SSL *Scandinavian Studies in Language, 12(1), 2021 - https://tidsskrift.dk/sss*

MULTILINGUAL SENTIMENT NORMALIZATION FOR SCANDINAVIAN LANGUAGES

Rebekah B. Baglini[1]
Interacting Minds Centre, Aarhus University
rbkh@cc.au.dk

Anita Kurm[2]
Aarhus University

Lasse Hansen[1,2,3]
Aarhus University

Kenneth Enevoldsen[1,2]
Aarhus University

Kristoffer L. Nielbo[1, 2]
Aarhus University

Abstract*
In this paper, we address the challenge of multilingual sentiment analysis using a traditional lexicon and rule-based sentiment instrument that is tailored to capture sentiment patterns in a particular language. Focusing on a case study of three closely related Scandinavian languages (Danish, Norwegian, and Swedish) and using three tailored versions of VADER, we measure the relative degree of variation in valence using the OPUS corpus. We found that scores for Swedish are systematically skewed lower than Danish for translational pairs, and that scores for Norwegian are skewed higher for both other languages. We use a neural network to optimize the fit between Norwegian and Swedish respectively and Danish as the reference (target) language.

Keywords: Multilingual Sentiment, Sentiment Lexicon, Normalization

1. Introduction

Sentiment analysis and classification encompasses a set of NLP techniques for extracting and quantifying emotional valence from linguistic data. These techniques originate in affective computing and social psychology, and, today, we see a wide range of application domains ranging from customer profiling (M. Hu and Liu 2004) and

*This research was supported the ”HOPE - How Democracies Cope with COVID-19”-project funded by The Carlsberg Foundation with grant CF20-0044 and NeiC’s Nordic Digital Humanities Laboratory project, and DeiC Type-1 HPC with project DeiC-AU1-L-000001.

1: Interacting Minds Centre, Aarhus University, Denmark; 2: Center for Humanities Computing Aarhus, Aarhus University, Denmark; 3: Department of Clinical Medicine - Department of Affective Disorders, Aarhus University, Denmark

social media monitoring (Chew and Eysenbach 2010) to literary analysis (M. Hu and Liu 2004). Like many NLP techniques, sentiment analysis depends on the availability of large data sets, and therefore favors high resource languages, especially contemporary English. While language models based on deep learning techniques are showing promising developments in multilingual models, e.g., (Devlin et al. 2019), traditional approaches to sentiment analysis are tailored to capture sentiment patterns in a particular language and multilingual analysis is achieved through more or less uncorrected automated translation. In this study, we addressed the challenge of multilingual sentiment analysis with traditional lexicon and rule-based tools with a case study of three closely related Scandinavian languages (Danish, Norwegian, and Swedish). For this task, we measured the relative degree of variation in valence for three tailored versions of VADER (Valence Aware Dictionary for sEntiment Reasoning) (Hutto and Gilbert 2015) using the OPUS corpus (Tiedemann 2012). We found evidence of a systematic skew in the compound sentiment scores produced for parallel texts: scores for Swedish are systematically skewed lower than Danish for translational pairs, and that scores for Norwegian are skewed higher for both other languages. To correct the skew, we approached alignment across languages as a function approximation problem and trained a neural network to optimize the fit between languages. This study represents a new method for comparative lexicon-based sentiment analysis for Scandinavian languages without the use of automated translation.

1.1 Applications of sentiment analysis

Modern applications of sentiment analysis (after 2000) combine central insights from affective computing (Picard 1997) and social psychology (Pennebaker, Francis, and Booth 1999) in order to identify and extract expressions of emotion, mood and tone in textual data, and quantify their intensity, or in the case of sentiment classification, categorize them according to their polarity. While the development of sentiment analysis is driven by computational linguistics and NLP, (e.g., Nielsen 2011, Devlin et al. 2019), the search applications are many and the update in social sciences and humanities research has been growing over the last decade (e.g., Thelwall 2011, O’Connor 2010, Hu). Business applications have similarly been increasing rapidly as more out-of-the-box sentiment tools have been made available either as open, e.g., (Nielsen 2011, Hutto and Gilbert 2015, Turc et al. 2019), or closed source (Pennebaker et al. 2007).

Early approaches to sentiment analysis were a response to the growth of e-commerce in the early 2000s. As online marketplaces became the norm, the need for automated techniques for customer profiling became apparent (Pang, Lee, and Vaithyanathan 2002). These techniques typically involved a series of steps that resulted in a classification of the opinion orientation (positive/negative) of the sentences and paragraphs in relatively short texts, e.g., customer reviews (M. Hu and Liu 2004). The sentiment component underlying these techniques consisted of lexical matching between a text and a full forms list of English words and their associated sentiment polarity, category or score, e.g., Opinion Lexicon (Hu and Liu 2004) or LIWC (Pennebaker, Francis, and Booth 1999).

As global uptake of social media and micro-blogging accelerated, the value of fast techniques for opinion mining in short texts became even more apparent. This resulted in a multitude of sentiment instruments that combined lexical matching with combinatorial rules to manage negations and order, see (Reagan et al. 2015). Particularly the social media platform Twitter has, due to its liberal data policy, been the target for many applications. These studies use Twitter sentiment to track how emotions impact stock market behavior (Bollen, Mao, and Zeng 2011), political position (Tumasjan et al. 2010, O'Connor et al. 2010), the evolution of catastrophic events (Chew and Eysenbach 2010) and, more general, communicative event responses (Thelwall, Buckley, and Paltoglou 2011). Figure 1 show an example of an event- based Twitter analysis during the Covid-19 vaccine rollout in Denmark. Initially, the sentiment reflects a positive attitude towards the vaccination as Pfizer and Moderna are approved and vaccination commences. The initial optimism is replaced by negativity as the rollout is delayed early January. The approval of AstraZeneca (AZ) mitigates the negativity temporarily until AZ is put on hold and, later, withdrawn. Several studies have confirmed that sentiment analysis of tweets captures offline opinions and behavior (Tumasjan et al. 2010, Chew and Eysenbach 2010).

Coinciding with the global adoption of social media was the development of contemporary machine learning that utilized (semi-)supervised learning to train algorithms that can predict sentiment distributions of texts (Thelwall, Buckley, Paltoglou, et al. 2010). In many cases, these techniques outperform lexicon-based techniques on sentiment analysis tasks (Socher et al. 2011, Le and Mikolov 2014), but their reliance on specific large data sets (Reagan et al. 2015) can make them less appealing for low resource languages.

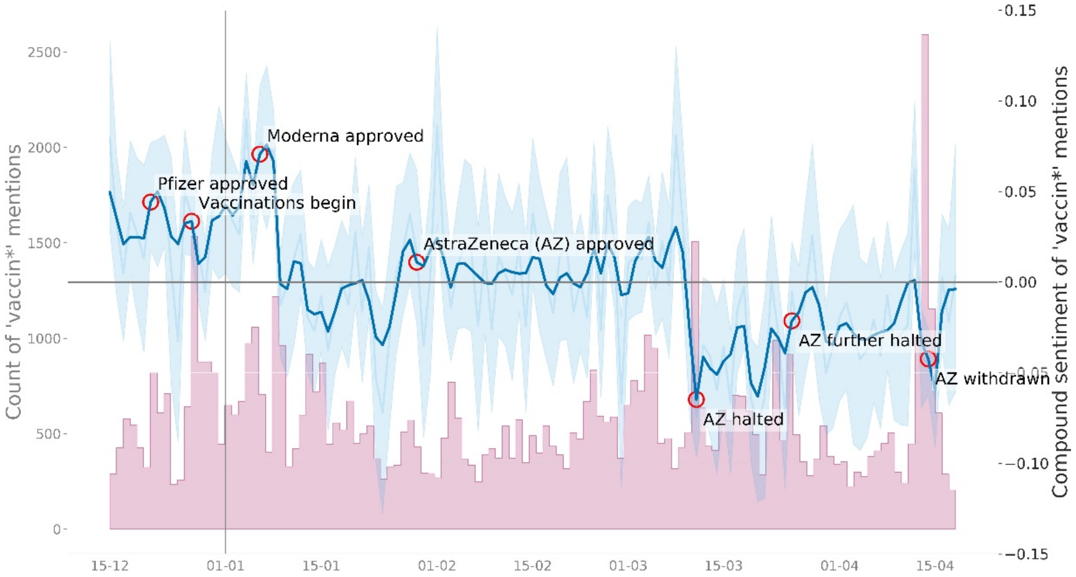

Figure 1: Compound sentiment scores (blue line, right ordinate) from Danish tweets containing mentions of 'vaccin' during the early Covid-19 vaccine rollout. Number of tweets are shown with bars (red, left ordinate). Red circles with labels indicate significant events during the vaccine rollout.

The motivation for the present study comes from our participation in the ongoing HOPE Project (How Democracies Cope with Covid-19), which provides continuous monitoring and analysis of trends in social and behavioral aspects of the Covid-19 pandemic in several societies, including Denmark, Norway, and Sweden. One topic of interest is the co-variance of windowed average sentiment scores on Twitter with other variables, such as lockdown stringency or infection rate. The Scandinavian countries provide a particularly interesting case study for cross-linguistic sentiment comparison, as they adopted very different strategies for mitigating the pandemic during the first wave (in particular, Denmark and Norway versus Sweden). This prompted us to pursue new methods for improving existing tools for comparative sentiment analysis between Norwegian, Danish, and Swedish.

1.2 Lexicon-based sentiment analysis

Contemporary methods for sentiment analysis are either lexicon and rule-based or machine learning-based. Both categories can be further subdivided depending on their reliance on specific corpora, incorporation of rules, and for machine learning, whether they use supervised, semi-supervised or unsupervised learning. In this study we focus on lexicon-based approaches to Nordic languages for several reasons. First, they are predominantly corpus agnostic and do not rely on training data, which makes them less dependent on a specific purpose or context than machine learning-based approaches (Reagan et al. 2015). Second, the scoring of lexicon-based approaches is transparent and can always be explained in terms of combinations of word-level sentiments. This again is in stark contrast to machine learning, where the effect of atomic elements is, in many cases, black boxed (Reagan et al. 2015). Finally, the focus on multilingual sentiment analysis for Nordic languages is due to the high degree of linguistic affinity between these languages, which makes them easier to compare, and that they are, relatively speaking, low-resource languages.

Lexicons, that is, mappings between words or lemmas and sentiment values, are typically generated manually or through surveys that assign sentiment scores to words. Criteria for word selection vary from expert-based to purely statistical, and validation procedure range from qualitative assessment to performance comparisons on curated corpora and sentiment analysis tasks (Reagan et al. 2015, Le and Mikolov 2014). In spite of accuracy issues stemming from contextual dependencies between words and the finite size of lexicons, lexicon-based tools are widely used in automated sentiment analysis and opinion-mining tasks, and remain the most accessible and easy to validate method of performing sentiment analysis on a text corpus.

Valence Aware Dictionary and sEntiment Reasoner (VADER) (Hutto and Gilbert 2015) is one of the most widely used lexicon-based sentiment instruments, and represents the cur- rent state of the art within this class. VADER was developed for social media data, but has been shown to generalize to other domains (Hutto and Gilbert 2015). Despite the existence of more sophisticated ML methods for sentiment analysis, the simplicity of VADER gives it several advantages and explains its continued popularity. First, VADER is computationally economical and fast to run, even on a standard laptop computer. Second, it is designed to be domain-general, based on grammatical and syntactic rules and a broad sentiment lexicon that generalized across many different types of texts. The lexicon itself is human-validated, and the sentiment instrument has been shown to outperform other similar instruments and human coders across a variety of text genres, including social media texts, without requiring any training datasets nor fine-tuning[1]. The third advantage is transparency and interpretability. The lexicon and rules for VADER are directly accessible to the user, making it easy to understand how the sentiment scores are produced and to inspect, extend, and modify the rules and lexicon. This is particularly important for researchers in areas of the social sciences and humanities, where the use of automated computational research tools is less common and where it is important to be able to directly examine and interpret the scoring instrument. Finally, unlike modern ML sentiment analyzers which assign only a degree of probability that a text falls into a given class, VADER assigns a more fine-grained compound score based on the scores of composite words (potentially augmented by rules), reflecting the predicted degree of positivity or negativity of the text.

Prior to this study, cross-linguistic comparative sentiment analysis using VADER required the use of a multilingual version of VADER, which uses integration with Google Translate's API to automatically translate source languages into English to produce sentiment scores. However, translation-based approaches are inadequate for high-fidelity multilingual sentiment analysis given the high degree of variation in the valency in the translational equivalents of evaluative, subjective, and expressive words and expressions (Jackson et al. 2019, Mohammad, Salameh, and Kiritchenko 2016, Rouces et al. 2018). To avoid the problem of translation, we propose an alternative approach to cross-linguistic VADER sentiment analysis which applies a normalization algorithm, trained on a parallel corpus, to adjust language-specific VADER model scores for Danish, Norwegian, and Swedish texts making them more comparable.

2. Methods

2.1 VADER - Lexicon for Danish, Swedish, and Norwegian

The sentiment instrument used in this study builds on VADER, which combines a substantial full forms list (7520 words and emoticons for English) and rules for sentence handling (e.g., negations, degree modifiers, word-shape) (Hutto and Gilbert 2015). The Danish VADER developed for this study builds upon previous Danish lexicon-based

[1] A limitation of VADER is that is has only been designed for and validated against relatively recent corpora. Given the nature of language change and semantic shifts among high-sentiment words in particular, it is likely that VADER and other lexicon-based sentiment tools are unsatisfactory compared with a approach which is fine-tuned on data from the target historical era.

sentiment instruments, e.g., AFINN (Nielsen 2011) and SENTIDA (Lauridsen, Dalsgaard, and Svendsen 2019), and extends upon them by adding support emoji and replacing the use of stems with lemmas. The Swedish VADER utilized in the study is the publicly available VADER implementation of Swedish, and the Norwegian (Bokmål) sentiment instrument is a translation of the Danish corrected and validated by to native speakers.[1] Because the lexicons were relatively small (DK: 5264, SW: 5501, NO: 3214), all models used lemmatization, as opposed to the full form English instrument, to obtain wider coverage.

Note that for Danish, comparatively better performance on sentiment analysis tasks has been found using machine learning sentiment instruments with attention-based language models such as BERT (Devlin et al. 2019), but these only provide sentiment classification, e.g., (Wang et al. 2019).

2.2 OPUS - Parallel translation corpus

To compare sentiment instruments across languages, we used the collection of parallel corpora OPUS, which is widely used for statistical machine translation (Tiedemann 2012). OPUS is open and consists of a large collection of translated texts from the web. We specifically used the OpenSubtitles corpus (Lison & Tiedemann 2016) with > 1.9 million parallel sentences for the three languages in question, Table 2 for descriptive statistics.

2.3 Task, feature selection and model architecture

To learn a normalization function for the three languages, we defined a simple prediction task for a neural network mimicking an auto-encoder. For each sentence pair, the network had to map a sentiment score from the source language, e.g., *Swedish*, onto the target language, e.g., *Danish*, where the target language is constant for all models. Experiments with several network architecture (Sequential models with 2-3 dense layers composed of 16, 32, 64 nodes), showed that adding an additional one-hot encoded context feature to the input resulted in a reliable performance improvement. Grid search was used to identify the final sentiment ranges for the context feature, see Table 1[2].

[2] The sentiment instruments for Danish, Swedish and Norwegian are available on https://github.com/centre- for-humanities-computing/text to x.

Class	**Range**
extremely negative	$[-1.0, -0.4]$
negative	$(-0.4, -0.2)$
neutral	$[-0.2, 0.2]$
mildly positive	$(0.2, 0.4)$
positive	$[0.4, 1.0]$

Table 1: Final input feature that provide context information for ranges of sentiment scores. The feature can also be used to transform output to a sentiment classifier.

To learn the normalization function, we used *Autokeras* 1.0.2 (Jin, Song, and X. Hu 2019) to automate architecture selection with all features. Hyper-parameter selection was performed with sweep from *Weights and Biases* (Biewald 2020) on a validation set consisting of 20% of the data. This was done using a Bayesian grid search over the following parameters: number of units in each layer l1-5 (16-254), batch size (250-1000), dropout, and learning rate (0.0005- 0.002). The resulting optimal parameters were: l1 111 units, l2 174 units, l3 225 units, l4 247 units, l5 36 units, batch size 623, dropout 0.06, learning rate 0.0005.

3. Results

Initially we conducted a pairwise comparisons for the Danish, Swedish and Norwegian sentiment instruments using OpenSubtitles v.2018. From the corpus we sampled ~ 1.9 million parallel sentence pairs for each language combination. Table 2 shows exact sizes for the parallel corpora and non-normalized average comparison for compound sentiment scores. Observe the systematic skew from Table 2, that is, in comparison to Danish VADER ($Danish_M$), the Norwegian sentiment instrument ($Norwegian_M$) has a tendency to assign higher sentiment values $Norwegian_M > Danish_M$ to paired sentences, while the Swedish instrument assigns lower scores $Danish_M > Swedish_M$.

Language pair	**Dataset size**	*Danish$_M$*	*Swedish$_M$*	*Norwegian$_M$*
Danish - Swedish	1,902,685	0.023	0.013	-
Danish - Norwegian	1,920,409	0.023	-	0.046
Swedish - Norwegian	1,909,422	-	0.012	0.047

Table 2: Total number of OpenSubtitles parallel sentences and mean VADER compound score.

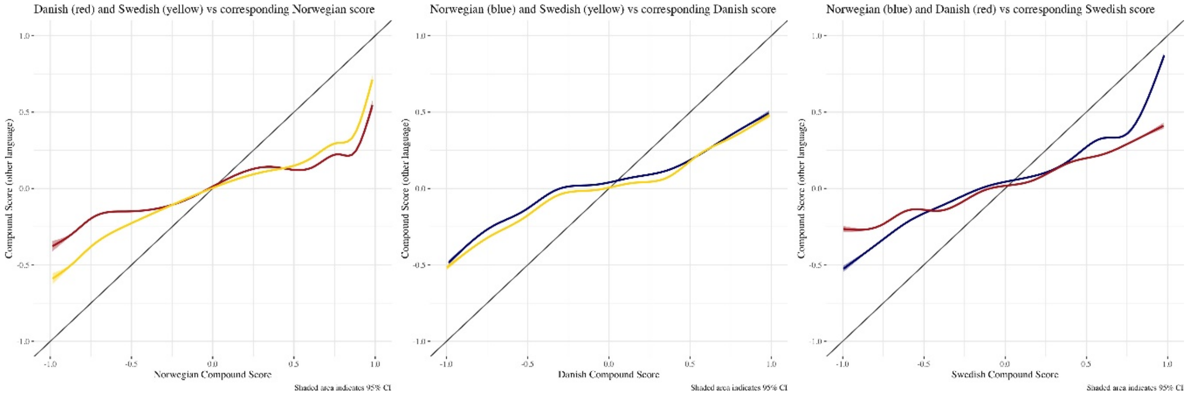

Figure 2: Non-normalized sentiment scores in Norwegian (blue), Swedish (yellow) and Danish (red) plotted against each other.

To further explore the systematic skew, we fitted two source languages on a target language for each possible combination using smooth bivariate spline approximation, see Figure 2, that is, *Danish* and *Swedish* on *Norwegian* in the left column, *Norwegian* and *Swedish* on *Danish* in the centre column, and *Norwegian* and *Danish* on *Swedish* in the right column. The fits on *Danish* for both source languages are similarly aligned to the target, which provides the rationale for selecting the Danish sentiment instrument as the target baseline for normalization. In what follows we therefore developed *Swedish to Danish* and *Norwegian to Danish* models in order to improve alignment for the two other sentiment instruments.

Language pair	*Danish*	*Swedish*	*Norwegian*
Danish - Swedish	$M = 0.025, SD = 0.264$	$M = 0.025, SD = 0.104$ $(M = 0.014, SD = 0.272)$	-
Danish - Norwegian	$M = 0.022, SD = 0.258$	-	$M = 0.022, SD = 0.094$ $(M = 0.049, SD = 0.083)$
Swedish - Norwegian	-	$M = 0.022, SD = 0.104$ $(M = 0.011, SD = 0.273)$	$M = 0.024, SD = 0.094$ $(M = 0.048, SD = 0.284)$

Table 3: Adjusted compound scores (mean and standard deviation) after normalization. Non- normalized scores in parentheses. (Data: the 50K sample from the OpenSubtitles test set for each language pair).

Finally, we trained neural network models to perform the pairwise function approximation between the source language scores as input and target language as output. For every language pair, a random sample (*n* = 50, 000) was selected from the testing set for evaluation. As shown in Table 3, the normalization produces almost identical mean sentiment values in parallel sentences, but also decreases variation of sentiment scores for normalised data. To further compare the normaliziation effect,

Language pair	**Original alignment**	**Normalized alignment**
Danish - Swedish	$r_S = 0.297$ $RMSE = 0.301$	$r_s = 0.316(\uparrow +6.4\%)$ $RMSE = 0.339(\downarrow -20.6\%)$
Danish - Norwegian	$r_S = 0.273$ $RMSE = 0.320$	$r_s = 0.288(\uparrow +5.5\%)$ $RMSE = 0.240(\downarrow -25.0\%)$
Swedish - Norwegian	$r_S = 0.377$ $RMSE = 0.305$	$r_s = 0.395(\uparrow +4.8\%)$ $RMSE = 0.099(\downarrow -67.5\%)$

Table 4: Spearman correlation r_s and RMSE before and after normalization using the developed models. All reported correlation tests yielded p < 0.001 and for pairwise comparisons with Fisher's z-transform p <0.001. (Data: the 50K sample from the OpenSubtitles test set for each language pair)

we correlated sample sentiment pairs pr. language using Spearman's rank correlation coefficient, see Table 4. As expected, the normalization reliably increased the overall association between languages.

4. Discussion

While machine learning techniques are superior to lexicon and rule-based sentiment instruments, the latter remains hugely popular in social sciences and humanities because of their versatility and transparency. Lexicons are tailored to sentiment patterns of specific languages and multilingual instruments therefore relies on automated translation. With this study, we proposed a different approach to comparative sentiment analysis for these Scandinavian languages. We showed that we can approximate a normalization function using a learning approach (function approximation) that aligns source languages to a target baseline thereby effectively eliminating potentials skews and biases.

Several issues remain: Corpus-dependency, performance validation, variation squashing, and real cultural effect. First, by relying on a specific corpus for training the model, the approach indirectly introduces corpus-dependency in lexicon-based approaches. The issue however is less severe compared to machine learning-based approaches, because it only impacts normalization, and inference is still carried out by the lexicon. Second, neither the baseline nor normalized scores produced by the three VADER models have been properly validated against an evaluation corpus. This issue reflects that our approach is work in progress and a benchmarking task, comparing normalized scores with compound scores of the three original VADER models and VADER multilingual, are the next task. Third, from Table 3 it could be observed that the model learned a 'variance squashing' function in order to align languages. This however can be remedied by re-scaling all scores to $[-1, 1]$, which will retain the alignment effect. Fourth, we have implicitly assumed that the systematic skew in the language comparison is an artifact of the VADER. It is however possible that there are real cultural differences in Scandinavia that impact linguistic expressions of emotion, mood and tone. At a first glance, an interpretation of Figure 2 could be that Norwegians (on average) are more positive than their Scandinavian neighbours. Disregarding the potentially stereotype re-enforcing aspect of this interpretation, it is also badly aligned with data from the *World Happiness Report* and ignores the impact of random variation in development of sentiment lexicons. Nevertheless, there are indications of cultural differences both within and between countries in affective meanings, e.g. (Heise 2014, Jackson et al. 2019). While anecdotally, we have observed that Danish emotion expressions in social media seem to utilize the sentiment range in a less continuous fashion than Norwegian and Swedish, and cluster them around a narrow set of ranges.

We predict that this approach to sentiment score alignment is best-suited for languages which are typologically, semantically, and culturally similar. For the purpose of cross-linguistic sentiment analysis of the Scandinavian languages Danish, Norwegian, and Swedish, we find that automatically aligning sentiment scores produced by each language's version of VADER provides a useful alternative to using automatic translation and a single sentiment instrument. Since the language-specific validated VADER lexicons and rules can be examined independently and the score-normalization is achieved using simple function approximation, this method provides more transparency than automatic translation. Where the expected scope of semantic and cultural divergence is relatively small (as is the case with the Scandinavian languages, given their shared linguistic lineage and close historical and socio-cultural ties), important semantic variance is less likely to be obscured by producing aligned scores through normalization.

References

Biewald, Lukas (2020). Experiment Tracking with Weights and Biases. URL: https://www.wandb.com/

Bollen, J., H. Mao, and X. Zeng (2011). Twitter mood predicts the stock market. In: Journal of Computational Science 2.1, pp. 1–8. DOI: https://10.1016/j.jocs.2010.12.007

Chew, C. and G. Eysenbach (2010). Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. In: PLoS ONE 5.11. DOI: https://10.1371/journal.pone.0014118

Devlin, Jacob et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805v2 [cs.CL]

Heise, David R (Dec. 2014). Cultural variations in sentiments. In: *SpringerPlus 3.1*, pp.1-11. DOI: https://10.1186/2193-1801-3-170

Hu, Minqing and Bing Liu (2004). Mining and Summarizing Customer Reviews. In: *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177 https://doi-org.ez.statsbiblioteket.dk:12048/10.1145/1014052.1014073

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

Jackson, Joshua Conrad, et al. (2019). Emotion semantics show both cultural variation and universal structure. In: *Science* 366.6472: 1517-1522. https://10.1126/science.aaw8160.

Jin, Haifeng, Qingquan Song, and Xia Hu (2019). Auto-Keras: An Efficient Neural Architecture Search System. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, pp. 1946–1956.

Lauridsen, G. A., Dalsgaard, J. A., & Svendsen, L. K. B. (2019). SENTIDA: A New Tool for Sentiment Analysis in Danish. Journal of Language Works - Sprogvidenskabeligt Studentertidsskrift, 4(1), 38–53. Retrieved from https://tidsskrift.dk/lwo/article/view/115711

Le, Q. and T. Mikolov (2014). Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on Machine Learning*, PMLR 32(2), pp.1188-1196.

Lison, Pierre and Jorg Tiedemann (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), pp. 923-929.

Mohammad, Saif M, Mohammad Salameh, and Svetlana Kiritchenko (2016). How Translation Alters Sentimen". In: The Journal of artificial intelligence research 55, pp. 95–130. ISSN: 1076-9757.

Nielsen, Finn A˚ rup (2011). "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs". In: arXiv preprint arXiv:1103.2903. (Visited on 01/06/2017).

O'Connor, B. et al. (2010). "From tweets to polls: Linking text sentiment to public opinion time series". In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 122–129.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan (2002). Thumbs up?: sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 conference on*

Empirical methods in natural language processing - EMNLP'02. Vol. 10. Association for Computational Linguistics, pp. 79–86. DOI: https://10.3115/1118693.1118704 (Visited on 04/30/2021).

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001).

Picard, Rosalind W. (Sept. 1997). Affective Computing. English. Second Edition 1998. Cam- bridge, Mass: The MIT Press. ISBN: 978-0-262-16170-1.

Reagan, A. et al. (2015). Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. In: arXiv preprint arXiv:1512.00531.

Rouces, Jacobo et al. (2018). SenSALDO: Creating a Sentiment Lexicon for Swedish. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, p. 7.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., & Manning, C. D. (2011). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 151-161).

Thelwall, M., K. Buckley, and G. Paltoglou (2011). Sentiment in Twitter events. In: *Journal of the American Society for Information Science and Technology* 62(2), pp. 406–418. DOI: https://10.1002/asi.21462

ThelWall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2011). Sentiment in Short Strength Detection Informal Text (vol 61, pg 2544, 2010). *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, *62*(2), 419-419. DOI: https://10.1002/asi.21416

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA)*, pp. 2214–2218.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010, May). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 4, No. 1).

Turc, Iulia et al. (2019). Well-Read Students Learn Better: On the Importance of Pre- training Compact Models. In: arXiv:1908.08962 [cs]. arXiv: 1908.08962. (Visited on 04/30/2021).

Wang, A. et al. (2019). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: arXiv:1804.07461 [cs]. arXiv: 1804.07461.