Machine Learning Analysis of Problems Encountered by STEM Students from Underrepresented Groups During the Covid-19 Pandemic

Anna Protisina, Nuremberg Institute of Technology, Germany
Beate Neumer, Nuremberg Institute of Technology, Germany
Patricia Brockmann, Nuremberg Institute of Technology, Germany

The Barcelona Conference on Education 2021
Official Conference Proceedings

Abstract
During the current Covid-19 pandemic, STEM students from underrepresented groups have been disproportionately affected. These include women in STEM degree programs, "first generation" students from non-academic families, students with a migration background, students with physical or psychological disabilities and students with children. A control group of university students who do not belong to any of the categories above was defined. This work presents concrete problems reported by students from underrepresented groups as ascertained during interviews. The interviews were first recorded as audio files and then transcribed using speech-recognition software. Transcripts from interviews were analyzed with machine learning methods in an to attempt to identify whether specific patterns of problems were experienced by members of one of the underrepresented groups, or whether the difficulties encountered were uniform across all types of student groups, including the control group. The problems identified in these interviews were compared and contrasted to those previously presented in published literature before the pandemic. These results will be used to define requirements for the design of future digitalization measures to specifically support university students from underrepresented groups.

Keywords: STEM, Underrepresented, Pandemic, Machine Learning

Introduction

During the Covid-19 pandemic, lockdowns and contact minimization regulations were implemented in most countries to slow the spread of the virus. For universities and other educational institutions, this meant a rapid, unplanned introduction of distance learning. Both educators and students had to do the best they could in a difficult situation. Financial and technical resources as well as emotional support and advice available to cope with the increased challenges of distance learning are not equally distributed among all university students.

This work investigates two research questions:
1. Subject-oriented: Did university students in Germany from underrepresented groups in STEM subjects experience changes in their psychological or social situations during the pandemic?
2. Technically-oriented: Can machine learning methods discover text patterns with a limited data set?

First, background information about underrepresented groups is presented. Next, the methods used in this investigation are explained. Experimental results are described and discussed. Finally, conclusions are drawn and plans for future work are presented.

Underrepresented Groups

University students from underrepresented groups already reported feeling disadvantaged before the pandemic (Whitcom, et al., 2021). Underrepresented groups in STEM subjects are defined as

- Women
- First generation
- Migration background
- Parents
- Financially disadvantaged.

Although women represent approximately 50% of the total population, they are still a minority of students in STEM subjects. At the Nuremberg Institute of Technology, only 21% of IT students are female (THN, 2018). First generation students are defined as the first ones in their family to attend a university. Especially in Germany, a person's chances of going to university is highly correlated with the academic level achieved by their parents. 80% of students with academic parents attend university, while only 48% of first generation students do so (Heine 2010). People with a migration background were either born in a different country or one of their parents was. People with a migration. In Germany, although 24% of the population identifies as coming from a migration background, only 11% of university students does so (Berthold, 2018). 6% of university students have one or more children. 41% of them report that they are single parents (Berthold, 2018). Students with disabilities could not be included in this study, because too few respondents self-identified as belonging this group. This would have violated EU data privacy regulations, because individuals could be personally identified. Instead, financially disadvantages students were included as an additional group.

Methods

A number of process steps were conducted to gather and analyze empirical data.

1. Structured interviews were conducted with students from different underrepresented groups and recorded as audio files.
2. These audio files were converted to text using speech recognition software.
3. The text data was pre-processed to normalize capitalization and remove stop words.
4. N-gram terms were analyzed to calculate their significance within the interviews.
5. Machine learning algorithms classified the sentiments associated with n-grams.

23 structured interviews were conducted with students from underrepresented groups. Students self-identified as to whether they belonged to one or more of the groups, as shown in Table 1. Approximately half of the participants identified as female, about half as male. None self-identified as non-binary or diverse. Slightly more than half reported that German was their native language, the other half identified as non-native speakers. Eight of the participants claimed a migration background and eight said they were the first in their family to attend university. Two of the students identified as parents. Five students said they were financially disadvantaged.

Self-Identification	M	F	German Native Speaker	German as 2nd Language	Migration Back-ground	1st Gener-ation	Has Child (-ren)	Financial Difficulties
of Interview Subjects	12	11	12	11	8	8	2	5

Table 1: Interview Subjects

All of the students answered the same questions in the same order. The interviews were record as audio files and uploaded to a secure university server.

Next, these audio files were converted to text characters using the online speech recognition tool Amber script. Each interview had an average of 667 words. In total, approximately 15,000 words were available in total. One inherent weakness encountered here arose during the speech to text conversion. Due to differing accents, the interviews with non-native speakers resulted in a higher number of conversion errors than those conducted with native speakers.

Before the text could be analyzed by the machine learning algorithms, a number of pre-processing steps were required. First, the text was divided up into individual words, a called tokenization. Next, all of the text was converted to lower case, so that a capitalized instance of a word would be recognized as the same word non-capitalized. Stop words, such as “the”, “is” or “a”, are so commonly used that they do not contribute much additional information. These stop words are removed to speed up the analysis. Here, it is important not to remove all three letter words, such as “not”, since this would completely change the meaning of the sentence (Manning, et al., 2008).

Next, so-called “n-grams” were generated. An N-gram is a sequence of “n” words. One example of a 2-gram, or bi-gram, could be “child care”. An example of a 3-gram, or tri-gam, could be “lack child care”. These n-grams are used in Natural Language Processing (NLP) in order to predict the next word in a sequence, based on probabilities calculated from past

examples. The importance or significance of a particular n-gram was calculated using Term Frequency – Inverse Document Frequency (TF-IDF) value. The Term Frequency (TF) is calculated as the number of times a specific n-gram is used in one interview, divided by the total number of words in that interview. The Inverse Document Frequence (IDF) measures how often a certain n-gram appears in all of the interviews. For example, a term such as "pandemic" would probably appear several times in all of the interviews conducted for this study, since the study was conducted during a pandemic, about issues related to the pandemic. As a result, the term "pandemic" would probably not supply much significant additional information. On the other hand, the 3-gram "lack child care" would probably only appear in interviews with student parents and thus be judged by the TF-IDF value to be highly significant (Qaiser, Ali 2018).

After identifying the most significant n-grams, or semantic word groups, the next step was to try to identify which sentiments were associated with each of these terms. Sentiment Analysis is a machine learning method, which attempts to identify the context in which a term is used. Subjective information is extracted to infer the sentiment of a user when using a certain term. The goal is to classify the context in which a term is used in order to predict whether the term is associated with a positive, neutral or negative sentiment. For example, the term "good financial situation" would probably be associated with a positive sentiment, while "financial difficulties" would probably be associated with a negative sentiment. (Pang, Lee 2008).

The sentiment analysis was conducted using a neural network with supervised training. Similar to the way humans learn, supervised learning neural network is trained on a large number of past examples, in order to make predictions for new data. The machine learning tool RapidMiner was utilized to perform this analysis. A neural network needs a large number of past examples (many thousands to millions) in order to be trained correctly. For this project, only a relatively small data set of approximately 15,000 words was available from the interviews. If the training data set is too small, the neural network often fails to produce good predictions, due to a problem called overfitting (Goodfellow 2016). The second research question addressed by this work addresses this aspect: Can machine learning methods discover text patterns with a limited data set?

To improve the performance of the neural network in spite of the small amount of training data, a method called k-fold cross-validation was used, as shown in Fig. 1. Instead of just entering the entire training data set several times in the same order, the training data is partitioned into a number of groups, called folds. The letter k is used to designate the number of partitions, or folds. First, the training data is fed into the neural network in the order: Fold 1, Fold 2, Fold 3, Fold 4. Next, the training data is fed into the neural network in a different order of folds: Fold 2, Fold 3, Fold 4, Fold 1, and so on. This method has been shown to help avoid the problem of overfitting when using small training data sets in supervised learning of neural networks (Berrar 2018).

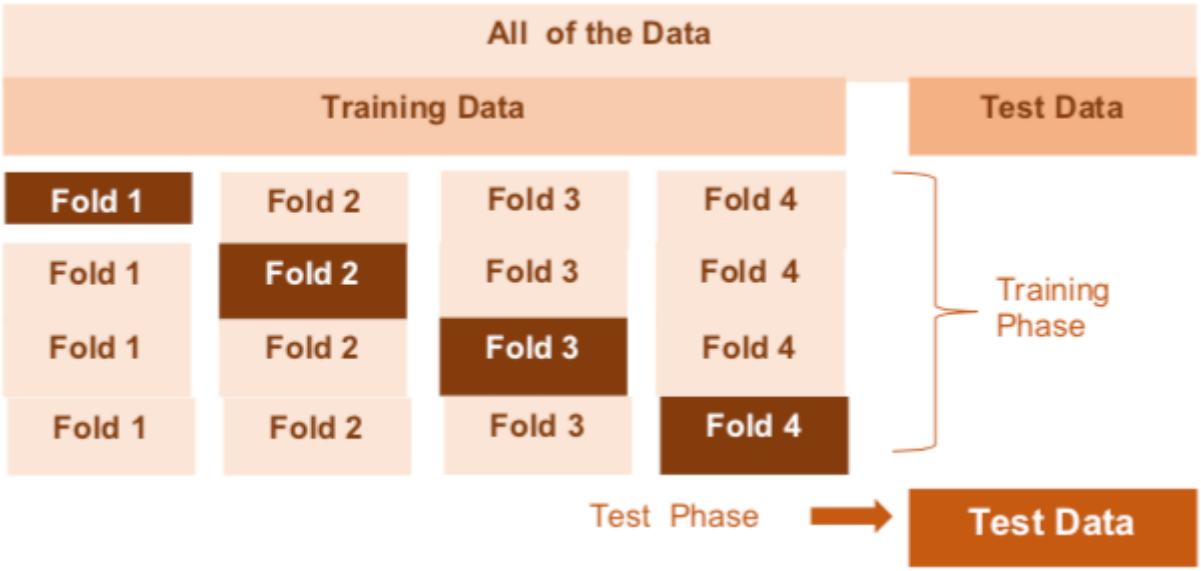

Figure 1: K-Fold Cross Validation

Results

As described in the Section "Methods", after the pre-processing was completed, the first task was to generate n-grams of the text data. The largest words in the word cloud have the highest TD-IDF values and thus judged to have the highest level of significance.

Figure 2 shows a word cloud generated by 1-grams in the text data. "books", "university" and "financial" appear highly significant. Interpretation of this 1-gram word cloud is made more difficult by the high significance of both "good" and "not". Depending on context, each of these 1-grams could completely change the meaning of a sentence.

Figure 2: 1-Gram Word Cloud

Figure 3 shows the bi-grams with the highest TD-IDF score for significance. Here, we see that the bi-grams “financial situation”, “technical difficulties”, “psychological situation” and “changeover presence” are judged to be highly significant. Some terms, such as “cared for family members” were bi-grams in the original German text. The translation of German compound words into English resulted in four words instead of two.

Figure 3: Bi-Gram Word Cloud

Figure 4 shows the tri-grams with the highest TD-IDF scores for significance. Here, we can ascertain quite a number of tri-grams which deliver significant information, such as “thoughts terminate studies”, “distance held younger” and “children home schooling”.

Figure 4: Tri-Gram Word Cloud

After the n-grams with the highest TF-IDF scores were identified, a neural network was used to classify these n-grams according to three different sentiments:

1. Positive
2. Neutral
3. Negative.

Figure 5 shows the results of this classification by the neural network using sentiment analysis. Values to the left of zero signify negative sentiments, values to the right of zero signify positive sentiments. Especially the negative n-grams were classified correctly: “family caused”, “psychological situation” “need time adjust”. A number of neutral sentiments were also classified correctly, such as “situation interim aid”, “relatively managed”. Most of the positive sentiments were classified correctly, such as “luck financially”, “rather well setup” and “sources obtained”. One glaring classification error occurred for the term “abandon studies”. This should definitely be classified as a negative sentiment, rather than a positive sentiment. This error is probably due to overfitting of the neural network, due to the insufficient size of the training data set.

The machine learning algorithm used in this experiment was able to recognize some of the common problems reported by students from underrepresented groups. Worsening psychological and financial situations, a lack of motivation and difficulty concentrating when studying at home due to the presence of family members or children, feelings of isolation due to lack of contact with peers and difficulties in communication with professors were mentioned most frequently.

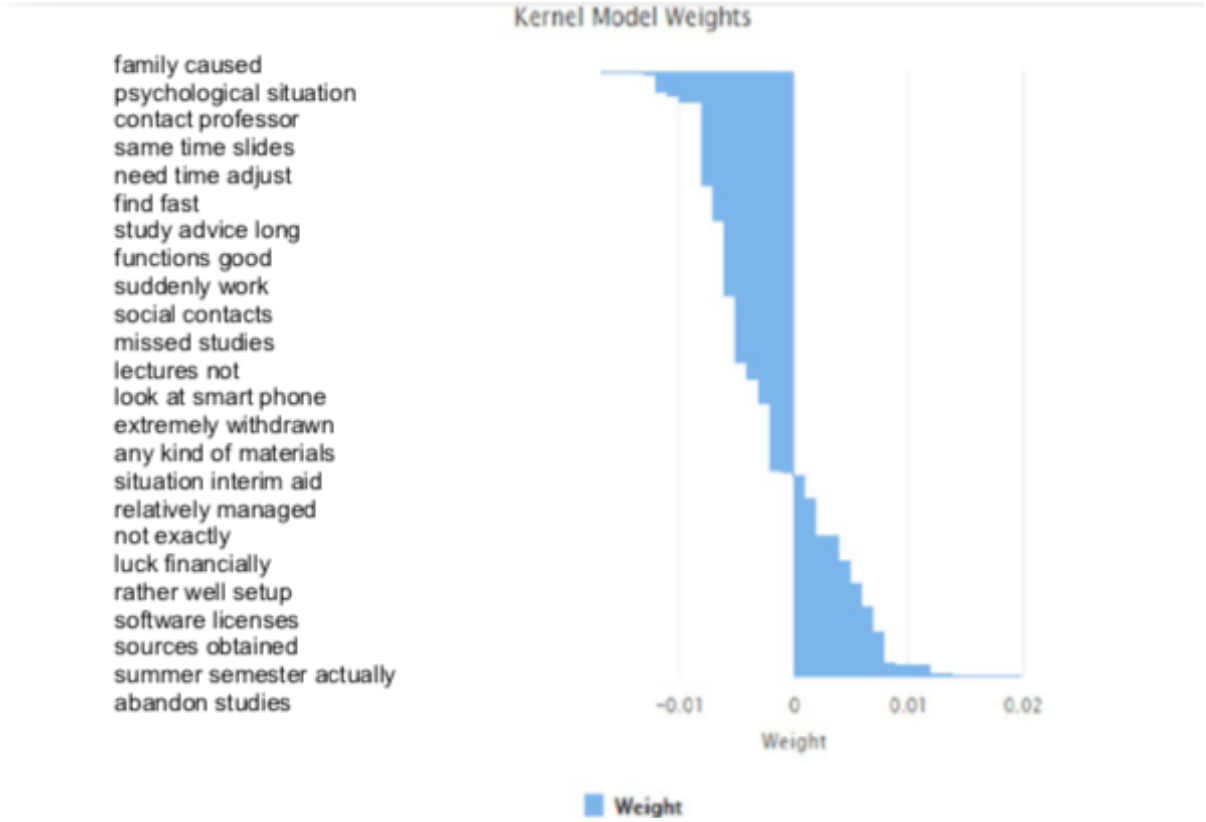

Figure 5: Results of Sentiment Analysis

Positive changes were also noted, such as the rapid digitalization of learning and library materials and excellent support from the computing center. Especially students caring for children and family members appreciated the increased flexibility of distance learning.

Conclusions and Future Work

In conclusion, students from underrepresented groups reported experiencing negative sociological and psychological changes during the pandemic. Machine learning methods, such as neural networks and sentiment analysis can discover unknown patterns in data. Limitations of this study due to the relatively small data set of 23 interviews hindered mapping of correlations between individual underrepresented groups and specific problems. Future work will conduct analysis with larger data sets to increase confidence levels.

Acknowledgements

This work was conducted with support from a research grant from the Nuremberg Institute of Technology in Germany: “Teaching Research – Exploration-Based Learning”.

References

Berrar D. (2018) Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, Volume 1, Elsevier, pp. 542-545.

Berthold, C., Leichsenring, H. (Eds.), Diversity Report, *Studierende mit Migrations-hintergrund*, https://www.checonsult.de/fileadmin/pdf/publikationen /CHE_Diversity_Report_B1.pdf, Retrieved on 10.10.2018.

Goodfellow, I., et al. (2016). *Deep Learning*. Cambridge: MIT Press.

Heine, C. (2010) Soziale Ungleichheiten im Zugang zu Hochschule und Studium: Expertise für die Hans-Böckler-Stiftung, Arbeitspapier, *Demokratische und Soziale Hochschule, No. 213*, Hans-Böckler-Stiftung, Düsseldorf 2010.

Manning, C., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge: University Press.

Pang, B. and Lee, L. (2008). *Opinion Mining and Sentiment Analysis*, Boston: Now Publishers.

Qaiser, S., Ali, R. (2018). Text Mining: Use of TD-IDF to Examine the Relevance of Words to Documents, *International Journal of Computer Applications* (095 -8887), Vol. 181, No. 1, July 2018.

Statista https://de.statista.com/statistik/daten/studie/1236/umfrage/migrationshintergrund-derbevoelkerung-in-deutschland/

Technische Hochschule Nürnberg, Diversity an der Hochschule, https://www.th-nuernberg.de/hochschule-region/strategie-und-profil/hochschule-der-vielfalt/ Retrieved on 12.10.18.

Vernetzungsworkshop: Integration von Studierenden mit Migrationshintergrund an deutschen Hochschulen - Bestandsaufnahme und Vernetzung, BAMF 2011.

Whitcomb, K., Cwik, S., Sing, C. (2021). Not All Disadvantages Are Equal: Racial/Ethnic Minority Students Have Largest Disadvantage Among Demographic Groups in Both STEM and Non-STEM GPA, American Educational Research Association Open, Jan.- Dec. 2021, Vol. 7, No. 1, pp. 1-16.

Contact email: patricia.brockmann@th-nuernberg.de