A Diachronic Analysis of Using Sentiment Words in Scandinavian Literary Texts from 1870-1900

Ali Al-Laith, Kirstine Nielsen Degn, Bolette Sandford Pedersen, Daniel Hershcovich, Jens Bjerring-Hansen
University of Copenhagen, Denmark

Diachronic corpora, or collections of texts spanning a significant time period, are useful computational linguistics tools for studying language change and evolution. They can be used to investigate changes in vocabulary [1, 2], grammar [3], and usage patterns over time [4]. Additionally, they can be used to analyze the development of different language varieties, and dialects [5, 6]. They can also be used to understand how language is used in different contexts and how language use changes in response to social, cultural, and historical factors [7, 8, 9, 10]. Other potential applications of diachronic corpora in computational linguistics include the creation of language processing tools and systems that consider the historical context in which a text was produced [11].

To track the cultural development in society through literature analysis, one can study the themes and ideas present in the literature over time and look for trends, and changes [12]. This includes examining shifts in how these themes and ideas are presented and changes in the style and form of literature and subjects addressed. It is also essential to consider the social, political, and economic context in which the literature was produced, as these factors can influence the culture and development of society [13].

There are several ways to track the use of emotional language over time in literature [14, 15]. One method is to conduct a content analysis of the text, in which the frequency of emotional words and phrases is counted [16]. Another approach is to use thematic analysis, which involves examining the themes related to emotions in the text and how they are presented [17, 18]. A third option is to employ sentiment analysis, which uses computational tools to analyze the emotional content of the text through natural language processing algorithms or the use of dictionaries or lexicons of emotional words and phrases [19, 20].

Given the large collection of diachronic literary texts that is currently available, we expect to see variations in the usage of sentiment-bearing words in different time periods and in relation to the shifting discussions and themes over time. In this research, we examine the evolution of sentiment words' use in the MEMO corpus, a collection of almost 900 Danish and Norwegian novels from the latter part of the 19th century [21].

A dynamic BERTopic model is a powerful tool for analyzing the evolution of topics in a col- lection of documents over time. It uses transformers and class-based TF-IDF to identify clusters of words and phrases representing the main topics discussed in the corpus. It also incorporates important words in the topic descriptions for improved interpretability. By tracking the use of sentiment words, the dynamic BERTopic model allows us to gain a deeper understanding of the changes and developments in the discussions over time. To further analyze these patterns, we em- ploy the Danish Sentiment Lexicon (DDS)1 [22, 23] to identify any changes in the use of sentiment words over time.

This research aims to track the evolution of sentiment towards a specific topic over time and the evolution of which words are used to express sentiment. The goal is to understand how public sentiment or attitudes towards the topic have changed, identify trends and patterns in the way the topic is discussed, and provide historical context that helps explain how the topic has been represented.

Keywords— Sentiment Analysis, Sentiment Lexicon, Topic Modeling, Scandinavian Literature, Diachronic Corpora, Danish Text, Norwegian Text

1https://github.com/dsldk/danish-sentiment-lexicon

References

P. Cassotti, P. Basile, M. de Gemmis, and G. Semeraro, "Analysis of lexical semantic changes in corpora with the diachronic engine.," in CLiC-it, 2020.

T. McEnery and A. Wilson, “Diachronic Corpora in the Study of Vocabulary Change: A Review,” Corpus Linguistics and Linguistic Theory, vol. 11, no. 2, pp. 203–226, 2015.
C. Mair, “Tracking ongoing grammatical change and recent diversification in present-day standard english: the complementary role of small and large corpora,” in The changing face of corpus linguistics, pp. 355– 376, Brill, 2006.
I. Renau and R. Nazar, “Automatic extraction of lexical patterns from corpora,” in En EURALEX International Congress: Lexicography and Linguistic Diversity, pp. 823–830, 2016.
A. Karjus, R. A. Blythe, S. Kirby, and K. Smith, “Challenges in detecting evolutionary forces in language change using diachronic corpora,” arXiv preprint arXiv:1811.01275, 2018.
A. Jatowt and K. Duh, “A framework for analyzing semantic change of words across time,” in IEEE/ACM Joint Conference on Digital Libraries, pp. 229–238, IEEE, 2014.
M. Hilpert, “The great temptation: What diachronic corpora do and do not reveal about social change,” Corpora and the Changing Society: Studies in the Evolution of English. Amsterdam and Philadelphia: John Benjamins, pp. 3–27, 2020.
G. M. Alessi and A. Partington, “Modern diachronic corpus-assisted language studies: methodologies fro tracking language change over recent time.,” 2020.
M. Liakata and P. Rayson, “Using Diachronic Corpora to Study Language Change and Evolution: A Review,” Corpus Linguistics and Linguistic Theory, vol. 8, no. 2, pp. 227–250, 2012.
S. Kemmer and E. Zaretsky, “Diachronic Corpus-Based Approaches to the Study of Semantic Change: A Review,” Corpus Linguistics and Linguistic Theory, vol. 11, no. 1, pp. 29–62, 2015.
M. Piotrowski, “Natural language processing for historical texts,” Synthesis lectures on human language technologies, vol. 5, no. 2, pp. 1–157, 2012.
M. L. Jockers and D. Mimno, “Significant themes in 19th-century literature,” Poetics, vol. 41, no. 6, pp. 750–769, 2013.
G. Blix, “The Social Role of Literature in Society,” Acta Universitatis Upsaliensis, vol. 11, no. 3, pp. 53–63, 1986.
J. Petzold and M. Dickinson, “Tracking Emotional Language in Literature over Time: A Corpus-Based Approach,” Corpus Linguistics and Linguistic Theory, vol. 8, no. 2, pp. 267–291, 2012.
T. L. C. Jockers and J. D. Porter, “A Quantitative Approach to Tracking Emotional Language in Literary Texts over Time,” Literary and Linguistic Computing, vol. 29, no. 1, pp. 115–132, 2014.
S. Fiedler and K. Kunz, “The Use of Content Analysis in the Study of Emotional Language in Literary Texts,” Methods of Empirical Linguistics, vol. 21, no. 1, pp. 17–38, 2014.
S. Fiedler and K. Kunz, “Thematic Analysis of Emotional Language in Literary Texts: A Diachronic Corpus-Based Study,” Corpus Linguistics and Linguistic Theory, vol. 5, no. 2, pp. 195–224, 2009.
S. Fiedler and K. Kunz, “Thematic Analysis of Emotional Language in Literary Texts: A Review of Methodological Approaches,” Methods of Empirical Linguistics, vol. 18, no. 1, pp. 1–23, 2012.
M. Mu¨ller and A. Panchenko, “Sentiment Analysis of Emotional Language in Literary Texts: A Compar- ative Study of Machine Learning and Dictionary-Based Approaches,” Corpus Linguistics and Linguistic Theory, vol. 7, no. 2, pp. 227–250, 2011.
S. Fiedler and K. Kunz, “Sentiment Analysis of Emotional Language in Literary Texts: A Review of Methodological Approaches,” Methods of Empirical Linguistics, vol. 21, no. 1, pp. 17–38, 2014.
J. Bjerring-Hansen, R. D. Kristensen-McLachlan, P. Diderichsen, and D. H. Hansen, “Mending fractured texts. a heuristic procedure for correcting OCR data,” 2022.
S. Nimb, S. Olsen, B. S. Pedersen, and T. Troelsg˚ard, “A thesaurus-based sentiment lexicon for danish: The danish sentiment lexicon,” in Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 2826–2832, 2022.
B. S. Pedersen, S. Nimb, and S. Olsen, “Dansk betydningsinventar i et datalingvistisk perspektiv.,” Danske Studier, 2021.