Scaling-up the Resources for a Freely Available Swedish VADER (*svVADER*)

Anonymous ACL submission

Abstract

With widespread commercial applications in various domains, sentiment analysis has become a success story for Natural Language Processing (NLP). Still, although sentiment analysis has rapidly progressed during the last years, mainly due to the application of modern AI technologies, many approaches apply knowledge-based strategies such as lexicon-based, to the task. This is particularly true for analyzing short social media content, e.g., tweets. Moreover, lexicon-based sentiment analysis approaches are usually preferred over learning-based methods when training data is unavailable or insufficient. Therefore, our main goal is to scale-up and apply a lexicon-based approach which can be used as a strong baseline to Swedish sentiment analysis. All scaled-up resources are made available, while the performance of this enhanced tool is evaluated on 2 short datasets, achieving adequate results.

1 Introduction

Sentiment analysis is the computational study of people's opinions, sentiments, emotions, and attitudes towards entities such as products and services, and their attributes. Sentiment analysis allows tracking of the public's mood about a particular entity to create actionable knowledge (Ligthart et al., 2021) and has found numerous applications, ranging from digital humanities (Kim & Klinger, 2022) to gaining insight into customers' feedback about commercial products and services (Rashid & Huang, 2021). Sentiment analysis can occur at the document, sentence, or word level, while the sentiment types usually assigned are *Very positive*, *Positive*, *Neutral*, *Negative* or *Very negative*. E.g., the sentiment for the sentence *Att känna stödet från publiken och folket är väldigt smickrande* 'To feel the support of the audience and the people is very flattering' will be usually assigned a positive sentiment while the sentence *Föräldrar i chock efter bluffen i basketlaget* 'Parents in shock after the hoax in the basketball team' will be assigned a negative one.

In this paper we discuss an enhancement of a popular off-the-shelf (unsupervised) dictionary-based approach to Sentiment analysis using VADER (*Valence Aware Dictionary and sEntiment Reasoner*). VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media (Hutto & Gilbert, 2014). VADER is fully open-sourced, available e.g., from the NLTK package (Bird et al., 2009), which can be applied directly to unlabeled text data. Furthermore, VADER can efficiently handle large vocabularies, including the use of degree modifiers and emoticons. These qualities make VADER a good fit for use on social media platforms to rapid sentiment text analysis. Therefore, the need for previous training as in machine or deep learning models, is eliminated.

Our main aim of this work is to make VADER a strong baseline for Swedish sentiment analysis, by rapidly scaling-up the already translated to Swedish resources (lexicons and processing tools), and to make them more reusable by further increasing and improving the lexical coverage of this resource. We further evaluate the coverage by applying and comparing the original VADER translation with the enhanced version on two small datasets, one with Swedish tweets and one random sample from the *Swedish ABSAbank-Imm*[1] which is an annotated Swedish corpus for aspect-based sentiment analysis (Rouces et al., 2020).

[1] https://spraakbanken.gu.se/en/resources/absabank-imm.

2 VADER

The *Valence Aware Dictionary for sEntiment Reasoning* (VADER[2]) is a parsimonious rule-based model for sentiment analysis of specifically social media text (Hutto & Gilbert, 2014). Since its release VADER has been extensively used in various applications and domains; from the analysis stock news headlines (Nemes & Kiss (2021); to the assessment of sentiments expressed in customers' e-mails (Borg and Boldt, 2020); and further to the analysis of tweets on covid-19 vaccine hesitancy (Verma et al., 2022).

2.1 VADER translations

VADER lexical components have been translated into several languages, such as German, French, and Italian [3]. The Swedish translation of the VADER sentiment lexicon, along with the VADER application's negators and booster words, were translated from English to Swedish, using the Google Cloud Translation API by Gustafsson (2019). However, one third of the original English sentiment lexicon remained untranslated during this process, which in a sense decrease the quality of the analysis. According to Gustafsson (2019) the original English VADER lexicon contained 7517 words, slang words, abbreviations, and emoticons. Out of these, 2435 could not be translated to Swedish because no translation could be found; for instance, many words in the English lexicon had inflections that did not exist in the Swedish counterpart; polysemy created problems, as well as English idiomaticity, e.g., slang words. The original Swedish version of the VADER sentiment resources can be found in Github[4].

2.2 Enhancements of the translated Swedish VADER: *single words, lexicalized idioms, and other multiword expressions*

The original Swedish translation of VADER was the starting point for developing and enhanced version of VADER (svVADER [5]). In general, VADER is based on a few key points when determining the sentiment of a text. Some of these points are the use of:

- *degree modifiers or booster words*, that is dampeners and intensifiers, i.e., words or characters that affects the magnitude of the polarity by either increasing or decreasing the intensity of the sentiment;
- *negations,* words which reverse the semantic orientation in a text and thus also its polarity score;
- *capitalization*, which increases the intensity of polarity, and the sentiment becomes intensified, and,
- certain types of *punctuation*, specifically exclamation marks which increase the intensity of polarity without affecting the semantic feeling.

We started refining and adapting the VADER script, in which booster words and negation items are hard coded. We both added new booster lexical items (e.g., *knappast*; *minimalt*; *svagt*; and *måttligt*) and deleted several dubious words (e.g., *effing*; *flippin*; *frackin*; *fuggin* and *hella*); similarly, some missing Swedish negation words (e.g., *icke*; *inget*; *inga* and *ej*) were also added to this script.

The characterization of the multiword expressions (MWE) and their idiomaticity play an important role in lexically based sentiment analysis. For instance, Moreno-Ortiz et al. (2013) discuss that MWEs, being units of meaning, their relative weight to the calculated overall sentiment rating of texts needs to be accounted for as such, rather than the number of component lexical units. Therefore, we added a list of 100 sentiment laden idioms, that is multiword expressions the meaning of which cannot be deduced from the literal meaning of constituent words (e.g., the Swedish idioms *blåst på konfekten* 'to be cheated on' and the Swedish idiom *tomtar på loftet* which is used to refer to someone who is stupid or crazy). The lexicalized idioms originate from the available list of the NEO lexicon DB [6] that contains a large number (over 4,000) of lexicalized idioms; the selection was made by matching all items on Tweeter and Flashback corpora, extracting the matches, and browsing manually the matched idioms annotating relevant items as positive or negative. Moreover, we manually annotated and

[2] github.com/cjhutto/vaderSentiment.

[3] See here German (Tyman et al., 2019) (github.com/KarstenAMF/GerVADER); French (github.com/vr0nsky/vadersentiment_fr) and here details for Italian (Martinis et al., 2022).

[4] https://github.com/AlexGustafsson/vaderSentiment-swedish.

[5] https://[<removed-for-review>.

[6] https://spraakbanken.gu.se/en/resources/neo-idiom.

added over 200 phrasal verbs, (e.g., *spränga ihjäl sig* ‘to blow yourself up’; *skälla ut* ‘to scold’; *rusta ner* ‘to gear down’ and *rusta upp* ‘to gear up’). Statistically significant collocations were also added, these were extracted from the analysis of the two larger collections where the two datasets originate from (cf. Section 3). Also, common medical terminology[7] (i.e., roughly 500 symptoms and frequent disease names) where added with negative polarity to svVADER’s main lexicon. Finally, we created an emoj[8] list (3,500) with Swedish expansion (meaning) downloaded and refined from various Internet sites; i.e., 😂 *ansikte med glädjetårar* ‘face with tears of joy’.

Table 1 shows the current lexical content of the original and enhanced versions of svVADER.

Name	Size	License
Original translation	5,501	MIT License
Enhanced *single words*	58,100	CC BY 4.0*
Enhanced *MWE*	2,400	

Table 1: The size of the Swedish lexicons with and without lexical enhancements (single words: includes inflected forms; MWE: Multi-Word Expressions; ‘*’: license of the new, enhanced lexical resources).

3 Application scenario: Swedish tweets about mRNA vaccines and Flashback posts on immigration

As an application scenario for the evaluation of svVADER we selected two small datasets. The first one consists of Swedish tweets posted in 2022 that discuss vaccine skepticism, and particularly, anxiety about possible side effects and concerns related to novel vaccine technologies, such as the messenger RNA (mRNA) which has be used as a reason for not receiving (the COVID-19) vaccine (Leong et al., 2022). The extracted Swedish tweets were collected with the keywords m-?RNA.* (‘?’ the preceding character is optional.; ‘.*’: ≥ 0 characters) or the hashtag #mRNA and lang:sv (Swedish content). From the extracted tweets (ca 1,800), a random selection of 200 tweet was selected for the svVADER evaluation.

The second dataset originates from the ABSAbank-Imm (where ABSA stands for "Aspect-Based Sentiment Analysis" and Imm for "Immigration", a subset of the Swedish ABSAbank) annotated dataset (Rouces et al., 2020) we randomly extracted 315 posts. In ABSAbank-Imm, texts and paragraphs are manually labelled according to the sentiment (on 1-5 scale) that the author expresses towards immigration in Sweden (a task is known as aspect-based sentiment analysis or stance analysis). The 315 posts come from the Flashback Forum[9], a popular Swedish discussion platform. For simplicity, the extracted posts consisted of only one delimited with full-stop segments; roughly a post with only one or two sentences; posts that consisted of long paragraphs were excluded. Moreover, the selected posts were labelled as positive if their manually assigned score in ABSA was 5.0 (very positive) or 4.0 (positive) and negative if their manually assigned score was 1.0 (very negative) or 2.0 (negative). Posts that lied in the middle scale with ratio 3.0 were labelled as neutral. For practical reasons, we collapsed the scores 5.0 and 4.0 to positive sentiment and 1.0 and 2.0 to negative.

3.1 Experimental results and evaluation

The ABSAbank-Imm dataset was already manually labelled, while the Tweeter dataset was manually labelled by one of the authors and a Master student, the inter-annotator agreement[10] was high (Fleiss’ $\kappa \approx 0.839$).

VADER’s sentiment score is returned in both as a compound score or as *pos*itive, *neg*ative, and *neu*tral. The compound score is computed by summing the valence scores of each word in the text, adjusted according to the rules, and then normalized to be between –1 (very negative) and +1 (very positive). Specifically, VADER's compound sentiment score determines the underlying sentiment of a text (i.e., each tweet or Flashback post) according to the following schema:

- positive, compound score ≥ 0.05
- negative, compound score ≤ – 0.05
- a neutral, the compound score is between > – 0.05 and < 0.05

We use the *original* Swedish VADER translation to automatically classify each tweet and each Flashback post according to its semantic orientation and we then proceed to classify the

[7]Motivated by the fact that there is an interest to analyzed social media with health-related content.
[8]https://emojipedia.org/sv/.

[9]https://www.flashback.org/.
[10]For the interrater reliability and agreement, we applied the R package *irr* 0.84.1.

same data with the enhanced resources. Table 2 summarizes the results of the evaluation, which clearly shows, as expected, that the enhanced approach improved the compound score results based on the original VADER translation.

Modell	CS: Swedish Tweets	F1 ABSAbank
VADER	36,7%	37,2%
svVADER$_{single\ words}$	50,8%	48,2%
svVADER$_{all}$	51%	48,1%

Table 2: Evaluation results for the original Swedish translation of VADER and the enhanced flavors of svVADER. (CS: Compound Score; svVADER$_{single\ words}$: original translation with the addition of the new non-MWE words).

For the evaluation of (sv)VADER's performance, we apply a slightly adopted version of the SemEval-2017 Task 4 (Rosenthal et al., 2017), evaluation script[11]. As with other approaches to sentiment analysis there are several pros and cons to the task. The approach is relatively easy to implement and understand, and, given the magnitude of customer experience for products and services available online it becomes doable to capture relevant datasets. However, since the model is primarily designed for use with social media content in mind, the analysis may easily overlook important words or usage. Social media input is usually loaded with typos, misspellings, slang, and grammatical mistakes, including the misinterpretation of ironic or sarcastic statements. Moreover, (sv)VADER ignores the context of the words it analyzes, particularly when word order and discontinuous structures involve cases where the insertion of e.g., one or more lexical items, appears between a lexicalized multiword entry and at a longer distance than the very near context.

4 Conclusions and future work

VADER offers a simple process for sentiment classification with a design focus on social media texts, where no training data is required, and can be used as a baseline method to evaluate and compare other methods. In this paper we outlined the scaling-up process for a dictionary approach to Swedish sentiment analysis using the VADER, a less resource-consuming lexicon and rule-based sentiment analysis tool that consumes fewer resources as compared to learning models as there is no need for vast amounts of training data. As such, VADER can serve as a good starting point to sentiment analysis before diving into more elaborated machine learning (such as transfer learning; Prottasha et al., 2022); semiautomatic lexicon based (Chanlekha et al., 2018) or (pre-trained) deep learning models[12] which stand out in terms of usage the last years, and compare their results (Dang et al., 2020). Therefore, for higher level of accuracy, it may be worth evaluating alternatives or even better a combination of alternative models using the VADER's sentiment scores as input feature to ensemble learning approaches (Kazmaier & van Vuuren, 2022).

The performance of svVADER was further evaluated on two, rather small, but characteristic Swedish social media datasets. One that contains 200 tweets and one with 200 single-sentenced posts from Flashback and the achieved results were adequate. We have also shown several ways to augment and expand the resources, and there is a strong indication in which MWEs can slightly contribute to the improvement of the results (semantic orientation) of the texts. Perhaps evaluation on much larger and varied datasets could achieve better performance.

Limitations

There are many challenges with this approach. For instance, the representativity of the selection of tweets or the social media sample (and their size) might be low and polarized, so further experimentation is necessary on larger, manually curated datasets to verify the efficacy of the tool and resources on different domains, text genres, more formal language and other, more neutral, collections of texts. Apart from the text selection process, the major consideration is that this paper didn't provide a comparison with learning methods, a task we left for future research. Moreover, although we have had initial

[11]https://github.com/cardiffnlp/xlm-t/blob/main/src/evaluation_script.py

[12]A starting point could be the Swedish BERT models for sentiment analysis: Recorded Future & AI Sweden https://huggingface.co/RecordedFuture/Swedish-Sentiment-Fear. The two models are based on the KB/bert-base-swedish-cased model (https://huggingface.co/KB/bert-base-swedish-cased) and have been fine-tuned to solve a multi-label sentiment analysis task.

experimentations with a Swedish emolex (Mohammad, 2021), we leave the description and evaluation of the approach also as a future task.

Acknowledgments

This work is part of the project *XXXXXXX XXXXXXX* (*XXXXXXX* -NNNN) financed by the *XXXXXX and XXXXXXX*.

References

Hassan Abdirahman Farah and Arzu Gorgulu Kakisim. 2023. Enhancing Lexicon Based Sentiment Analysis Using n-gram Approach. *Smart Applications with Advanced Machine Learning and Human-Centred Problem Design*. ICAIAME. Engineering Cyber-Physical Systems and Critical Infrastructures, vol 1. Springer, Cham. https://doi.org/10.1007/978-3-031-09753-9_17

Valentin Barriere and Alexandra Balahur. 2020. Improving Sentiment Analysis over non-English Tweets using Multilingual Transformers and Automatic Translation for Data-Augmentation. *Proceedings of the 28th International Conference on Computational Linguistics*. Pages 266–271 Barcelona, Spain (Online).

Steven Bird, Ewan Klein and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Anton Borg and Martin Boldt. 2020. Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Syst. Appl.,* 162. 113746, https://doi.org/10.1016/j.eswa.2020.113746

Nhan Cach Dang, María N. Moreno-García, and Fernando De la Prieta. 2020. Sentiment Analysis Based on Deep Learning: A Comparative Study. Electronics 9:3: 483. 10.3390/electronics9030483

Hutchatai Chanlekha, Wanchat Damdoung and Mukda Suktarachan. 2018. The Development of Semi-automatic Sentiment Lexicon Construction Tool for Thai Sentiment Analysis. *Advances in Intelligent Systems and Computing.* Vol 684. Springer, Cham. https://doi.org/10.1007/978-3-319-70016-8_9

Marcus Gustafsson. 2020. Sentiment analysis for tweets in swedish: Using a sentiment lexicon with syntactic rules. Bachelor's thesis. [Online]. http://www.diva-portal.org/smash/record.jsf?pid=diva2:1391359

Clayton J. Hutto and Eric Gilbert. 2014. A Parsimonious rule-based model for sentiment analysis of social media text. *18th International Conference on Weblogs and Social Media* (ICWSM-14). Stanford, USA.

Jacqueline Kazmaier and Jan H. van Vuuren. 2022. The power of ensemble learning in sentiment analysis. *Expert Systems with Applications*. Vol 187. https://doi.org/10.1016/j.eswa.2021.115819.

Evgeny Kim and Roman Klinger. 2022. A Survey on Sentiment and Emotion Analysis for Computational Literary Studies. *arXiv preprint* https://doi.org/10.48550/arXiv.1808.03137 (v4).

Ching Leong, Lawrence Jin, Dayoung Kim, Jeongbin Kim, Yik Ying Teo and Teck-Hua Ho. 2022. Assessing the impact of novelty and conformity on hesitancy towards COVID-19 vaccines using mRNA technology. *Com Med.* 2:61. https://doi.org/10.1038/s43856-022-00123-6.

Alexander Ligthart, Cagatay Catal and Bedir Tekinerdogan. 2021. Systematic reviews in sentiment analysis: a tertiary study. *Artif Intell Rev* 54, 4997–5053. https://doi.org/10.1007/s10462-021-09973-3

Maria Chiara Martinis, Chiara Zucco, Mario Cannataro. 2022. An Italian lexicon-based sentiment analysis approach for medical applications. *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. https://doi.org/10.1145/3535508.3545594

Saif M. Mohammad. 2021. Sentiment Analysis: Automatically Detecting Valence, Emotions, and Other Affectual States from Text. *arXiv preprint* https://doi.org/10.48550/arXiv.2005.11882 (v2).

Antonio Moreno-Ortiz, Chantal Pérez-Hernández and Maria Del-Olmo. 2013. Managing Multiword Expressions in a Lexicon-Based Sentiment Analysis System for Spanish *Proceedings of the 9th Workshop on Multiword Expressions*. Pp. 1-10. Atlanta, USA. https://aclanthology.org/W13-1001

László Nemes and Attila Kiss. 2021. Prediction of stock values changes using sentiment analysis of stock news headlines. *J. Inf Telec*, 5. pp. 375-394, https://doi.org/10.1080/24751839.2021.1874252

Nusrat Jahan Prottasha, Abdullah As Sami, Md Kowsher, Saydul Akbar Murad, Anupam Kumar Bairagi, Mehedi Masud and Mohammed Baz. 2022. Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning. *Sensors* 22:11: 4157. https://doi.org/10.3390/s22114157

Aamir Rashid and Ching-Yu Huang. 2021. Sentiment Analysis on Consumer Reviews of Amazon Products. *J of Comp Theory & Eng*. Vol. 13:2. https://doi.org/10.7763/IJCTE.2021.V13.1287.

Sara Rosenthal, Noura Farra and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. *Proceedings of the 11th International Workshop on Semantic Evaluations* (SemEval-2), pp. 502–518. Vancouver, Canada. c ACL. https://aclanthology.org/S17-2088.pdf

Jacobo Rouces, Lars Borin and Nina Tahmasebi. 2020. Creating an Annotated Corpus for Aspect-Based Sentiment Analysis in Swedish. *Proceedings of the 5th conference in Digital Humanities in the Nordic Countries*. Riga, Latvia. http://ceur-ws.org/Vol-2612/short18.pdf

Karsten Michael Tymann, Matthias Lutz, Patrick Palsbröker and Carsten Gips. 2019. GerVADER - A German adaptation of the VADER sentiment analysis tool for social media texts. *Proceedings of the Conference Lernen, Wissen, Daten, Analysen* (LWDA). Berlin, Germany

Ravi Verma, Amit Chhabra & Ankit Gupta. 2022. A statistical analysis of tweets on covid-19 vaccine hesitancy utilizing opinion mining: an Indian perspective. *Soc. Netw. Anal. Min.* 13, 12. https://doi.org/10.1007/s13278-022-01015-2.