Natural Language Processing for Content Analysis

An Explorative Study of the Swedish Folk High School

Johannes Fagerlind, Julia Grentzelius, Helena Gustafsson, Marcus Malmberg, Gustaf Norberg, Carlos Palomino Casseres

Abstract

This study examines Swedish folk high school (SFHS) course descriptions from the years 1952-2007, to explore how naturalistic textual data could be analyzed using language technology. The course descriptions were analyzed using natural language processing (NLP) methods, with the aim of providing further insight into how NLP can be used effectively in content analysis. Using the results from these methods, the authors reflect on how the content might reflect changes in society. The following methods were implemented: readability metrics, word clouds, topic modelling, sentiment analysis, and word embeddings. The methods were grouped into shallow analysis and deep analysis methods. The course descriptions were extracted from the raw data, along with their corresponding school. From the school's name, the county and the owner were inferred as well. The extracted data was later pre-processed, before being examined using the aforementioned NLP methods. The study showed that shallow analysis methods are not sufficient for drawing meaningful conclusions on their own, but can provide an overview of the data, which can supplement results from deeper analysis. The deeper analysis gave insights to the course descriptions tone, word use, and topics. We conclude that some of the results stemming from this deep analysis, combined with contextual knowledge, can be seen as a mirror image of happenings in Swedish society. Furthermore, we conclude that there is a crucial necessity for familiarity with the dataset, both in order to pre-process it in a timely manner and to draw credible conclusions from the results of NLP methods.

1. Introduction

Big data commonly refers to datasets containing huge volumes of digital information, whose sheer size makes the processing of the dataset go beyond traditional processing techniques (Mediratta, 2015). A characteristic of the field of computational social science is that it makes use of computational methods to generate patterns and insights into big and complex datasets (Zamith & Lewis, 2015). Lazer et al. (2009) claim that there is a need for novel ways of handling and processing data to access the information stored within, which will give rise to joint efforts between social and computational scientists. While the fields remain separate, joint efforts are needed to propel computational social science (Lazer et al. 2009) with a common ground for realistic expectations and communication.

This paper aims to highlight the possibilities of this joint endeavor as well as to review common tools and avenues within computational analysis of text. We will provide tangible results from using different natural language processing (NLP) tools to analyze a corpus (collection of text data) consisting of historical course descriptions (1952-2007) from Swedish folk high schools (SFHS). In addition to this, we will present ideas on how to analyze and contextualize the data given from the tools. We hope this will help provide a contribution towards bridging the gap between the relevant computational and sociological fields.

The research questions we have used as the foundation of our explorative study have been:

- What information do commonly used NLP methods provide on this type of data?
- Are there differences that can be analyzed between SFHS course descriptions based on differing geographical locations, owners, or years?

By answering these questions, we hope to provide a response to the broader question: Could historical SFHS course descriptions be used as a mirror image of societal changes?

2. Theoretical Background

Content analysis

Bengtsson (2016) states that content analysis can be performed using both qualitative and quantitative measures, and that the inferences

drawn from the analysis can be either inductive or deductive. Krippendorff (2019) contrasts this claim by suggesting that all (textual) content analysis should be regarded as qualitative even if quantitative measures are utilized, since reading and interpreting processed data are qualitative endeavors themself. In addition to this he states that inductive and deductive inferences are not crucial to content analysis and that it is the abductive inferences which should be premiered.

There is a consensus regarding the purpose of content analysis - to transform data from secondary sources into information (Leetaru, 2012; Krippendorff, 2019) - even though the analytical techniques and methodological aspects differ.

Preparatory stages

Krippendorff (2019) states that the context of the data being analyzed (and the knowledge of the analyst) has a significant impact on the kind of inferences that can be drawn from the data. Leetaru (2012) states that data collection is commonly seen as the easiest part of the analytical process but claims that it is in fact one of the hardest parts, where every choice made heavily impacts the analytical resource. Different acquisitional and preprocessing techniques must be utilized depending on the source of the data. This study uses scanned documents, that have been converted into text using Object Character Recognition (OCR) software. This technique often suffers from "noise" such as typographical errors in words and random characters appearing in the text, resulting in data that could be laden with faulty words.

Jurafsky and Martin (2009) name three common normalization procedures for textual data: tokenization, normalizing word format, and sentence segmentation. Tokenization consists of taking a longer text and dividing the words into separate pieces, tokens (Krippendorff, 2019). Common steps in normalizing the word format are case folding – turning letters to lower case, stemming – removing word-final affixes, and lemmatization – turning different inflections of a word to the same root (Jurafsky & Martin, 2009). Sentence segmentation, marking sentence borders, is an important step in preparing the data for deeper or structural analysis.

Shallow analysis

Quantitative content analysis is a statistical approach in describing content, often reported with occurrence frequencies or percentages (Bengtsson, 2016). Leetaru (2012) points out that word level differences can be used to separate different authors, topics, or time periods among other things. It also enables for documents to be compared to and distinguished from each other. This report has chosen three readability metrics and one-word visualization method, to help understand and analyze the data on a shallow level. The readability metrics give information about vocabulary load (OVIX), information density (Nominal Ratio), and overall readability (LIX). The word visualization that this report used was word clouds, which is a specific way of organizing and visualizing words so that the most frequent words are enhanced and made prominent (Hearst et al., 2020). It can be a useful tool for text analysis research, although it is best suited for primary analysis that provides direction, rather than for in-depth analysis (McNaught & Lam, 2014).

Deep analysis

When performing a latent analysis on text data, the analyst is trying to find the underlying meaning of the text (Bengtsson, 2016). This often leads to textual representations of previously unobservable data. As for our deep analysis we choose topic modelling, sentiment analysis and word embeddings. Topic modelling is an unsupervised machine learning technique that builds on statistical measures to extract topics and hidden semantic structures that may otherwise be hidden when manually processing texts. Sentiment analysis extracts polarities in opinion through categorizing text as either positive, negative, or neutral. The method allows the researchers to get an overview of the tones of the texts that it processes. Finally, word embeddings are mathematical representations of words (Camacho-Collados & Pilehvar, 2018). The method can then extract patterns of how a word is used and which words they appear together with. The results can be plotted to visualize the use and meaning of words.

3. Method & Implementation

The following section in the report will outline methodical choices and discuss the data at hand in terms of its limitations and how it was processed. We will also go into detail on how the analytical natural language processing tools were

implemented for the purpose of answering the research question.

Data extraction

The dataset used in this paper stems from a collection of catalogs containing course descriptions from various folk high schools in Sweden. Each document was published ahead of the academic year, with years ranging from 1952 to 2007 (Nylander, 2019). Since the documents did not solely consist of course descriptions, but also excess information irrelevant to our study, such as addresses, phone numbers, and general information about the various schools, an effective way to extract all course descriptions was needed. Python code was used to extract course descriptions from the data. To perform advanced analysis of the data, individual course descriptions had to be tagged with their respective schools, and the school's owners and geographic location. To evaluate the constructed tagger, a measurement of accuracy was performed. An accuracy measure shows how large fraction of the total number of classifications is correct (Jurafsky & Martin, 2009). The gold standard was manually annotated by examining the actual pictures of the OCR scanned documents, for a set of randomly selected course descriptions from each year. Multiple evaluations were performed, and the tagger was iteratively improved. The final accuracy was measured to 81%.

Preprocessing

When all course descriptions had been extracted, the data was enriched using Sparv. Sparv is a text analysis tool developed by Språkbanken, used to preprocess text. The Sparv features used for this paper were the tokenization, lemmatization, sentence segmentation, part-of-speech tagging, and word senses (Borin et al., 2016). These enrichments were used to further process the data for use in the various models. Using the detected part-of-speech tags, the course descriptions were filtered to only include nouns, verbs, adverbs, and adjectives. The lemmas provided by Sparv were used to create lemmatized versions of the course descriptions to allow analysis of all inflections of a word as one single reoccurring item.

Data

Analytical categories were created from the data depending on the year, geographical county, and owner that each description was tagged with. The first category consists of the year that the description was published. The years that the catalogs were published were grouped together into their respective decades. The second category used in this study is the organizer/owner (sw. huvudman). The owners were tagged with a more abstract characteristic, linking different owners together into groups, which we will refer to as the SFHS owner going forward. Grouping of the religious (Church of Sweden and Independent churches) and non-religious (all other) SFHS owners were categorized for analytical purposes. The last analytical category was based on geographical location of the school. Each school is located within a county (sw. län), and we have grouped the counties into three regions, southern, northern, and middle of Sweden.

Readability metrics

Readability metrics were calculated using StilLett API Service (SAPIS) (Rennes & Jönsson, 2015). Course descriptions from each year was sent to the API and from the response, Nominal ratio, LIX, and OVIX values were saved and plotted over time using the python libraries seaborn (Waskom, 2021).

Word clouds

Word clouds were implemented using the python libraries *wordcloud* (Mueller, 2020) and *matplotlib* (Hunter, 2007). Swedish stop words (excluded words) from the Natural Language Toolkit (NLTK) corpora (Bird et al., 2009) were used, along with a manually created list of stop words.

Topic Modelling

The topic modelling used Latent Dirichlet Allocation (LDA) and was implemented using the python library *Gensim* (Řehůřek & Sojka, 2010). Perplexity and coherence values were calculated for different k-values (number of topics), to determine which models were worth analyzing.

Sentiment Analysis

For the sentiment analysis we used VADER, a lexicon and rule-based tool for implementing sentiment analysis from the NLTK library (Bird et al., 2009). Vader inspects both polarity of a word and its intensity, and then returns a sentiment score for each sentence (Hutto & Gilbert, 2014). The mean sentiment score of all

the sentences for each category of the dataset was plotted.

Word Embeddings

Word embeddings was implemented in python using the gensim library (Řehůřek & Sojka, 2010). One model was created for each category of the data. Vectors of similar words to some keywords for each model were saved and transformed to 2 dimensions using t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten & Hinton, 2008), and plotted with matplotlib (Hunter, 2007).

4. Results

Due to a lack of space, only a small selection of the results will be described below. The full results, complete with a more in-depth explanation and visualizations, are available in the full report and its appendices. The reader is strongly encouraged to read the full report to get a better grasp of the results.

Readability metrics

In the full report, changes in the readability measures over time are represented with graphs. The LIX value begins with a drop from just over 50 to just over 40 between the years 1952 and 1957, before spiking the following year. This peak remains until 1968 when a sharp drop occurs. The LIX value then hovers around 50 – 55 for the remainder of the time. The nominal ratio begins with a noticeable increase in value in the years 1958 and 1965. This peak (at about 12) remained until 1968 before dropping sharply. The value remains stable up until 1975, before dropping and stabilizing at around 4 in the '80s and forward. The OVIX value begins with a spike and returns to its previous value (around 50) between the years 1954 and 1958. From this point onwards, there is a dwindling increase (peaking at 70).

Word clouds

Due to space limitations, no word clouds will be shown, although a few will be described. In the word cloud for the entire dataset, the most prevalent words were: "give", "could", "will", and "study". The word cloud for the 2000s looked very similar. In the word cloud for southern Sweden, "mathematics", "Swedish", and "English" were most prevalent. For both northern and middle Sweden, "opportunity" was amongst the most prevalent and were equally prominent.

Topic modeling

The LDA models isolated topics for all the categories defined in the method section above, as well as for the entire dataset. Many coherent topics were isolated, such as a topic consisting almost entirely of words related to music by the model trained on data from the sobriety movement and a topic related to gender segregation in the model trained on the entire dataset.

Sentiment analysis

The results extracted by sentiment analysis revealed that the sentiment in the course descriptions has grown more positive over time. The course descriptions have a neutral score up until reaching the 1980's, which changes into an increasing positive score with a steeper incline decade after decade.

Word-embeddings

Word embeddings allowed us to view the most semantically similar words to "course", "society", "education", "culture", "will", "achieve", and "work". Models were trained on data belonging to the different categories described in the method sections. In the model for southern Sweden for example, words like "bible", "faith", and "church" were found to be semantically similar to the word "society".

5. Discussion

Shallow analysis

The readability metrics were used in this project with the intention of detecting points of interest (Rourke & Andersson, 2004), or changes in language on a larger scale (Leetaru, 2012). Krippendorff (2019) states that word clouds could be a useful tool to familiarize oneself with the data, given that they visualize and highlight the most frequently used words within the text (Hearst et al., 2020). Therefore, the main effect of the generated word clouds was not a foundation for in-depth analysis, but rather a steppingstone to get to know the dataset. Looking at the readability metrics together with the most frequent words in the different categories (SFHS owner, decade, and region) was interesting, because it gave hints to possible

further paths to take in the deep analysis described below.

Deep analysis

An interesting phenomenon that was observed using heatmap visualization of topic distribution over time was the disappearance of gender segregation within SFHS. The LDA model trained on the full dataset isolated a topic that was seemingly consisting of words related to said segregation as it contained words like "men", "women", and names of typical women-only courses like "weaving", "childcare" etc. This topic was highly prevalent in the fifties but went away during the late sixties and seventies. This phenomenon can be said to reflect changing gender norms in society as men and women came to be treated more like equals.

The cause of the observed incline in sentiment over time can be hypothesized to three underlying reasons. Firstly, the municipal adult education system Komvux was founded in 1968 giving the SFHS competition as it offers somewhat similar courses (Fejes et al., 2018). It can be implied that a higher sentiment is angling the courses in a positive light, as an effect of attempting to market students to apply for the SFHS. Another hypothesis for the sentiment incline could be traced to the change of publisher, which occurred twice during the entire time period (Nylander, 2019). Lastly, we must consider the quality of data which is marginally worse in the first two decades of the corpus. VADER, the lexicon-based tool used for the analysis, needs complete words and sentence structures. We are aware that the course descriptions from earlier years have incomplete or split words. This most likely has impacted how those words got a sentiment assigned through VADER.

When observing the cosine similarities between words in the word embedding model trained on data from the southern region of Sweden, words like "bible", "church", and "faith" appear as some of the closest words to "society". This semantic proximity between the word "society" and words with religious undertones are not present to the same degree in models trained on other parts of the dataset. This might be because large parts of the Swedish south, which is sometimes referred to as "the bible belt", contains a population which is largely Christian (HBL., 2018). A hypothesis like this is of course speculative but if it were to be true it would mean that course descriptions from SFHS reflect Swedish demography.

Conclusion

The purpose of this paper was to examine NLP models in terms of what information they can offer to researchers interested in using a computational approach to content analysis. In the process we wanted to examine whether the SFHS can be used as a mirror to society at large. Our results show that while different techniques have value on their own, they can be used in conjunction to further augment the use of one another. A combination of deductive, abductive and inductive reasoning is both possible and important in order to use learnings from one method in other methods. Approaching the paper in this manner showed us that themes which do in fact align with Swedish society at large, can be seen in the way in which the SFHS has presented itself between the '50s and '00s.

References

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *NursingPlus Open, 2*, 8–14. https://doi-org.e.bibl.liu.se/10.1016/j.npls.2016.01.001

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python.* O'Reilly Media Inc. https://www.nltk.org/book/

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., & Schumacher, A. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure, in *SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University.* https://people.cs.umu.se/johanna/sltc2016/abstracts/SLTC_2016_paper_31.pdf

Camacho-Collados, J., & Pilehvar, M. T. (2018). From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research, 63*, 743-788. http://arxiv.org/abs/1805.04032

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society. 2.* https://doi-org.e.bibl.liu.se/10.1177/2053951715602908

Fejes, A., Olson, M., Rahm, L., Dahlstedt M., & Sandberg, F. (2016) Individualisation in

Swedish adult education and the shaping of neo-liberal subjectivities. *Scandinavian Journal of Educational Research*, doi: 10.1080/00313831.2016.1258666

HBL. (September 2 2028). HBL besökte svenska bibelbältet: Flit, frälsning och flyktingar. https://www.hbl.fi/artikel/91e9324d-c1a2-403e-ae89-a08e2daa6c24

Hearst, M. A., Pedersen, E., Patil, L., Lee, E., Laskowski, P., & Franconeri, S. (2020). An Evaluation of Semantically Grouped Word Cloud Designs. *IEEE Transactions on Visualization and Computer Graphics*, 26(9), 2748–2761. https://doi.org/10.1109/tvcg.2019.2904683

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science Engineering, 9(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 216-225. https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Jurafsky, D. & Martin, J.H. (2009). Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition. (2nd ed.) Upper Saddle River, N.J.: Prentice Hall.

Krippendorff, K. (2019). *Content analysis: an introduction to its methodology.* (4th ed,). Thousand Oaks, California: SAGE.

Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabasi, A.-L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; Jebara, T.; King, G.; Macy, M.; Roy, D. & Van Alstyne, Ml. (2009). Computational social science. *Science 323*(5915). 721-3. https://doi-org.e.bibl.liu.se/10.1126/science.1167742.

Leetaru, K. (2012). *Data mining methods for the content analyst: an introduction to the computational analysis of content.* Routledge. https://ebookcentral.proquest.com/lib/linkoping-ebooks/detail.action?docID=1075229

McNaught, C., & Lam, P. (2014). Using Wordle as a Supplementary Research Tool. The Qualitative Report. https://doi.org/10.46743/2160-3715/2010.1167

Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605.

Mediratta, A. (2015). Big data: terms, definitions, and applications. *EMC²*. https://2masteritezproxy.skillport.com/skillportfe/assetSummaryPage.action?assetid=RW$30681:_ss_book:97487#summary/BOOKS/RW$30681:_ss_book:97487

Mueller, A. (2020). w*ordcloud.WordCloud.* WordCloud for Python documentation. https://amueller.github.io/word_cloud/generated/wordcloud.WordCloud.html

Nylander, E. (2019). Folkhögskoledatabasen. Linköping University Electronic Press. https://doi.org/10.3384/db.fhdb

Okasha, S. (2002). Philosophy of science: a very short introduction. Oxford: Oxford University Press.

Rennes, E., & Jönsson, A. (2015). A Tool for Automatic Simplification of Swedish Texts. *Proceedings of the 20th Nordic Conference of Computational Linguistics (NoDaLiDa-2015)*, Vilnius, Lithuania, 317–320. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-128185

Rourke, L. & Anderson, T. (2004). Validity in quantitative content analysis. *Educational Technology Research and Development. 52*(1), 5-18. https://www-jstor-org.e.bibl.liu.se/stable/30220371

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. doi:10.21105/joss.03021

Zamith, R., & Lewis, S. C. (2015). Content Analysis and the Algorithmic Coder: What Computational Social Science Means for Traditional Modes of Media Analysis. *The Annals of the American Academy of Political and Social Science*, *659*(1), 307–318. https://doi.org/10.1177/0002716215570576