SenSALDO: a Swedish Sentiment Lexicon for the SWE-CLARIN Toolbox

Jacobo Rouces   Lars Borin
Nina Tahmasebi   Stian Rødven Eide
Språkbanken Text
University of Gothenburg, Sweden
`jacobo.rouces|lars.borin@gu.se`
`nina.tahmasebi|stian.rodven.eide@gu.se`

Abstract

The field of *sentiment analysis* or *opinion mining* consists in automatically classifying text according to the positive or negative sentiment expressed in it, and has become very popular in the last decade. However, most data and software resources are built for English and a few other languages. In this paper we compare and test different corpus-based and lexicon-based methods for creating a sentiment lexicon. We then manually curate the results of the best performing method. The result, SenSALDO, is a comprehensive sentiment lexicon for Swedish containing 7,618 word senses as well as a full-form version of this lexicon containing 65,953 items (text word forms). SenSALDO is freely available as a research tool in the SWE-CLARIN toolbox under an open-source CC-BY license.

1 Introduction

The field of *sentiment analysis* or *opinion mining* consists in automatically classifying text according to the positive or negative sentiment expressed in it, and has become very popular in the last decade (Pang and Lee, 2008). However, most data and software resources are built for English or a few other languages, and there is still a lack of resources for most languages. While often discussed in the NLP literature as a business-intelligence tool – helping online businesses keep track of customer opinion about their goods and services – there have also been a number of studies where sentiment analysis has been applied to research data in the humanities and social sciences (HSS) (Bentley et al., 2014; Eichstaedt et al., 2015; Sprugnoli et al., 2016; Thelwall, 2017). This has prompted inquiries by Swedish HSS researchers as to whether the Swedish CLARIN infrastructure could provide this kind of tool also for Swedish textual data. For this reason, at the CLARIN B center Språkbanken Text (University of Gothenburg) we initiated a concerted effort aiming at the development of a Swedish sentiment lexicon for the SWE-CLARIN toolbox.

The development of this resource – *SenSALDO* – has been done in three steps, described below: (1) creation of a gold standard word-sense list (section 3); (2) implementation and evaluation of different automatic methods for creating the sentiment lexicon (section 4); and (3) manual curation of the results of the best performing method (section 5).

In SenSALDO, each word sense has two annotations: a coarse-grained label with three possible values (‘positive’, ‘neutral’, and ‘negative’) and a more fine-grained score in the range $[-1, 1]$.

2 Lexical resources for sentiment analysis: state of the art

In recent years, a wide variety of methods has been used for building sentiment lexicons. Unsurprisingly, most of this work has focused on English, although some efforts targeting other languages have also been reported in the literature.

Some methods rely on corpus analysis (making use of word co-occurrence, syntactic patterns, or distant-supervision signals) and others on existing lexicons (usually utilizing some sort of sentiment label propagation exploiting the structure of the lexicon), although both approaches can be combined (Devitt

Jacobo Rouces, Lars Borin, Nina Tahmasebi and Stian Rødven Eide 2019. SenSALDO: a Swedish Sentiment Lexicon for the SWE-CLARIN Toolbox. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 177–187.

and Ahmad, 2013). In some cases manual annotation is used as well, either to create seeds or to curate results.

Among the English sentiment lexicons built with mostly-automatic lexicon-driven methods, SentiWordNet (Baccianella et al., 2010) has become a popular resource. It was created by combining a semi-supervised learning step that uses existing relations between WordNet 3.0 entries (Fellbaum, 1998), (such as *synonymy*, *antonymy*, and *related with*), and a random-walk step over a graph built using the implicit *definiens–definiendum* relation between words in the entries and words in the glosses of the entries (Esuli and Sebastiani, 2007). However, the use of these relations requires a structure like that of WordNet, which in turn requires a considerable amount of manual effort by trained lexicographers.

Among the English lexicons built using corpus-driven approaches, SENTPROP (Hamilton et al., 2016) constitutes a recent state-of-the-art approach where a directed weighted graph of terms is constructed using the nearest neighbors in the space of word embeddings obtained from applying singular value decomposition to the positive pointwise mutual information matrix obtained from the corpus. Then, it uses random walks in a similar fashion to SentiWordNet.

A common problem when using label propagation is that words far away from seeds get low values just by virtue of their distance, which should not be the case. Yazidi et al. (2015) propose a solution: at each iteration, a fixed number of "informative words" are selected as new seeds for labeling according to different criteria.

For Swedish, the language of interest in our case, much less work has been reported in the literature. However, two openly available sentiment lexicons existed prior to the work reported here, presented by Rosell and Kann (2010) and Nusko et al. (2016). In addition, some Swedish sentiment lexicons or word lists have been produced by automatic translation of corresponding English resources, e.g., by Mohammad and Turney (2010)[1] and Chen and Skiena (2014). Looking at these existing resources, there are obvious ways in which they can be improved. With automatically translated resources, there are more than a few strange translations or translation errors. In many of these resources the lexical items are undisambiguated text words or lemmas, i.e., all senses of polysemous words are conflated in the lexicon, even though the different senses may of course have different sentiment values. In the case of the resources organized by lemmas, there is generally no attempt at full-form expansion, i.e., having the lexicon cover all inflected forms of a lemma, which obviously makes the lexicon less useful for automated text processing. Both these issues are addressed in the work reported here, at the same time that we have been able to draw inspiration from the work of both Rosell and Kann (2010) and Nusko et al. (2016). In this sense, our lexicon builds on their earlier efforts.

3 Step 1: a gold-standard sentiment word-sense list for Swedish

The first step of our work consisted in producing a gold-standard list of sentiment-annotated Swedish word senses. This work has been reported elsewhere, and here we just provide some necessary background information. For the details, see Rouces et al. (2018a) and Rouces et al. (2018b).

SenSALDO is based on SALDO, a computational lexicon for Swedish composed, among other components, of word senses as entries and semantic relations – called *descriptors* – connecting word senses. There are two kinds of descriptors: The *primary descriptor* is obligatory. It connects an entry to exactly one other word sense (also a SALDO entry[2]). This parent word sense is a close semantic neighbor which is also more central, which means that is typically structurally simpler, stylistically more neutral, acquired earlier by a first-language learner and more frequent in usage. Any number of secondary descriptor relations provide additional semantic properties of the entry that are not conveyed by the primary descriptor, such as an inversion or negation or some important semantic argument in a hypothetical definition. The primary descriptor structure forms a tree and the secondary descriptors define a directed acyclic graph. For a detailed description of the organization of SALDO and a discussion of the underlying linguistic

[1] `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`

[2] Except in the case of 41 top entries, which are given an artficial primary descriptor in order to make all of SALDO into a single rooted tree.

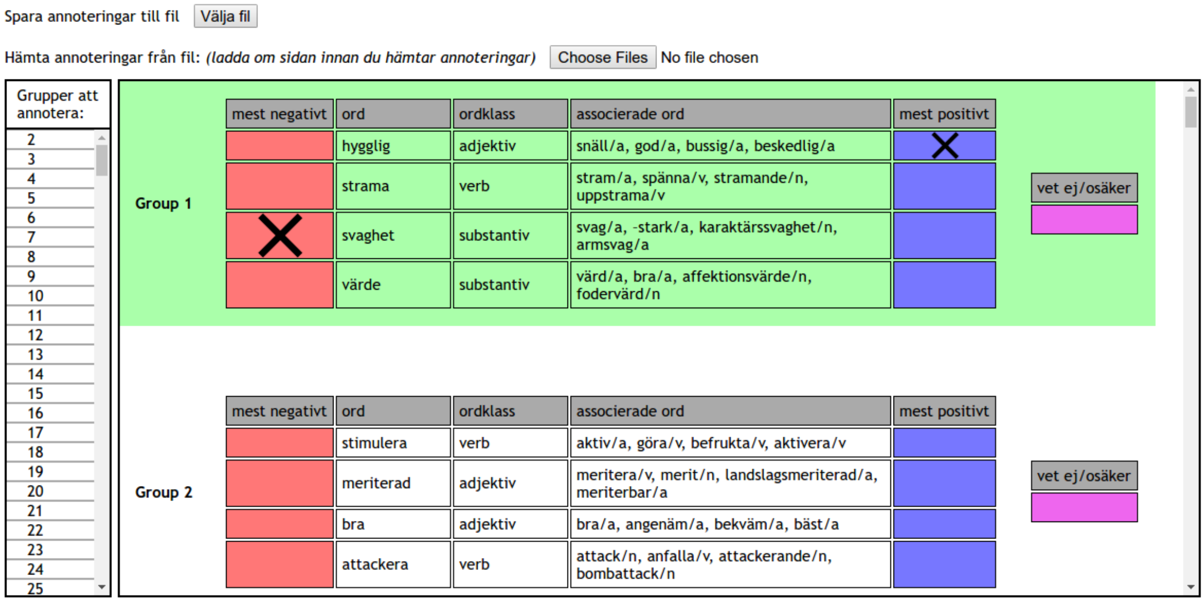


Figure 1: Screenshot for the best–worst scaling annotation interface. The labels displayed for each group are (from left to right) 'most negative', 'word', 'part of speech', 'associated words', 'most positive', 'don't know/uncertain'.

theoretical and methodological principles informing it, see Borin et al. (2013). For the work described here, we used the current stable version SALDO v. 2.3, which contains 131,020 word senses.[3]

First, an initial sampling from SALDO was done following the distribution given by the estimated frequency of each word sense in the *Swedish Culturomics Gigaword Corpus* (Eide et al., 2016), which is a one-billion-word mixed-genre corpus of written Swedish.[4] We sampled ∼2,200 open-class words (nouns, verbs, adjectives and interjections), which were annotated by three annotators (the three last authors of this paper), with 200 overlapping items in order to estimate interannotator agreement. This annotation was done using discrete labels with three possible values (-1, 0, and $+1$, for negative, neutral, and positive sentiment, respectively).

After this, four external annotators were employed to annotate a subset of the 2,200 items, such that at least two of the initial annotators had asigned a non-neutral value to each item. The resulting 278 word senses were annotated using *best–worst scaling* (Kiritchenko and Mohammad, 2016) through a web interface developed for this purpose, shown in figure 1.

The histograms in figure 2 show the distribution of the sentiment values obtained with direct and best–worst scaling annotation, illustrating the effectiveness of the preliminary filtering steps in ensuring that the best–worst scaling annotators were presented mainly non-neutral items.

4 Step 2: method evaluation

The methods that we compare can be divided in two categories: graph-based algorithms using the SALDO descriptors and other lexicon-based relations, and corpus-based methods using dimensions from word embeddings as features for different classifiers.

We model the sentiment associated to a word sense using a value in the interval $[-1, 1]$, where $+1$ represents a totally positive sentiment and -1 represents a totally negative sentiment. After having considered using a three-dimensional model like that of SentiWordNet (Baccianella et al., 2010), we found that experimental evidence indicated that the average overlap between positivity and negativity in the same word was very low (Rouces et al., 2018b).

[3]SALDO is freely available (under a CC-BY license) at `https://spraakbanken.gu.se/eng/resource/saldo`.

[4]The corpus is freely available (under a CC-BY license) at `https://spraakbanken.gu.se/eng/resource/gigaword`.

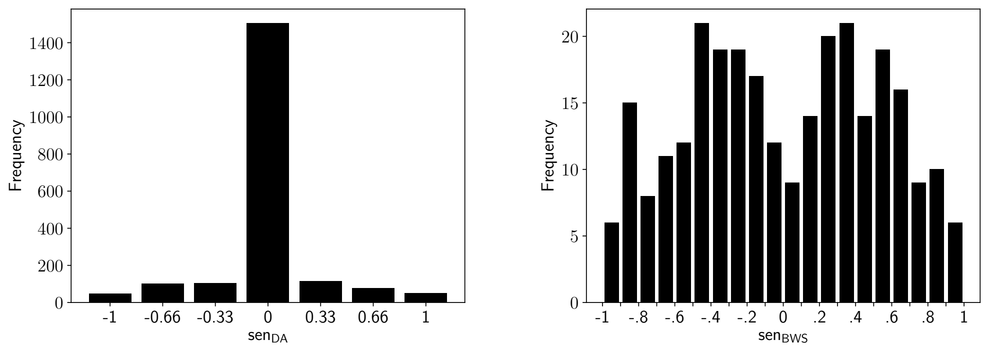


Figure 2: Histograms of the sentiment values resulting from direct annotation (left) and best–worst scaling annotation (right)

We experiment with different approaches, which we describe below. We extend the methods of Rosell and Kann (2010) and Nusko et al. (2016) and we also try a corpus-oriented approach similar to the one described by Hamilton et al. (2016). For all methods, we produce continuous scores as well as discrete labels: $+1$ (positive), 0 (neutral), -1 (negative). What is relevant about the continuous scores is not their magnitudes but the relative order that they produce. The values and their distributions depend on idiosyncrasies of the methods employed and do not necessarily resemble what would be produced by direct human annotations, but they can be fit to any desired distribution. The discrete labels are less fine-grained, but may be more appropriate for certain applications.

4.1 Inheritance over graph

Nusko et al. (2016) propose a tree traversal method on the tree defined by the primary descriptor relation between SALDO entries. This method starts with 6 seeds with a manually assigned polarity and recursively calculates the sentiment of children based on the sentiment of the parent.[5] The algorithm calculates a confidence score for each sentiment, which decreases at a constant rate with the distance to the original seed (steps of -0.25 from a confidence of 1 for the descendants of the core words), and sets a threshold of 0.5 as the lowest acceptable confidence. It also uses secondary descriptors, but only when the secondary descriptor is *inte* ‘not’, which indicates that the child and parent have opposite semantic values and therefore the sign on the sentiment value should also be inverted, or a strength modifier like *lite* ‘a little’, or *enastående* ‘outstanding’. It obtains a sentiment for 2,133 entries. Three annotators independently labeled 150 of the entries as positive, negative or neutral, and for 117 of these the annotators were in full agreement. From these a 71% precision was obtained. The 150 entries were sampled using equally sized stratification over the three confidence levels.

Our first method is a modified and extended version of this method. The extensions allows it to inherit semantic values from the secondary descriptors too. In particular, the traversal occurs over the directed acyclic graph defined by using both primary and secondary descriptors. In this way, the secondary descriptors of an entry are used not only for polarity inversion or intensification, but also their sentiment value is used, although with a lower weight.

The algorithm cannot use a simple breadth-first exploration over this graph, because for a given node, in general, some incoming neighbors will be at a different distance from the seed set than others, and the node will be reached before all the incoming neighbors have been calculated. This prevents all elements in the frontier to be expanded in a single iteration. Depth-first is not suitable for similar reasons.

Therefore, different partially-successful passes over the frontier have to be tried. However, even when doing this, the algorithm will stagnate as soon as some secondary descriptors are not reachable from the given seed words. For this reason, the algorithm applies a best-effort mechanism when the previous approach stagnates, whereby the node with the lowest possible number of unreached secondary-descriptor

[5] In Nusko et al. (2016) the seeds and their children are referred to as “core words” and “seeds” respectively.

incoming nodes is chosen, and the sentiment is calculated for this node ignoring its unreached incoming nodes, and a new pass is performed over the frontier. Primary descriptors are never ignored. A priority queue is used for the nodes with unreached secondary-descriptor incoming nodes. If it is still not possible to calculate new nodes from all their parents, the process is repeated until it is possible, or the queue and the frontier are empty.

This method outputs scores, so in order to obtain discrete labels we apply thresholds. The thresholds are obtained from the percentiles of each class in a training set obtained from sampling two thirds of the gold standard, which constitutes a very basic kind of supervised learning. The other third is used for testing.

4.2 Random paths over graphs

For our second experimental setup, we develop an adaptation of the method by Rosell and Kann (2010), who developed a Swedish sentiment lexicon using random walks over a graph of synonyms and 4 positive and 4 negative seed words. The graph was built using the *Synlex/People's Dictionary of Synonyms* (Kann and Rosell, 2005), which used Swedish–English lemma pairs concatenated with their inverse relation to generate candidate synonym pairs. The pairs were filtered by grading and then averaging the grades. The result of Synlex was 16,006 words with 18,920 weighted pairs, which were used as edges of the graph in the random walks.

Our modification consists of adapting Synlex to use SALDO word senses instead of Swedish sense-ambiguous lemmas (the adaptation was done by a trained linguist, adapting the original weights to the $(0, 1]$ interval), and the union of the following sets of edges with an heuristic weight of 0.5.

- The edges defined by primary descriptors in SALDO. This component ensures that there are no isolated nodes, since every node has one primary descriptor.
- The edges defined by secondary descriptors in SALDO.
- The edges that connect SALDO entries that have the same primary descriptor (siblings). This creates a relation which is often often tantamount to co-hyponymy.

The discrete labels are obtained using the same thresholding method as in the inheritance-based method.

4.3 Classification over word2vec

As opposed to the previous methods, which are purely lexicon-driven, the third approach is partly corpus-based. We use already existing vector representations of SALDO word senses derived from *word2vec* lemma embeddings (Johansson and Nieto Piña, 2015) by means of solving a constrained optimization problem. The vector space dimensionality is 512 and the source for the vector representations was the Gigaword corpus (see section 3). Because the elements of the vector space are SALDO word senses, and the problem solved in Johansson and Nieto Piña (2015) uses the SALDO descriptor relations, this is not a purely corpus-based approach but a mixed one. We train a logistic regression classifier (word2vec-logit) and a support vector classifier with a radial basis function kernel (word2vec-svc-rbf). All the classifiers used a one-vs-rest approach of the three-class classification. For the classifiers we used 5-fold cross-validation stratified by the (positive, neutral, negative) classes. For each fold, the SVM/RBF meta-parameters (C, γ) were estimated using 5-fold cross-validation over the training set.

The classifiers' final output are discrete labels (positive/pos, neutral/neu, negative/neg), but scores are obtained computing $p((pos) - p(neg))$, where p is the probability for a given entry to belong to the positive or negative classes. For the logit classifier, p is straightforward. For the support vector classifier, we use an extension of Platt scaling for multiple classes (Wu et al., 2004).

4.4 Results

For training and testing the different methods, we used the direct annotation gold standard developed by Rouces et al. (2018b) (see section 3), which contains of 1,998 entries from SALDO entries labeled as negative (-1), neutral (0), or positive $(+1)$. The values were averaged over three annotators (so if an entry is labeled as positive by two annotators and as neutral by one, the final value would be $2/3$).

Table 1 shows the results for each method. We employ two different sets of measures for measuring the quality of the gold standard: ones based on ranks and (Spearman rank-order correlation coefficient ($\rho \in [-1, 1]$, the p-normalized Kendall tau distance ($\tau_{\in}[0, 1]$, and Kendall's tau-b (τ_b)) others based on discrete labels (precision, recall and confusion matrix).

- The rank-based measures are the Spearman rank-order correlation coefficient (ρ) (Kokoska and Zwillinger, 2000), in the interval $[-1, 1]$ (Myers and Well, 2003), the p-normalized Kendall tau distance (τ_p) (Fagin et al., 2004) in the interval $[0, 1]$ (the one used in (Baccianella et al., 2010)), and Kendall's tau-b (τ_b) (Kendall, 1945) (the one used in (Rothe et al., 2016)). Both τ_p and τ_b are suited to handle ties —which in our case means word senses with equal sentiment values— but they do so in different ways. For additional testing, in addition to the direct annotation values in the test set, we also use more fine-grained sentiment values of 278 entries that are available as part of the same gold standard (Rouces et al., 2018b), which were obtained using Best-Worst Scaling (BWS) and also comprised in the $[-1, 1]$ range. The reason for this is that these values are more fine-grained than the Direct Annotation (DA) values (which due to the use of 3 annotators, they range over only 7 possible values), and therefore ties are less common in the gold standard, making some ranking comparison algorithms more suitable. Since the BWS values were created only for the entries annotated as non-neutral by the DA scoring ($|\text{value}| \geq 0.5$), they cannot all be used for testing (or else the training set would be too biased towards neutral elements). Therefore, the intersection of the DA test set and the entries with BWS value is used for applying the rank-based measures.

- The measures based on discrete labels are the precision and recall values for each label, derived from the confusion matrix.

	DA						confusion matrix				BWS
	ρ	τ_p	τ_b	precision	recall	acc.	GS	SL pos	SL neu	SL neg	τ_b
graph inheritance	0.39	0.39	0.38	pos: 0.28	pos: 0.26	0.82	pos	10	28	1	0.49
				neu: 0.91	neu: 0.90		neu	23	391	21	
				neg: 0.33	neg: 0.42		neg	3	12	11	
graph inheritance ext	0.33	0.42	0.32	pos: 0.22	pos: 0.21	0.81	pos	8	30	1	0.46
				neu: 0.90	neu: 0.89		neu	26	386	23	
				neg: 0.27	neg: 0.35		neg	2	15	9	
graph random paths	0.30	0.31	0.24	pos: 0.25	pos: 0.23	0.82	pos	9	29	1	0.46
				neu: 0.90	neu: 0.90		neu	26	390	19	
				neg: 0.39	neg: 0.50		neg	1	12	13	
word2vec +logit	0.47	0.21	0.38	pos: 0.37	pos: 0.54	0.84	pos	15	13	0	0.61
				neu: 0.93	neu: 0.88		neu	25	301	15	
				neg: 0.46	neg: 0.52		neg	1	11	13	
word2vec +svc /rbf	**0.55**	**0.15**	**0.45**	pos: 0.65	pos: 0.46	**0.89**	pos	13	15	0	**0.62**
				neu: 0.92	neu: 0.96		neu	7	328	6	
				neg: 0.65	neg: 0.44		neg	0	14	11	

Table 1: Results for evaluating the different methods for constructing the sentiment lexicon in Swedish. Note that the Kendall tau τ_p is a distance, and therefore it is inversely related to the Spearman correlation ρ. GS and SL stand for gold standard and sentiment lexicon respectively.

SentiWordNet is reported to have τ_p values of 0.281 and 0.231 for positive and negative dimensions (their sentiment model has 2 degrees of freedom). All our embeddings-based methods outperform both measures (τ_p is a distance, and therefore lower values are desired). (Rothe et al., 2016) reports $\tau_b = 0.654$. We obtain $\tau_b = 0.45$ when testing against the DA values, which is significantly lower. However, this probably owes to τ_b penalizing the big amount of ties in the DA values (61.95% of the possible pairs), as the method obtains $\tau_b = 0.63$ (a very close value) when testing against the BWS values, where ties are much less common (0.63%).

word sense ID	gloss	value	label
ond..4	‘bad’	-0.9959	neg
farlig..1	‘dangerous’	-0.9919	neg
villa..2	‘illusion’	-0.9878	neg
kriminalitet..1	‘criminality’	-0.9838	neg
skrämma..1	‘frighten’	-0.9797	neg
fel..2	‘wrong (a)’	-0.9757	neg
problem..1	‘problem’	-0.9716	neg
misskreditera..1	‘discredit’	-0.9675	neg
reaktionär..1	‘reactionary’	-0.9635	neg
angrepp..1	‘attack (n)’	-0.9594	neg
förfördela..1	‘wrong (v)’	-0.9554	neg
brottslig..1	‘criminal (a)’	-0.9513	neg
risk..1	‘risk (n)’	-0.9473	neg
steka..2	‘dismiss’	-0.9432	neg
absurd..1	‘absurd’	-0.9391	neg
server..1	‘server’	-0.0426	neu
ställe..1	‘place (n)’	-0.0385	neu
förhållande..1	‘relationship’	-0.0345	neu
markägare..1	‘land owner’	-0.0304	neu
radio..1	‘radio’	-0.0264	neu
sälja..1	‘sell’	-0.0223	neu
offentlighet..1	‘public (n)’	-0.0183	neu
manus..1	‘manuscript’	-0.0142	neu
positiv..2	‘positive (charge)’	-0.0101	neu
älvstrand..1	‘riverside’	-0.0061	neu
molnet..1	‘the cloud’	-0.0020	neu
flagga..2	‘flag (v)’	0.0020	neu
reglera..2	‘regulate’	0.0061	neu
resenär..1	‘traveller’	0.0101	neu
läge..1	‘position (n)’	0.0142	neu
inkomstskatt..1	‘income tax’	0.0183	neu
kurator..1	‘therapist’	0.0223	neu
land..2	‘field, plot (n)’	0.0264	neu
distrikt..1	‘district’	0.0304	neu
likartad..1	‘similar’	0.0345	neu
fusion..3	‘fusion (music)’	0.0385	neu
surdeg..1	‘sourdough’	0.0426	neu
uppryckning..1	‘improvement, recovery’	0.9635	pos
god..2	‘tasty’	0.9675	pos
riktig..2	‘genuine’	0.9716	pos
överraska..1	‘surprise (v)’	0.9757	pos
hjälpa..1	‘help (v)’	0.9797	pos
välsignelse..1	‘blessing’	0.9838	pos
stöd..2	‘support, aid (n)’	0.9878	pos
bra..3	‘good’	0.9919	pos
äga..3	‘rock, excel (v)’	0.9959	pos
fantastisk..1	‘fantastic’	1.0000	pos

Table 2: Examples of sentiment values and labels.

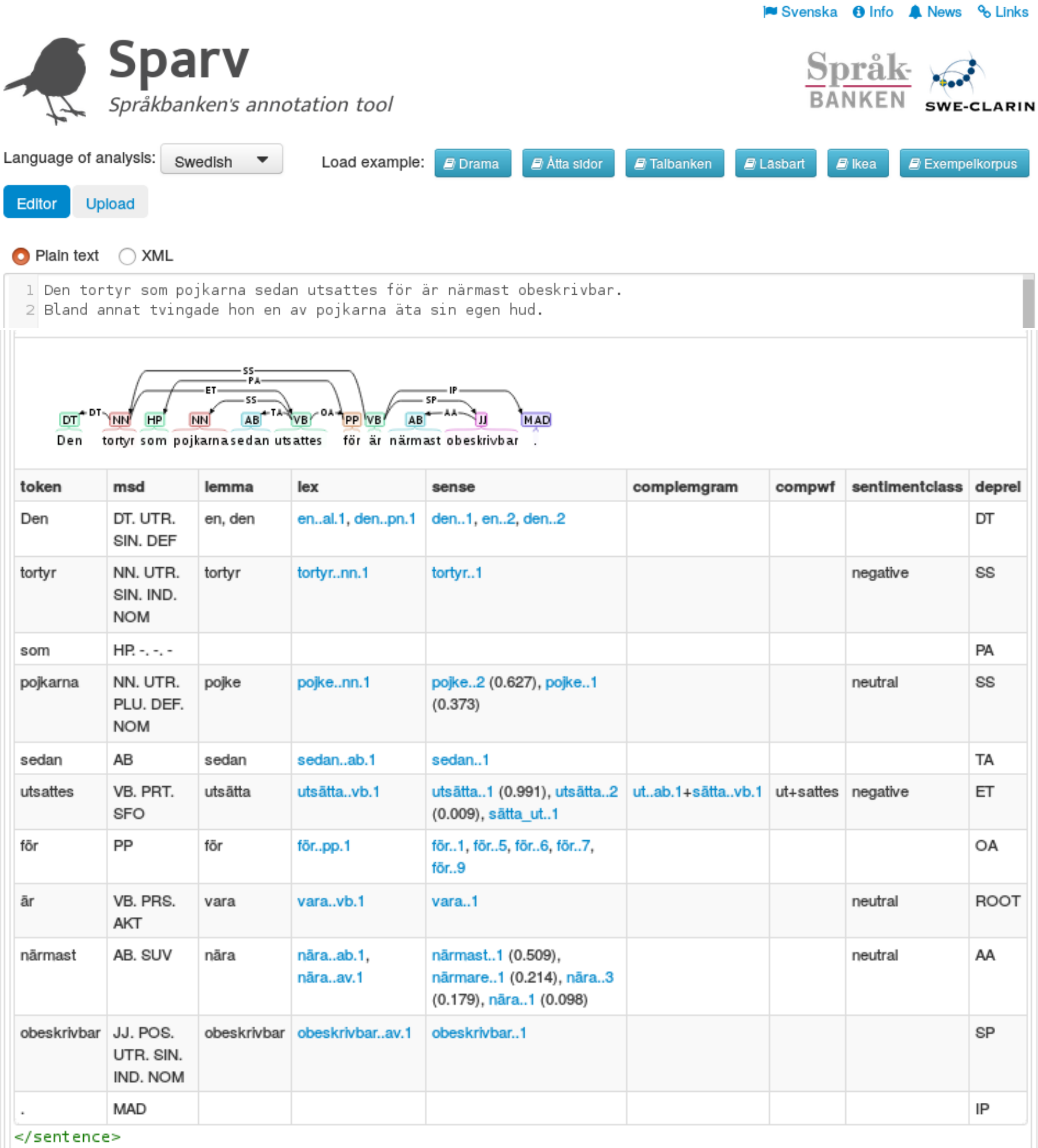
The method word2vec-svc-rbf performed consistently better than the rest, and therefore we have used it for the input to the manual curation step. Table 2 shows some examples of sentiment scores obtained using this method.

5 Step 3: Manual curation

In order both to get a better sense for the accuracy of the word2vec-svc-rbf method and in order to enhance the quality of the resulting dataset, this has been manually curated, as described in the following.

The outcome of the automatic sense-label assignment was a list of SALDO word senses labelled with a score in the interval $[-1, 1]$ assigned by the word2vec-svc-rbf method, and a sentiment label – one of -1, 0 or $(+)1$ – computed on the basis of the score. The resulting list contained 69,785 word senses, out of which 5,118 were labeled as non-neutral (3,508 negative and 1,610 positive items).

For the manual curation, we took all non-neutral items, plus the top 2,500 neutral items as determined



token	msd	lemma	lex	sense	complemgram	compwf	sentimentclass	deprel
Den	DT. UTR. SIN. DEF	en, den	en..al.1, den..pn.1	den..1, en..2, den..2				DT
tortyr	NN. UTR. SIN. IND. NOM	tortyr	tortyr..nn.1	tortyr..1			negative	SS
som	HP. -. -. -							PA
pojkarna	NN. UTR. PLU. DEF. NOM	pojke	pojke..nn.1	pojke..2 (0.627), pojke..1 (0.373)			neutral	SS
sedan	AB	sedan	sedan..ab.1	sedan..1				TA
utsattes	VB. PRT. SFO	utsätta	utsätta..vb.1	utsätta..1 (0.991), utsätta..2 (0.009), sätta_ut..1	ut..ab.1+sätta..vb.1	ut+sattes	negative	ET
för	PP	för	för..pp.1	för..1, för..5, för..6, för..7, för..9				OA
är	VB. PRS. AKT	vara	vara..vb.1	vara..1			neutral	ROOT
närmast	AB. SUV	nära	nära..ab.1, nära..av.1	närmast..1 (0.509), närmare..1 (0.214), nära..3 (0.179), nära..1 (0.098)			neutral	AA
obeskrivbar	JJ. POS. UTR. SIN. IND. NOM	obeskrivbar	obeskrivbar..av.1	obeskrivbar..1				SP
.	MAD							IP

Figure 3: Sentiment annotation in Sparv

by corpus frequency in the Gigaword Corpus (described in section 4.3 above). The curation consisted simply in checking the sentiment labels for all the 7,618 word senses in the resulting list, and correcting them if needed.

The resulting list has more neutral, and consequently less positive and negative items than the original: 2,640 neutral, 1,584 positive, and 3,394 negative items. A detailed analysis of the differences is still pending.

6 Summing up and looking ahead

We have described the development of SenSALDO, a Swedish sentiment lexicon containing 7,618 word senses as well as a full-form version of this lexicon containing 65,953 items (text word forms), for the SWE-CLARIN toolbox.[6]

Merely providing the downloadable lexicon is generally not sufficient for the user community targeted by CLARIN. For this reason, as a first step in the direction of more user-friendliness we have included

[6] The first version of this resource – SenSALDO v. 0.1 – is freely available for downloading under a CC-BY license from Språkbanken Text: `https://spraakbanken.gu.se/eng/resource/sensaldo`

sentiment annotation based on SenSALDO in Språkbanken Text's online annotation tool *Sparv*[7] (see figure 3) and the new document-oriented infrastructure component *Strix*, with the aim to provide document filtering based on sentiment (see figure 4).[8]

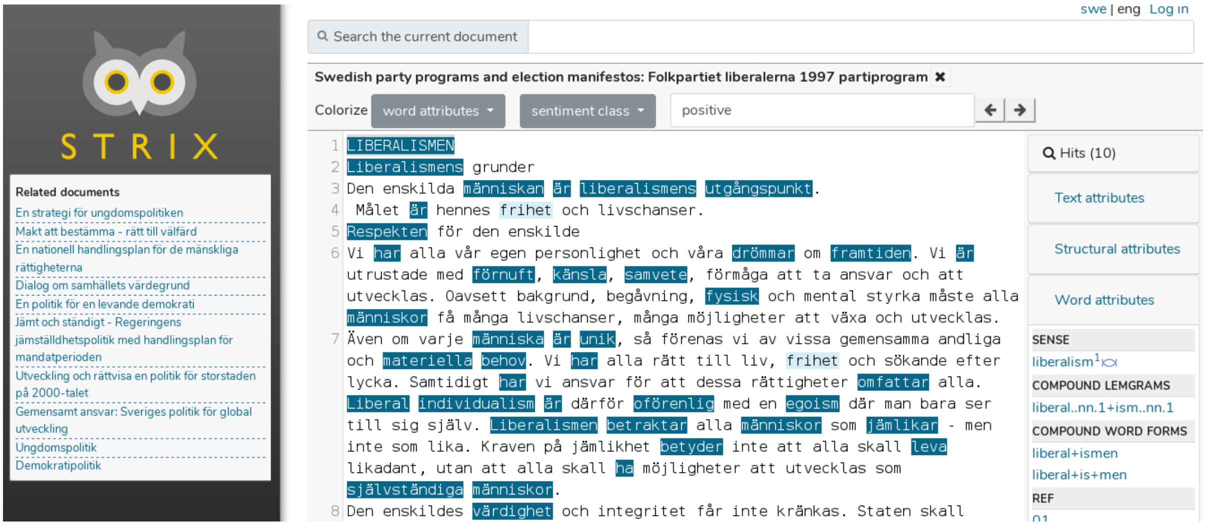


Figure 4: Sentiment annotation in Strix

In order to provide sentiment analysis as a standard tool in the SWE-CLARIN toolbox, we are currently pursuing two lines of development.

Firstly, there is a natural extension to the present version of SenSALDO, namely one where we ensure that all sentiment values for all lemmas present in it are accounted for. Because of the way SALDO is organized and because the lexical units considered in the work described here are *word senses*, there is no guarantee that this will be the case. For example, SenSALDO contains the information that the word sense `suga..2` carries a negative sentiment, which is correct (it means 'suck', as in *this situation sucks*). When generating the full-form version of SenSALDO, all forms of the lexeme *suga (v)* are given the sentiment label 'negative' (-1), but in fact the SALDO word sense `suga..1` 'suck (with mouth or instrument)' links to the same lexeme, and this word sense is arguably neutral wrt its sentiment value, so that SenSALDO ought to tell us that all forms of *suga (v)* occur with both negative and neutral sentiment. Complementing SenSALDO to reflect this is a straightforward enhancement which we are planning to implement in the next release of the resource.

Secondly, any sentiment-analysis software will require some means of evaluation, regardless of whether it is based on a sentiment lexicon or not. Pure machine-learning approaches to sentiment analysis rely on annotated training data. To meet both these needs, we are now in the final stages of preparing a Swedish gold-standard corpus for aspect-based sentiment analysis, consisting of approximately 1.5 million words from three different sources, two newspapers belonging to opposite ends of the political left–right spectrum, and an online discussion forum. The newspaper material consists of editorials and opinion pieces, and the topic for the whole corpus is immigration. This work is described in Rouces et al. (forthcoming).

An additional obviously very useful extension would be to add sentiment values to our diachronic lexicons (Borin and Forsberg, 2017), in order to support for instance historical research and studies in conceptual history such as that by Viklund and Borin (2016). This must remain a plan for the future, however.

Finally, we are in the process of using SenSALDO for developing both a sentence-level and an aspect-based sentiment analysis system for Swedish text, combining the polarity of terms according to syntax-based rules of compositionality. This will be complemented with information derived from annotated

[7]https://spraakbanken.gu.se/sparv
[8]https://spraakbanken.gu.se/strix/

corpora, which can cover cases that the lexicon-based approach cannot cover either due to limited coverage or non-compositional expressions.

With the already accomplished work presented above and the ongoing activities described in this section, we will soon be able to offer sentiment-analysis tools to Swedish researchers which are on a par with what is available for English.

Acknowledgements

This work has been supported by a framework grant (*Towards a knowledge-based culturomics*;[9] contract 2012-5738) as well as funding to Swedish CLARIN (*Swe-Clarin*;[10] contract 2013-2003), both awarded by the Swedish Research Council, and by infrastructure funding granted to Språkbanken by the University of Gothenburg.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC 2010*, pages 2200–2204.

R. Alexander Bentley, Alberto Acerbi, Paul Ormerod, and Vasileios Lampos. 2014. Books average previous decade of economic misery. *PLoS ONE*, 9(1):e83147.

Lars Borin and Markus Forsberg. 2017. A diachronic computational lexical resource for 800 years of Swedish. In Caroline Sporleder, Antal van den Bosch, and Kalliopi Zervanou, editors, *Language technology for cultural heritage*, pages 41–61. Springer, Berlin.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: A touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of ACL 2014*, pages 383–389. ACL.

Ann Devitt and Khursid Ahmad. 2013. Is there a language of sentiment? An analysis of lexical resources for sentiment analysis. *Language Resources and Evaluation*, 47(4):475–511.

Johannes C. Eichstaedt, Hansen Andrew Schwartz, Margaret L. Kern, Gregory Park, Darwin R. Labarthe, Raina M. Merchant, Sneha Jha, Megha Agrawal, Lukasz A. Dziurzynski, Maarten Sap, Emily E. Weeg, Christopherand Larson, Lyle H. Ungar, and Martin E. P. Seligman. 2015. Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26(2):159–169.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The Swedish culturomics gigaword corpus: A one billion word Swedish reference dataset for NLP. In *Proceedings of the* From Digitization to Knowledge *workshop at DH 2016, Kraków*, pages 8–12, Linköping. LiUEP.

Andrea Esuli and Fabrizio Sebastiani. 2007. Random-walk models of term semantics: An application to opinion-related properties. *Proceedings of LTC 2007*, pages 221–225.

Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. 2004. Comparing and aggregating rankings with ties. In *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, pages 47–58, New York, NY, USA. ACM.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of EMNLP 2016*, pages 595–605, Austin. ACL.

Richard Johansson and Luis Nieto Piña. 2015. Embedding a semantic network in a word space. In *Proceedings of NAACL-HLT 2015*, pages 1428–1433, Denver. ACL.

Viggo Kann and Magnus Rosell. 2005. Free construction of a free Swedish dictionary of synonyms. In *Proceedings of NODALIDA 2010*, Joensuu. University of Eastern Finland.

[9] `https://spraakbanken.gu.se/eng/culturomics`
[10] `https://sweclarin.se/eng`

Maurice G Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, pages 239–251.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of NAACL 2016*, pages 811–817, San Diego. ACL.

Stephen Kokoska and Daniel Zwillinger. 2000. *Standard Probability and Statistics Tables and Formulae*. Chapman & Hall / CRC.

Saif Mohammad and Peter Turney. 2010. Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles. ACL.

Jerome L Myers and A. (Arnold) Well. 2003. *Research Design and Statistical Analysis* . Mahwah, N.J. : Lawrence Erlbaum Associates, 2nd ed edition.

Bianka Nusko, Nina Tahmasebi, and Olof Mogren. 2016. Building a sentiment lexicon for Swedish. In *Proceedings of the* From Digitization to Knowledge *workshop at DH 2016, Kraków*, pages 32–37, Linköping. LiUEP.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Magnus Rosell and Viggo Kann. 2010. Constructing a Swedish general purpose polarity lexicon: Random walks in the People's dictionary of synonyms. In *Proceedings of SLTC 2010*, pages 19–20, Stockholm. KTH.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. *arXiv preprint arXiv:1602.07572*.

Jacobo Rouces, Lars Borin, Nina Tahmasebi, and Stian Rødven Eide. 2018a. Defining a gold standard for a Swedish sentiment lexicon: Towards higher-yield text mining in the digital humanities. In *Proceedings of DHN 2018*, pages 219–227, Aachen. CEUR-WS.org.

Jacobo Rouces, Nina Tahmasebi, Lars Borin, and Stian Rødven Eide. 2018b. Generating a gold standard for a Swedish sentiment lexicon. In *LREC 2018*, pages 2689–2694, Miyazaki. ELRA.

Jacobo Rouces, Lars Borin, and Nina Tahmasebi. forthcoming. Tracking attitudes towards immigration in Swedish media. In *Proceedings of DHN 2019*, Aachen. CEUR-WS.org.

Rachele Sprugnoli, Sara Tonelli, Alessandro Marchetti, and Giovanni Moretti. 2016. Towards sentiment analysis for historical texts. *Digital Scholarship in the Humanities*, 31(4):762–772.

Mike Thelwall. 2017. Sentiment analysis. In Luke Sloan and Anabel Quan-Haase, editors, *The SAGE Handbook of Social Media Research Methods*, pages 545–556. SAGE, London.

Jon Viklund and Lars Borin. 2016. How can big data help us study rhetorical history? In *Selected papers from the CLARIN annual conference 2015*, pages 79–93, Linköping. LiUEP.

Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005.

Anis Yazidi, Hugo Lewi Hammer, Aleksander Bai, and Paal Engelstad. 2015. On enhancing the label propagation algorithm for sentiment analysis using active learning with an artificial oracle. In Leszek Rutkowski, Marcin Korytkowski, Rafal Scherer, Ryszard Tadeusiewicz, Lotfi A. Zadeh, and Jacek M. Zurada, editors, *Artificial Intelligence and Soft Computing*, pages 799–810, Cham. Springer International Publishing.