, natural language text processing and summarization, user interfaces, multilingual and cross-language information retrieval, speech recognition, artificial intelligence, and expert systems (Chowdhury, 2003).

In particular, Louwerse (2011) proposed that language encodes a non-linguistic information and therefore extracting meaning from text is computationally possible.

In their work Fan, Wallage, Rich and Zhang (2006) claimed that technologies in text-mining processing include information extraction, topic tracking, summarization, categorisation, clustering, concept linking, information visualization and question answering help to extract text coherence.

Our study aims to apply these possibilities to a concrete cases study.

III. The Case study

The text data (in Dutch) of several years of Fontys National Student Surveys (2013-2018) was provided to Fontys students of Applied Data Science. This data was anonymised and made available in a temporary Fontys repository with access for students and their supervisors. The access to the repository was protected by a confidentiality agreement.

The goal of this study was translated to an Applied Data Science minor project for a group of 5 students with the following main question: *To what extent can computational analyses of text from the NSS add value to the existing analyses?*

The assignment was further detailed into an applied research covering available tools and methods for topic modelling and sentiment extraction from text. Students were encouraged to follow the applied research framework that Fontys is using (Methoden Toolkit HBO-i)

The outcomes of the assignment were defined as a set of tools that help to process and analyse given corpus of NSS data. The results of the analysis were to include topic and sentiment modelling across multiple years of survey data. Comparing multiple years was necessary to capture and visualize any trends that a human investigator may have missed while analysing the data by hand.

IV. The Automated Data processing

The data analysis steps undertaken were based on theory in text mining, analysis and data visualisation (Ware, 2008; Feldman & Sanger, 2007).

The students started with text cleaning process enabling the use of data by text analysis algorithms. In the data cleaning step stop words and punctuation was removed, all text was brought to a lower case, names and inappropriate language – such as swear words – were deleted.

The next data preparation step contained manual labelling of sentiment as preparation for classification algorithms. About 80% out of all records were labelled leaving 20% for validation purposes. This labor-intensive approach was chosen to increase relevance of analysis and counter sarcasm and off-topic text.

After basic pre-processing functionality was implemented, the students aimed to improve the results by selecting and applying multiple standard machine learning classification algorithms for sentiment analysis and open-source topic analysis methods. Finding any pretrained sentiment models for Dutch language proved to be cumbersome. This

process consisted mainly of drafting models and comparing their results.

The selection of the best classifier was based on their accuracy for sentiment prediction on the testset. The students were encouraged to understand the working of the algorithms to gain a confidence in their findings.

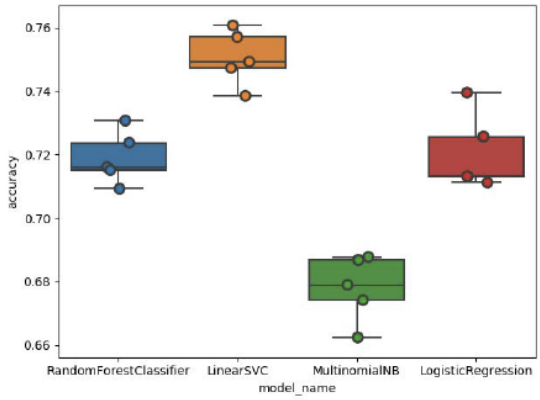

Figure 1: Testset accuracy of standard machine learning classification methods for predicting sentiment.

Based on results available in Figure 1 the students concluded that LinearSVC algorithm is the best for predicting the sentiment of an NNS open answer. Automated hyperparameter tuning and cross validation with 5-folds were further applied to increase the quality of the this model.

Use of open source tooling was stimulated for reproducibility of our study. One of these tools was based on Latent Dirichlet allocation (LDA). LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar (Blei, Ng & Jordan, 2003). For topic modelling the Gensim (Řehůřek, 2011) was used. Gensim is an open-source vector space modelling and topic modelling toolkit implemented in Python.

To complete the task, the prototype of the user interface for the tooling was created in Python. This step provided an integrated and automated text analysis tool with visualisation of sentiment and topics.

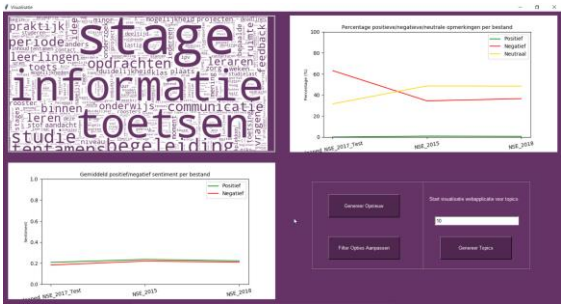

Figure 2: UI for analysis of the NSS open answers.

The prototype was showcased during minor graduation poster event and evaluated briefly for usability by the experts that are performing NSS analysis at Fontys.

V. RESULTS AND FUTURE OUTLOOK

The main goal of this study was to investigate if a computational analyses of text data from the NSS can add value to the existing, manual analysis. By focusing on a proper data pre-processing and automated text comprehension, the results show that automated modelling is possible. It correctly extracts topics which are discussed by students in the open questions of the NSS. Additionally, student sentiment is extracted and dissatisfaction trends across years are made visible.

The importance of a topic for a student is represented as a highest frequency of mentioning it. The key-words associated to that topic are counted towards a topic frequency as well. This frequency is different each year. For example, in 2013 it was *informatie* (information) and *toetsen* (exams) and in 2017 it was *stage* (internship) and *begeleiding* (coaching).

Remarkably, all extracted topics are related to themes defined by the NSS. This indicates that in general students' answers are related to topics of interest for educational institutions.

The extracted list of the words related to the topic is also relevant to this topic. The

visualisation of the relation among the topics further enriches the analysis. Topics which are close to each other have more relations with each other. For example, in 2013 the most important topic consisted of words *stage* (internship), *tijd* (time), *studenten* (students) and *tentamen* (exams), *rooster* (time schedule) have relations with each other and that they describe this topic.

Despite the fact that most of the results require further human expert interpretation, it is indicative to conclude that the computational analysis of the texts from the open questions of the NSS contain information which enriches the results of standard quantitative analysis of the NSS.

Future experiments with tuning of existing algorithms and application of more sophisticated text-mining techniques (LSA, deep-learning, etc.) can extract more specific information from these texts and make the work of analysts more efficient.

Results achieved by e.g., Jurafsky & Martin (2014), Louwerse (2018), Manning & Schutze (1999) are further supporting this belief and showed that this is no science fiction, but science fact.

VI. Acknowledgment

We would like to thank the students of the minor "Applied Data Science" who have worked on this project: Jelle Bouwmans, Alexander Colen, Thomas van Dongen, Bas van Zutphen, Niels van Oijen.

References

1. Chowdhury, G. G. (2003). Natural language processing. Annual review of information science and technology, 37(1), 51-89.
2. Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. Communications of the ACM, 49(9), 76-82
3. Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press
4. Řehůřek, Radim (2011). "Scalability of Semantic Analysis in Natural Language Processing
5. Landauer, T., Dumais, S. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review, 104(2), 211-240.
6. Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. Topics in Cognitive Science, 3, 273–302
7. Kwaliteit volgens Fontys (2012). Fontys Hogescholen.
8. Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). London:: Pearson.
9. Louwerse, M. M. (2018). Knowing the Meaning of a Word by the Linguistic and Perceptual Company It Keeps. *Topics in cognitive science*.
10. Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
11. Joshi, A. K. (1991). Natural language processing. *Science*, *253*(5025), 1242-1249.
12. Ware, C. (2008). Toward a perceptual theory of flow visualization. IEEE *Computer Graphics and Applications*, 28(2), 6-11.
13. Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.
14. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John,

ed. "Latent Dirichlet Allocation". Journal of Machine Learning Research. 3 (4–5):

15. https://onderzoek.hbo-i.nl/index.php/Methoden_Toolkit_HBO-i