Building a Swedish sentiment model

Michelle Ludovici
Department of Computer and Systems Sciences, (DSV), Stockholm University
Email: michelleludovici@gmail.com

***Abstract*—The proposed research is within the machine learning and data mining research area of sentiment analysis. Sentiment analysis is the analysis of attitudes, emotions and opinions. No open-source Swedish sentiment lexicon exists at the time being. There are knowledge gaps for evaluation of sentiments in Swedish texts, for methods to develop sentiment analysis techniques adapted to the Swedish language, as well as for the development of tools for Swedish sentiment analysis. In this study, the authors design, demonstrate and evaluates a Swedish general purpose sentiment lexicon and a model for word- and sentence-level sentiment analysis. The lexicon is constructed with seed word expansion of adjectives, adverbs and verbs. The sentiment model calculates word and sentence sentiments with the Naive Bayes maximum likelihood method (NB) and a simple averaging method (AM) that both rely on the created lexicon at the basis. The model takes account of negation and Swedish idiomatic expressions with verbs. The created lexicon and the word and sentence sentiment analysis perform comparable to results of previous research.**

I Introduction

The proposed research is within the machine learning and data mining research area of sentiment analysis.

Sentiment analysis is the analysis of attitudes, emotions and opinions, i.e. subjective expressions that can be extracted from written text and classified. The usefulness of sentiment analysis stems from the premise that lexical items carry affective and attitudinal information [Karlgren et.al. [1]]. Classification of sentiments is often within the categories ‘positive’, ‘negative’ and ‘neutral’. In addition, sentiment analysis may be also be used to extract "more complex information content such as emotional states and opinion content." [Rice and Zorn [2]] In research there are many names for the same concept of sentiment analysis, such as ‘opinion extraction’, ‘opinion mining’, ‘sentiment mining’, and ‘subjectivity analysis’.

The extraction of such information might lead to insights into consumers attitude towards a product which has direct applications in advertisement and recommendations to the individual consumer, for the development of adapted search engines and for propagation trends of product information. Sentiment analysis may also be used for predicting investment trends and for political forecasts. For example, Tweets have been used to predict stock market trends [Bollen, Mao and Zeng [3]] and public opinion monitoring about politicians can be used in order to calculate the preference of one candidate or party over others [Conover et.al. [4]]. Thematic opinion search with sentiment analysis could be used for monitoring security and even medicinal mood tracking for surveillance of depression patients would be possible.

II Previous Research

The article "Sentiment Analysis and Opinion Mining: A Survey" by Vinodhini and Chandrasekaran [5] shows that extensive research has been done in the field of sentiment analysis. According to Medhat et.al. [6], there exist two major techniques for sentiment analysis: supervised and unsupervised learning, with and without a sentiment lexicon as the basis.

Supervised learning makes use of sophisticated programs to infer rules about the polarity of words. These algorithms learn by training on a validated set of words and their polarities, called a semantic lexicon or semantic dictionary. Classifiers used in supervised learning for sentiment analysis encompass amongst other probabilistic classifiers, rule based classifiers, linear classifiers and decision tree classifiers [Medhat et.al. [6], Pand and Lee [7], Psomakelis et.al. [8], Santos and Gatti [9], Fuchs [10]]. Bootstrapping is often used to create labelled data to train supervised learning algorithms. Bootstrapping methods for learning require a small amount of supervision to seed the learning process.

Unsupervised learning in sentiment analysis often has as first task to build a sentiment lexicon, then attempts classification with the lexicon as basis. This approach has been employed by Turney to classify reviews into “thumbs up” or “thumbs down”. Turney

uses just two seed words[1], ”excellent” and “poor”, and calculates the polarity of a phrase with an algorithm[2] that measures the distance of pairs of words to both of the two seed words [Turney [11]]. For automatic creation of a lexicon, often a gold-standard is used and expanded, for example by bootstrapping algorithms [Turchi and Balahur [12], Taboada et.al. [13], Wei et.al.14, Neviarouskaya, Prendinger and Ishizuka [15]]. Expanding sentiment lexicons by growing an initial set with the help of dictionaries or thesauri has the disadvantage that opinion words are being classified without context. However, words and sentences may have different polarity in different contexts. For instance, as Pang and Lee point out, the sentence "We recommend reading the book" is positive for a book review, but negative in the context of a movie review [16]. Unsupervised corpus based techniques for building a sentiment lexicon take context into account when classifying sentiments. Often corpus based techniques use syntactic patterns and frequencies to classify opinion words. The premise is that similar opinion words appear frequently together. These statistical approaches to corpus based lexicon induction find co-occurrence frequencies of unlabelled words and labelled words with large amounts of texts (corpus) to train. Classification methods for corpus based semantic classification are bag-of-words, n-grams, pointwise mutual information, naive bayes, bayes network, maximum entropy and support vector machines amongst other [Psomakelis et.al. [8], Rice and Zorn [2], Cao and Zukerman [17]]. Corpus approaches are often combined with feature extraction by means of POS-tagging and heuristic linguistic rules. Example of such rules can be that the connector "but" indicates inverted polarity after the connector, just as a negation inverses polarity as well.

Domain unspecific lexicons usually perform worse than lexicons trained on a certain domain. However, general lexicons have the advantage to be able to be adapted to a certain domain in hindsight. Examples of accuracy of general lexicons are Turneys lexicon that achieved 66% accuracy [11]], 65.8% for Read and Carroll [18] and 77.1% to 80% for Lambov, Pais and Dias [19].

Rothfels and Tibshirani attempt unsupervised learning without a lexicon. They use the linguistic intuition that positive items occur frequently after negation, for instance in the example “not good”, and they occure even more frequently without negation. Their method statistically classifies items according to this assumption, but the authors find that this type of classification is “consistently unimpressive” [Rothfels and Tibshirani [20]]. Other unsupervised methods include automatic clustering of sentiments via linguistic heuristic rules, such as rules for negation and conjunctions[3], after which polarity is assigned, whereas positive orientation is assigned to the class with the highest average frequency [Hatzivassiloglou et.al. [21]]. Vohra and Teraiya [22] compared supervised and unsupervised methods for sentiment analysis and concluded that "supervised machine learning techniques have shown relative better performance than the unsupervised lexicon based methods".

Ensemble techniques and semi-supervised methods for sentiment analysis are a mixture of both supervised and unsupervised methods and often result in high precision as they employ several sentiment classification techniques or algorithms in combination [Balahur and Perea-Ortega [23], Cao and Zukerman [17], Kang, Yoo and Han [24], Chikersal, Poria and Cambria [25]].

III Research problem

As described above, a sentiment lexicon is often the basis for sentiment analysis. For the English language many sentiment lexicons exist and some attempts at creating lexicons in the most widely spoken languages such as Chinese and Spanish have been made [Balahur and Perea-Ortega [23], Oye [26], Vincze and Bestgen [27], Steinberger et.al. [28]]. However, an extensive search in databases, articles, and on the Internet in general have shown that no open-source Swedish sentiment lexicon exists at the time being, which was confirmed by the institute Språkbanken. There are knowledge gaps for evaluation of sentiments in Swedish texts, for methods to develop sentiment analysis techniques adapted to the Swedish language, as well as for the development of tools for Swedish sentiment analysis.

The problem that this thesis addresses is the lack of a Swedish sentiment lexicon on which to build models for sentiment analysis of Swedish words and sentences.

A. Aim and objectives

The aim of this research study is to build, evaluate and test a model for determining word and sentence

[1]seed words are words which are at the core either ’positive’ or ’negative’ independently of context

[2]Pointwise Mutual Information and Information Retrieval

[3]such as “and”, “but”

sentiments with a general purpose Swedish sentiment lexicon at the basis.

The objectives of the study are to:

- Build a general purpose Swedish sentiment lexicon by seed word expansion according to Kim and Hovy [29]
- Combine the lexicon with classifiers for Swedish sentence sentiment analysis
- Evaluate and discuss which change of parameters may improve the sentiment analysis

B. Research questions

This research study attempts to answer the following main research question:

- What is the performance of the general purpose lexicon-based Swedish sentiment classification model measured with accuracy, precision, recall and F1-score[4]

The sub-questions are as follows:

- How accurate is the created Swedish sentiment model for sentiment analysis of Swedish sentences for five different types of texts?
- How may parameters for the created model be tweaked in order to improve it?

C. Contribution

This research contributes with a first prototype of a functional open source Swedish sentiment classification model. It takes account of negation and typical Swedish idiomatic expressions with verbs[5].

It is expected that this research will increase knowledge on which factors are important in refining a rudimentary Swedish sentiment model based on a general purpose lexicon. It may also lead to insights on how to conduct more advanced sentiment analysis of Swedish texts, documents and for specific domains.

D. Limitations of the study

The study will implement a Swedish sentiment lexicon and evaluate it in comparison to similar English lexicons and human classification. Sentiment analysis will be done on word and sentence level only.

[4]for explanation of the terms, see appendix A

[5]An example of such an idiomatic expression is 'vara ute och cykla' to describe that one is confused or has misunderstood. Only idiomatic expressions with verbs are considered and included in the verb seed list

IV Methodology

Since there is no previous research for Swedish in this field, the authors adopt an approach which starts at the easiest level of sentiment analysis and incrementally increases the difficulty of the analysis from word to sentence level. Therefore this study is inspired by Kim and Hovy [29] since it is the only study found which starts at basic level and advances incrementally in its approach. This study attempts to recreate the seed word expansion by Kim and Hovy [29] for Swedish words. The report then diverges in the development of the naive bayes and averaging approach in order to account for negation and Swedish idiomatic expressions with verbs. Idiomatic expression are fixed expressions used in everyday speech, for instance 'ta hänsyn till' (take account of'). Only idiomatic expressions with verbs are considered and included in the verb seed list, i.e. treated as verbs. The sentence classifier is then tested on newspaper, blog, twitter, prose and thesis paper texts instead of the DUC 2001 corpus that Kim and Hovy [29] chose. This is done in order to test a variety of the language strata.

The general method for conducting research follows the design science research methodology process model (DSRM Process Model), by Peffers et.al. [30], see appendix **B**.

A. Problem identification and motivation

The DSRM process starts with problem identification and motivation. As described in section **III**, sentiment lexicons are often the basis for sentiment analysis, but no freely available Swedish sentiment lexicon exists, neither at word- or sentence-level. In order to be able to gradually built up the process of lexicon improvement, the model presented in this study starts at the simplest level: sentiment analysis at the word and sentence level. The model is tested for several types of texts in order to create a baseline for the accuracy of a first general purpose lexicon-based classification model.

1) Choice of lexicon creation and sentence classification method

In literature, several approaches for building a sentiment lexicon are discussed. To create a lexicon

manually is very time expensive. Translation methods where an English sentiment lexicon is translated into a Swedish sentiment lexicon or where Swedish texts are translated to English and then analysed for sentiments, induces errors due to the differences in linguistic features in different languages [Hedlund, Pirkola and Järvelin [31]]. Therefore the translation method was not chosen in this study.

Another approach is the use of previously human labelled data as a gold standard to create a lexicon automatically from the data [16]. Unfortunately no gold standard exists for Swedish.

The last and most used approach is to create a seed-list of positive and negative words by hand and then expand it with unsupervised algorithms in order to collect new terms for the dictionary [Turney [11], Turchi and Balahur [12], Taboada et.al. [13], Wei et.al. [14], Neviarouskaya, Prendinger and Ishizuka [15]]. Since this method has proven to be successful in previous research, it is also used for this study.

Naive Bayes maximum likelihood estimation (NB) and an averaging method (AM) are used for sentence classification. Although studies show that support vector machines and other approaches are more successful in sentence classification, NB and AM are a good indicator of the base accuracy that can be achieved in sentence classification. In addition the chosen methods are relatively fast implemented and easy to tweak in order to determine which parameters most influence sentiment analysis for Swedish. Finally, Kim and Hovy [29] use similar methods and therefore it is assumed that the results of this study will be comparable, which gives an even better estimation of Swedish sentence classification accuracy on which to build more advanced methods in later studies.

B. Objectives for a solution

The second step of the DSRM requires to define the objectives for a solution. Objectives for a first model of general sentiment analysis at the word level are:

- a recall of over 90% for adjective and over 80% for verb word classifications, comparable to [29]
- a sentence level sentiment analysis accuracy between 66% and 80%, which corresponds to accuracies delivered by previous models based on general English sentiment lexicons [Turney [11], Read and Carroll[18], Lambov, Pais and Dias [19]]

C. Design and development

The third step in the DSRM process model is the design and development of the artifact.

The design and development of seed words and seed word expansion is described in section **V-A**. The seed words are chosen according to previous studies and classified by three annotators. The seed word expansion is described in detail in section **V-A2**. It is based on a program that extracts synonyms and antonyms of the seed words from the website 'www.synonymer.se'. The design and development choice for the classification of new words is described in section **V-B**. This study uses the Naive Bayes maximum likelihood estimation and an averaging algorithm to classify new words. Finally the design choice for sentence level sentiment classification is described in section **V-C**. Sentence classification works by extracting sentiment-bearing words and classifying them with the above mentioned algorithms. The word classifications are aggregated to one sentence level classification. It includes an idiomatic expression handler and a negation handler. An overview of the sentiment model is given in appendix **D**.

D. Demonstration

Stage four of the DSRM is to demonstrate the use of the created artifact for problem solving. The prototype of the sentiment lexicon was built in Java in the IntelliJ IDEA [32], see appendix **C**. The prototype was extensively tested. After the prototype of the lexicon was functional, the word classification algorithms were implemented. The algorithm use data from the created lexicon in order to classify new words. The algorithms were tested on several new words. Finally the classification algorithm for sentence level sentiment analysis is demonstrated by testing it on different types of texts: newspaper, blog, novel, twitter, online forum and thesis paper.

E. Evaluation

Activity five in the DSRM model is evaluation. In this step, it is measured how well the artifact performs in solving a problem and the objectives of a solution are compared to the actual results. The evaluation methods are described in detail in section **VI**. Evaluation is six-fold: 1) IAA measure on seed word list 2) Accuracy of seed word extraction 3)

Accuracy of Naive Bayes (NB) and Average Method (AM) word extraction in comparison to [29] 4) Comparison of word classifications by the General Inquirer lexicon to the created Swedish lexicon 5) IAA of sentence classification 6) Accuracy of NB and AM sentence classification.

F. Communication

Finally, communication is the sixth activity of the DSRM model. In this part the problems, the influence of different parameters, the utility and novelty, design and effectiveness of the lexicon and classification model are discussed with regard to the above mentioned evaluation points.

G. Research Ethics

Research should be ethical, i.e. the researcher needs to respect the rights and dignity of participants, avoid harm, and to operate with honesty and integrity [Denscombe [33]]. All three principles in Denscombe [33] are observed. No private data or other personal information has been used for this study, since only single words without context are classified anonymously into the generic categories 'positive' and 'negative'. According to Denscombe [33], the impersonal and uncontroversial nature of the data gathered is grounds for not seeking informed consent. No incentives for any participation have been offered and the two volunteers who classified words by hand are treated anonymously and confidentially during and after the study. No software licensing agreement has been broken by this study. The nature of the data gathered does not infringe on the Swedish laws for data protection regarding personal information such as the 'Personuppgiftslagen' (PUL) and the 'Offentlighets- och sekretesslag'.

V The Artifact

In this chapter the design and development of the artifact are presented.

A. The seed word list

The sampling of the initial seed word list and the expansion of the seed word list are described.

1) Sampling the initial seed word list

Kim and Hovy [29] assemble a relatively small list of initial seed words by hand. The seed word lists consist only of adjectives, adverbs and verbs since those elements are sentiment-bearing, whereas other lexical elements, such as nouns, may change sentiment depending on the context [21]. The approach where adjectives, adverbs and verbs define sentiment, but nouns define the topic of the sentiment is appropriate to understand the narrators judgement about a topic. For example in the sentence 'This stupid dog is so ridiculous that even they feel it is fun', the narrators sentiment is negative ('stupid', 'ridiculous') about a positive topic ('fun').

The initial seed word list consists of nominal data, i.e. a collection of words and their corresponding polarity 'negative' or 'positive'. The choice of seed words affects the overall outcome of the sentiment lexicon and thereby the sentiment analysis. Kozareva and Hovy [34] choose seed words at random, Igo and Riloff [35] by picking the most frequent terms or Kim and Hovy [29] by asking humans.

It is hard to set a data collection frame for seed words, since there is no agreement or proof of which amount and which kind of seed words are most adequate for expanding a sentiment lexicon. Research shows that some approaches of sentiment analysis suffer from deterioration when using seeds that are infrequent and ambiguous. [Kozareva and Hovy [36]]

Purposive hand-sampling is used when the researcher selects particular data on purpose because "they are seen as instances that are likely to produce the most valuable data" [Denscombe [33, pp.16]]. Since adult humans are, at the time of writing, best able to judge if the sampled data is 'negative' or 'positive' regardless of context, purposive sampling is employed in this study. An initial list of positive and negative adjectives, adverbs and verbs were constructed by choosing the same seven adjectives as Turney [37] and adding other adjectives to the list to arrive at an initial seed list of 15 positive and 19 negative adverbs and adjectives, as well as 23 positive and 21 negative verbs, in order to have a seed word sample comparable to Kim and Hovy [29]. The remaining words of the seed list were added with the criteria that the list words should be unambiguous and commonly used words. In comparison, Hu and Liu [38] use 30 of the most common adjectives from WordNet as seed words, all of which are also used in the seed word list for this study. Kim and Hovy [29] uses three human annotators to rate their initial seed words and their seed word expansion. SentiWordNet, a large English sentiment corpus, has been evaluated by using only five human annotators.

Three annotators were therefore deemed appropriate for evaluating the total of 77 initial positive and negative adjectives, adverbs and verbs. The words on which the annotators agreed on were chosen as input data for the lexicon expansion. Although the process of seed word choice is rather subjective and therefore inherits cognitive bias, the measure of agreement strength in inter-annotator classification is an indicator of the justifiability of the gold-standard set for seed words, according to [Okasha 39, pp.35]].

2) Expansion of initial seed word list

The seed word list is expanded with a self-made Java program written in the IntelliJ Java IDEA [32]. For each word in the initial seed list, the program extracts synonyms and antonyms from the website 'www.synonymer.se' [40] and adds them to the initial seed word list. The list thus allows inclusion of several instances of one word, i.e. duplicates. For this process it is assumed that positive words have mostly positive synonyms and negative words mostly negative synonyms. Synonyms receive the same polarity as their seed parent and antonyms receive reversed polarity. Two iterations are used to expand the initial seed word list[6].

The effort for seed word expansion is low since the process of seed word extraction is automatized. www.synonymer.se is a database of currently 95 000 words based on Göran Walters synonymordbok (later published as Bonniers Synonymordbok), Folkets synonymlexikon Synlex and Saldo (Språkbanken) amongst other. The website is regularly maintained by the owners Sinovum Media and user input is evaluated before being used [40].

B. Classification of new words

Two methods for the classification of new words, i.e. words that do not exist in the expanded seed word list, are implemented and compared in their accuracy. The classification methods use the created expanded sentiment lexicon and are the same as in Kim and Hovy [29] in order to be able to compare results.

The first algorithm is a NB approach. For each new word, it calculates the strength of synonym polarity of a class c by maximum likelihood estimation. It then assigns the classification according to the strongest polarity calculated. The second algorithm is a AM approach. It classifies new words by calculating the majority occurrence in a class c for each synonym of the new word. It then takes the arithmetic average [33] of the polarity frequencies of all synonyms of a class c and assigns the new word the class of the higher averaged polarity.

All algorithms are written in Java in the IntelliJ environment. The classes for classification are nominal and discrete: 'positive' and 'negative'. The preprocessing stage for new words requires stemming, i.e. a reduction of the new word to its root form[7]. This is automatically achieved by using the swemalt, a malt parser adapted to the Swedish language [Johan, Nilsson and Nivre [41]].

Both algorithms for seed word expansion and sentiment classification are linear and therefore the computation effort comes at low cost. The NB and AM classification algorithms for new words are a part of the sentence classification model.

1) Naive Bayes classification algorithm

The NB algorithm used by Kim and Hovy [29] is able to handle nominal data and relatively uncomplicated to implement for calculating positive and negative sentiment strength. For each new adjective, adverb and verb that is not in the lexicon, the program will get a synonym set of the word from synonymer.se and classifies the new words as follows:

For a given new word w, we want to know

$$\begin{aligned} \hat{c} &= argmax P(c|w) \\ &\cong argmax P(c|syn_1, syn_2, ..., syn_n) \end{aligned} \tag{1}$$

where ĉ is the estimation of a sentiment category 'positive' or 'negative' and syn_n are the synonyms of w. According to the NB theorem:

$$\begin{aligned} \hat{c} &= argmax P(c|w) \\ &= argmax P(c) P(w|c) \\ &= argmax P(c) P(syn_1, syn_2, ..., syn_n|c) \\ &= argmax P(c) \prod_{k=1}^{l} P(f_k|c)^{(f_k, synset(w))} \end{aligned} \tag{2}$$

$P(c)$ is the prior probability of the class c, i.e. the number of words in one class c divided by the total number or words in the lexicon. l is the length of the lexicon word list. f_k represents an instance of a word in the lexicon, for example one instance of 'bra', and if f_k is present in the lexicon, it might also be a member of the synonym set of w. $P(f_k|c)$ is the likelihood of obtaining an instance f_k in the synonym set belonging to class c. The NB assumption is applicable here: independence is assumed for the probabilities $P(f_k|c)$. $count(f_k, synset(w))$ is a

[6]Kim and Hovy [29] uses two iterations

[7]for example the verb 'plays' in 'he plays' is reduced to 'play'

smoothing parameter that assumes values of either 0 or 1 and thus ignores synonyms that are not recognized by the lexicon. $P(f_k|c)$ is the sum of counts of the instance f_k in the lexicon for class c divided by the total number of words in class c.

Equation **2** is rewritten in log space to avoid underflow in calculations:

$$\hat{c} = argmax \log P(c) + \sum_{i \in l} \log P(w_i|c) \tag{3}$$

In order to find the probabilities $P(c)$ and $P(w_i|c)$, the maximum likelihood estimation is used: What percentage[8] of the words in our lexicon are of class c? Let N_c be the number of words in class c and N_{lex} be the number of words in the lexicon.

$$\hat{P}(c) = \frac{N_c}{N_{lex}} \tag{4}$$

The probability $P(w_i|c)$ is the percentage of times the word w_i appears amongst all words in the lexicon of category c.

$$\hat{P}(w_i|c) = \frac{count(w_i, c)}{\sum_{w \in lex} count(w, c)} \tag{5}$$

In the implementation it is made sure that unknown synonyms in the synonym set are removed before the algorithm calculates probabilities, in order to avoid calculations with zero.

2) Averaging approach to classification

The AM approach first expands the inital seed word list in two iterations[9] by extracting the synonym set for each seed word in the list and adding the synonyms and antonyms to the seed-list. New words are then classified by extracting the synonym set for the new word. For each synonym its majority occurrences in a class c is calculated. The program then calculates the arithmetic average of the polarity frequencies of all synonyms of a class c and assigns the new word the class of the higher averaged polarity.

$$\begin{aligned} &argmax P(c|w) = P(c)P(w|c) \\ &= argmax P(c) \frac{\sum_{i=1}^{n} count(syn_i, c)}{count(c)} \end{aligned} \tag{6}$$

The occurrences of $w's$ synonyms in the list of c is counted and the synonym is assigned the polarity with majority counts.

[8]note that there are duplicates of words in the expanded seed word lexicon and there can be occurrences of one word in both the positive and negative word lists

[9]Kim and Hovy [29] uses two iterations

```
AdjAdv Bucket: [{word='själv', negated=false}, {word='därmed', negated=
verb Bucket: [{word='skola', negated=false}, {word='tjäna', negated=fal
noun Bucket: [högerstyrd, politik, moderat]
Synonymer: [ensam, ohjälpt, utan hjälp, personifierad, ingen mindre än,
Word: själv, P(+) = 0,064055  P(-) = 0,935944
Synonymer: [till följd därav, med detta, därmedelst]
Synonymer: [läroanstalt, undervisningsanstalt, lärosäte, plugg, grunds
Word: skola, P(+) = 0,918124  P(-) = 0,081876
Synonymer: [förtjäna, ha i lön, ha betalt, inhösta, erhålla såsom vinst
Word: tjäna, P(+) = 0,940616  P(-) = 0,059384
Synonymer: [sak som köps och säljs, artikel, försäljningsobjekt, varus]
Word: vara, P(+) = 0,178133  P(-) = 0,821867
Topic: [högerstyrd, politik, moderat] Sentence Sentiment: 0.64893264

Själv skulle jag tjäna på högerstyrd politik, därmed inte sagt att
moderaterna är mitt favoritparti.
```

Fig. 1: Example of Sentence Classification

C. Classification of sentences

For sentence classification, first sentences are checked for idiomatic expressions with verbs. These idiomatic expressions are taken from the largest online list of Swedish idiomatic expressions on Wikipedia [42]. 273 idiomatic expressions are subsequently classified by three Swedes and added to the verb lists of the lexicon. If an idiomatic expression in the sentence is matching an expression in the lexicon, the whole expression is replaced with one word of the same sentiment as the expression.

Afterwards, the sentence is passed to Stagger, a part-of-speech tagger developed by the Stockholm University [43] in order to mark each word according to its grammatical meaning. In that step the word 'inte' in front of an adjective or adverb flags that word as negative and reverses word classification by calculating $(1 - wordSentiment)$. This is calculated once with 'inte' up to three words and once up to five words in front of adjectives and adverbs. No other negators are implemented in this stage.

In the next step sentiment bearing words, i.e. adjectives, adverbs and verbs, are filtered from the sentence and classified with NB and AM. The classifications for a sentence are aggregated and normalized in order to obtain a total sentence classification.

The gold standard for sentence classification of the 8 chosen sentences is set by 13 Swedes who have classified the sentences on a scale from 1 to 10 independently from each other. The scale is then normalized to an interval between 0 and 1, where classifications < 0.5 are negative and classifications $<= 0.5$ are positive. In addition the topic of a sentence is extracted primitively by filtering nouns of the sentence, see figure **1**.

D. Classification Model

Figure **D** in the appendix gives an overview of the classification Model

VI Evaluation of the artifact

The lexicon and word sentiment classification model design evaluations will be six-fold:

- Inter-annotator agreement (IAA) measure on seed word list
- Accuracy of seed word extraction in comparison to human classifications
- Accuracy of Naive Bayes (NB) and Average Method (AM) word extraction in comparison to Kim and Hovy [29]
- Comparison of word classifications by the General Inquirer lexicon to the created Swedish lexicon
- IAA of sentence classification
- Accuracy of NB and AM sentence classification

Furthermore parameters are tweaked in order to see which change of parameter might improve the sentiment lexicon.

1) Evaluation of the initial seed word list

The annotators had the tasks to classify the words as positive or negative. According to McHugh [44], "the kappa statistic is frequently used to test interrater reliability". The inter-annotator agreement is thus measured with the Cohen's kappa coefficient:

$$\begin{aligned} k &= \frac{p_0 - p_e}{1 - p_e} \\ &= 1 - \frac{1 - p_0}{1 - p_e} \end{aligned} \tag{7}$$

were p_0 is the relative observed agreement, p_e is the hypothetical probability of a chance agreement [44].

2) Evaluation of the expanded seed word list

The expanded seed word list is evaluated by comparing the the expanded list with a sample of human classifications of the expanded words which serves as gold standard.

Humans have classified a sample big enough to give a 95% confidence level with a confidence interval of 5%. Since the sample population will be quite large, crowd-sourcing is used to find humans who want to classify words. A simple website is made that displays the words to be classified together with a 'positive' and 'negative' button and instructional text. The classifications are saved in a text sheet and once all expanded words have been classified once, the process restarts in order to get a second sample for classification. With help of a program, human classifications will be compared to the classifications in the expanded seed word list and the percentage of discrepancies will be visualized. Precision, recall, accuracy and F1-score[10] are calculated.

3) Comparison of word classification to [29]

4) Evaluation of Swedish lexicon by comparison to the General Inquirer sentiment lexicon

The expanded lexicon and its classification of new words is compared to the classification of words by the elaborate 'General Inquirer' sentiment lexicon [45]. The 'General Inguirer' is a lexicon for content analysis of textual data that has been developed by the USA National Science Foundation and Research Grant Councils of Great Britain and Australia. It is based on the 'Harvard' and 'Lasswell' general-purpose dictionaries amongst other and is deemed trustworthy for research. [45]

5) Evaluation of NB and AM sentence classification

The evaluation of NB and AM classification algorithms for classifying sentences, will be done by comparing the classification result to a human gold standard which has been created as explained in section **V-C**. 13 people have classified 8 sentences from the following types of texts: newspaper, blog, novel, twitter, online forum and thesis paper.
The accuracy in classification is calculated and the percentage of discrepancies between human and machine classification of sentences is visualized.

6) Tweaking parameters

After a first evaluation of a sentence classifier that takes account of negation, the following tweaks are implemented sequentially to see how sentence classification improves. The choice of which parameters to tweak are chosen with regard to recommendations from previous research and heuristics:

- Smooth the expanded seed word list to equalize the number of negative and positive words
- Increase iterations for expanding the seed word list from two to three
- Change the negation window from 3 to 5 words

[10]for explanation of the terms, see appendix A

VII Result

In this part the author presents the results of the IAA of the initial seed word list, the IAA of sentence classification, the accuracy, precision, recall and F1-score of the expanded seed list, the accuracy of NB and AM sentence classification, the accuracy of this lexicon in comparison to an English lexicon and the results of tweaking parameters.

A. IAA of the initial seed word list

The appropriateness of an initial seed word list, i.e. unambiguous and frequently used words, is measured with inter-annotator classification agreement of those words. The Cohen's kappa coefficient is used to illustrate the inter-annotator agreement and reveals that it is 100%, since the three volunteers assigned, independently from each other, the exact same polarity to all 77 prepared seed words:

$$k = \frac{1-0.5}{1-0.5} = 1 \tag{8}$$

B. Accuracy, precision, recall and F1-score of the expanded seed word list

In total, from 15 positive and 19 negative adjectives and adverbs, 5105 positive and 4453 negative adjective and adverb synonyms and antonyms have been extracted. From 23 positive and 21 negative verbs, a total of 2410 positive and 2152 negative verbs have been extracted from www.synonymer.se. Humans have classified each word in a sample from the expanded seed word list at least two times[11].

The total percentage of agreement between the human gold standard and the expanded seed word list is calculated for the categories positive and negative, where a rating of $>= 0.5$ is considered positive and < 0.5 is considered negative. The total percentage of accuracy in classification of adjectives and adverbs is 79.43%, and for verbs it is 72%. Recall, precision, accuracy and F1-score[12] are displayed in table **I**.

The overview of the difference in classification strength for adjectives and adverbs is displayed in figure **2**, the corresponding figure for verbs can be found in appendix **E**. Words are points and the y-axis is a measure of the difference in classification.

[11] each adjective and adverb has been classified three times

[12] for explanation of the terms, see appendix A

Precision, Recall, Accuracy and F1-score of the classified verbs, adjectives and adverbs

	Verbs	AdjeAdv	Total
Total Nr. Words	353	389	742
Total Positive	248	228	476
Total Negative	105	161	266
True Positive	203	190	393
True Negative	105	161	266
False Positive	0	0	0
False Negative	45	38	83
Recall	1	1	1
Precision	0.846	0.857	0.85
Accuracy	0.87	0.90	0.89
F1 score (balanced)	0.92	0.92	0.92

TABLE I: Recall, precision, accuracy and F1-score of the classified words

0 means that humans and the seed word list have the same classification strength, -1 means that the seed word list classifies a word as positive, while humans classify the same word as negative, and +1 means the opposite.

An example of word classifications between humans, NB and AM is displayed in figure **3**

C. Comparison to Kim and Hovy [29]

Figure **4** shows that all NB classifications are the same as the word examples shown in the article 'Determining the sentiment of opinions' in Kim and Hovy [29], and all but one classification for the word 'abysmal' (translates to 'bottenlös') are the same as in Kim and Hovy [29] for AM method.

D. Accuracy of classifications from the General Inquirer compared to the created Swedish lexicon

For the comparison of the General Inquirer English sentiment lexicon [45], 285 random sampled verbs, adverbs and adjectives have been translated with Google Translate into Swedish. The Swedish terms were then classified with the NB classifier. The comparison of random samples of verbs, adjectives and adverbs from the General Inquirer online sentiment lexicon [45] with the created Swedish lexicon shows that 90% of the words have the same classification in the English and the Swedish lexicon with a 95%

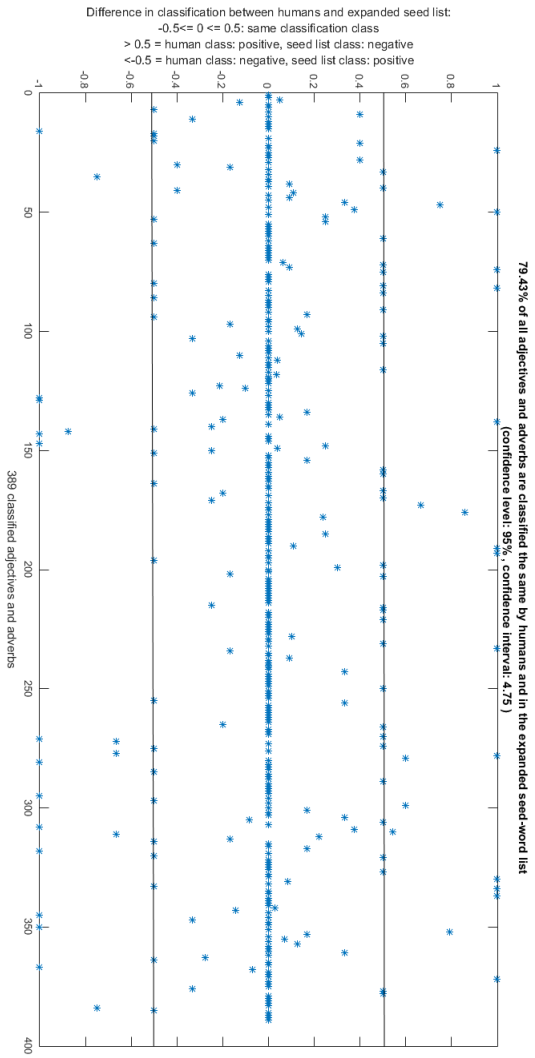


Fig. 2: Difference in classification strength between humans and the expanded seed list for adjectives and adverbs

confidence level and confidence interval of 3.44%.

E. IAA of sentence classification

The inter-annotator agreement is calculated as percentage of agreement on the categories 'negative', 'positive': The agreement total on all 8 sentences is 83%, see table **II**.

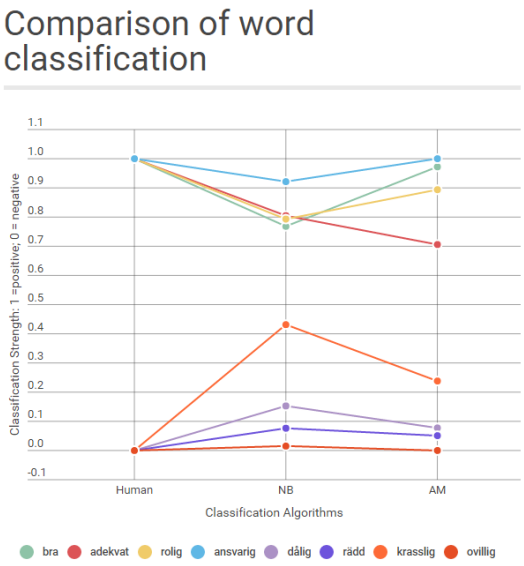


Fig. 3: Comparison of word classification between humans and the two classification algorithms

Swedish sentiment model	Kim and Hovy (2004)
Word: bottenlös, P(+) = 0,462648 P(-) = 0,537352 Classification NB_Adj = 0.46264768 Classification = NEGATIVE Word: bottenlös, P(+) = 0,764706 P(-) = 0,235294 Classification AM Adj = 0.7647059 Classification = POSITIVE	abysmal : NEGATIVE [+ : 0.3811][- : 0.6188]
Word: adekvat, P(+) = 0,804488 P(-) = 0,195512 Classification NB_Adj = 0.80448765 Classification = POSITIVE Word: adekvat, P(+) = 0,705882 P(-) = 0,294118 Classification AM Adj = 0.7058824 Classification = POSITIVE	adequate : POSITIVE [+ : 0.9999][- : 0.0484e-11]
Word: rädd, P(+) = 0,076114 P(-) = 0,923886 Classification NB_Adj = 0.076113544 Classification = NEGATIVE Word: rädd, P(+) = 0,050000 P(-) = 0,950000 Classification AM Adj = 0.05 Classification = NEGATIVE	afraid : NEGATIVE [+ : 0.0212e-04][- : 0.9999]
Word: krasslig, P(+) = 0,439214 P(-) = 0,560786 Classification NB_Adj = 0.43921426 Classification = NEGATIVE Word: krasslig, P(+) = 0,161290 P(-) = 0,838710 Classification AM Adj = 0.16129033 Classification = NEGATIVE	ailing : NEGATIVE [+ : 0.0467e-8][- : 0.9999]
Word: amusant, P(+) = 0,908202 P(-) = 0,091798 Classification NB Adj = 0.90820175 Classification = POSITIVE Word: amusant, P(+) = 1,000000 P(-) = 0,000000 Classification AM Adj = 1.0 Classification = POSITIVE	amusing : POSITIVE [+ : 0.9999][- : 0.0593e-07]
Word: ansvarig, P(+) = 0,923082 P(-) = 0,076918 Classification NB_Adj = 0.9230822 Classification = POSITIVE Word: ansvarig, P(+) = 1,000000 P(-) = 0,000000 Classification AM Adj = 1.0 Classification = POSITIVE	answerable : POSITIVE [+ : 0.8655][- : 0.1344]
Word: gripbar, P(+) = 0,644342 P(-) = 0,355658 Classification NB_Adj = 0.64434206 Classification = POSITIVE Word: gripbar, P(+) = 0,800000 P(-) = 0,200000 Classification AM Adj = 0.79999995 Classification = POSITIVE	apprehensible: POSITIVE [+ : 0.9999][- : 0.0227e-07]
Word: ovillig, P(+) = 0,015754 P(-) = 0,984246 Classification NB Adj = 0.015754454 Classification = NEGATIVE Word: ovillig, P(+) = 0,000000 P(-) = 1,000000 Classification AM Adj = 0.0 Classification = NEGATIVE	averse : NEGATIVE [+ : 0.0454e-05][- : 0.99991]
Word: skylla, P(+) = 0,088784 P(-) = 0,911216 Classification NB Verb = 0.08878382 Classification = NEGATIVE Word: skylla, P(+) = 0,000000 P(-) = 1,000000 Classification AM Verb = 0.0 Classification = NEGATIVE	blame : NEGATIVE [+ : 0.2530][- : 0.7469]

Fig. 4: Comparison of word classification between Kim and Hovy [29] and the Swedish sentiment model

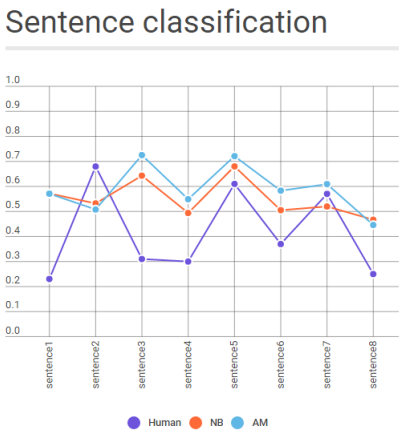


Fig. 5: Comparison of sentence classification between humans and program

IAA for sentence classification

	IAA (percentage of agreement)	Classification
Sentence 1	1	Negative
Sentence 2	0.85	Positive
Sentence 3	0.70	Negative
Sentence 4	1	Negative
Sentence 5	0.70	Positive
Sentence 6	0.70	Negative
Sentence 7	0.85	Positive
Sentence 8	0.85	Negative
Total	0.83	

TABLE II: Inter annotator agreement for sentences

F. Accuracy of NB and AM sentence classification

8 sentences have been chosen from different types of texts and classified by 13 humans. The types of texts are news, blog, twitter, novel, online forum and thesis paper, see appendix **F**. The comparison between human and program sentence classification can be seen in figure **5**. The inter-annotator agreement between humans for sentence classification is k=0.98. The accuracy for sentence classification into the categories positive and negative for the NB and the AM algorithm is poor at only 50% correctly classified sentences.

G. Tweaking parameters

Accuracy, precision, recall and F1-score of the expanded seed word list after smoothing

Smoothing is the process to equalize the number of negative and positive words in the expanded seed word list. After smoothing, the NB showed a 16% total improvement in accuracy, while the AM improved by 15% in total accuracy, see appendix **IV**. The total percentage of rightly classified sentences for the NB and AM classification increased to 62.5%, i.e. five out of 8 sentences.

Three iterations for expansion of seed word list
The increase of iterations from two to three (without smoothing) seems to have increased the total accuracy of sentence classification for the NB by 10% whereas the AM decreased in total accuracy to 38%, see appendix **IV**. However, the 10% improvement of accuracy for the NB affects all sentences and resulted in 100% of all sentences being rightly classified, whereas in the AM only 3 out of 8 sentences are rightly classified, which is roughly a chance outcome.

Change of negation window from 3 to 5 words
This tweak affects only sentence 3, since it is the only sentence with an adjective appearing up to five words after a negation. Changing the negation window from 3 to 5 words after negation improves the NB classification for sentence 3 by 15% while it did not affect the AM algorithm at all, see appendix **V**. For the NB algorithm a total of 0.75% of sentences, i.e. 6 out of 8 have been rightly classified with this tweak.

The rightly and wrongly classified sentences with their sentiment strengths after the tweak of parameters are shown in appendix **III**.

VIII Discussion

In this chapter, reasons for results, difficulties in sentence classification, biases and future improvements are discussed.

A. Reasons for results

1) Inter-annotator agreement on the initial seed word list

The initial seed word list of in total 77 adjectives, adverbs and verbs seems to contain seed words that can be unambiguously classified into negative and positive. Furthermore all words seem to be known, since the inter-annotator agreement shows

that there was no disagreement between the three independently working human annotators.

2) Accuracy, precision, recall and F1-score of the expanded seed word list

The classifications of the expanded seed word list is accurate in over 70% with human classifications which is considered reliable.

3) Accuracy of NB and AM word classification

It has been calculated that a recall of 1, a precision of over 80% and an accuracy of over 85% has been achieved for adjectives, adverbs and verbs. This fulfills the solution objectives **IV-B** of a recall of over 90% for adjective and over 80% for verb word classifications and exceeds the recall of Kim and Hovy [29]. The accuracy of both algorithms for word classification is therefore comparable with results from previous research.

4) Comparison with Kim and Hovy [29]

The word classifications of the Swedish model with NB give in 100% the same result as the words shown in Kim and Hovy [29] and in 89% it gives the same result as in Kim and Hovy [29] with the AM algorithm. This comparison strengthens the reliability of the results, since both lexicons have been created with a similar bootstrapping method.

5) Comparison to the General Inquirer sentiment lexicon

90% of all words have the same classification in the created Swedish lexicon as in the General Inquirer. This result of the comparison strengthens the reliability of the Swedish lexicon and shows that the Swedish lexicon can be used for sentiment analysis since English and Swedish words often have the same sentiment.

The 10% deviation in the comparison could be language specific. Three reasons for different sentiment classifications of English and Swedish words have been observed: 1) Homonymy 2) Translation and 3) Cultural differences

For example the word 'upphetsad' (exited in English) has been classified as positive in the General Inquirer but as negative in the Swedish lexicon. The reason for the negative Swedish classification is that 'upphetsad' not only means 'exited' in Swedish but can also bear the meaning 'ilsken' and 'arg' (angry). That one word can have different meanings is called homonymy and differs from language to language.

The translation is also a decisive factor in conveying the right sentiment of words. The word 'seize' in English has negative sentiment, but the Swedish translation 'ta vara på' ('profit from') is positive in Swedish. The difference in classification here is because the English word 'seize' can mean 'abduct' or 'occupy' which is negative, but also 'apprehend' and 'grasp' as in 'seize the day', which is positive. The first translation that Google Translate gave, determined the sentiment for the Swedish word.

Finally cultural differences can be a reason for differing sentiments for words in different languages. For example the word 'allvarlig' has been classified as negative in the Swedish lexicon and by Swedish annotators, while the word 'serious' is positive in the General Inquirer. Both classifications might be right and cultural differences in how these words are used could be the reason why sentiment differs. The theory is that Swedes might not use the word 'allvarlig' as often as 'serious' is used in English which is why the word bears more weight in Swedish.

6) Accuracy of NB and AM sentence classification

Sentence classification resulted in only 50% correctly classified sentences. The initial sentence classification therefore does not fulfil the set objectives of a sentence level sentiment analysis accuracy between 66% and 80% delivered by general English sentiment lexicons.

Sentences 1,3,4 and 6 were misclassified, and incidentally they correspond to four of 5 sentences that have been classified as negative by humans. This means that 80% of all negative sentences have been misclassified and although this is not in any way a representative sample for all possible sentences, this fact in combination with the fact that there were more positive than negative words in the seed list before smoothing leads to believe that sentence classification was biased towards a more positive classification. After smoothing sentence classification improved for both NB and AM to 62.5%. This confirms the believed bias towards positive classification.

Increasing iterations from two to three included in average about 50 000 more terms than before. This improved the NB sentence classification to 100% accuracy, i.e. all sentences have been correctly classified. However, this tweak decreased the accuracy of the AM classification to only 38%. Since three iterations will include a high amount of wrongly classified words, this decreases the average accuracy of the AM classification algorithm. In contrary, NB is based on frequency statistics, which get better by including more words through three iterations even though this means to also include wrongly classified words.

The accuracy improved for the NB by taking into account up to five words after the negation 'inte', while the AM accuracy did not change. Of course the result is just valid for the one classified sentence

in which this criteria applied and it needs to be tested if a window of five is in general better than a window of three.
The AM algorithm is not recommended for more complex sentiment analysis on sentence and document level. Complex sentences can not be classified by simply averaging the sentiments of words in the given sentence. Sentence sentiment classification therefore needs more refined methods to be accurate.

Genres of sentences seem to have no effect on the accuracy of classifications since sentences for different genres have been misclassified. No generalization can be made from only 8 classified sentences.

B. Difficulties during classification

It has been observed that composed words, such as 'favoritparti' are not recognized by the part-of-speech tagger and was therefore not tagged as a noun defining the topic of the sentence.

Furthermore, not all idiomatic expressions with verbs are part of the lexicon and will have to be added one by one.

Sentiments delivered by cultural contexts can not be recognized by the classification method explored in this study. For instance most humans will define the sentence 'Military drones were used in an airstrike' as negative, while the lack of sentiment-bearing adjectives, adverbs and verbs will define the sentiment of the sentence as neutral.

C. Bias

The process for the initial seed word choice (also choices made by previous research) is rather subjective and a different choice of initial seed words might result in better classifications.

The results of the sentence classifications will have to be tested on more than 8 sentences from different types of texts in order to be generalized. Only a general baseline can be gained for the accuracy of a first general purpose lexicon-based sentiment classification model.

Bias might have arisen in the crowd-sourcing approach for word classification since no control over the Swedish competences of the crowd could be exerted.

D. Future improvements

The sentence model as it is could be refined by taking into account sentence connectors such as 'and' and 'or' and adding more negator words. It could also be improved by adding the 'neutral' label for classification, which is not a problem, since the model prints out sentiment strengths for words and sentences. Furthermore diminisher words that weaken expressions and intensifier words which strengthen expressions can be included into the model. Example of diminishers are 'a bit', 'a little' and examples of intensifiers are 'rather' and 'very'. Future models for Swedish sentiment analysis could include entire corpus for sentiment calculations. More sophisticated classification models, such as support vector machines and neural networks could be tested for Swedish.

IX Conclusion

The aim of this research study was to build, evaluate and test a model for determining word and sentence sentiments with a general purpose Swedish sentiment lexicon at the basis, and have been fulfilled. A general purpose lexicon was built by seed word expansion, the lexicon was combined with a classifier for Swedish sentence sentiment analysis and changes of parameters to improve sentiment analysis have been evaluated and discussed.

Therefore the main research question and subquestions about the accuracy, precision, recall, F1-score and tweaks of parameters of the lexicon-based sentiment classification model for different types of texts have been answered: The best accuracy for sentence classification for NB was established after three iterations where all sentences were rightly classified. A window of five words for negation resulted in 75% rightly classified sentences, which is also within the norm of previous results in the field. The AM algorithm performed best after smoothing with a maximum accuracy of 63.5%, i.e. five out of eight sentences. The AM algorithm is the simplest algorithm and performs accordingly unimpressive. However, the results can not be generalized, since only eight sentences of different domains have been tested.

This research has contributed with a first open-source Swedish sentiment model for word and sentence classification that is comparable to English sentiment models in performance according

to the tests that have been done during this research. The model takes account of Swedish idiomatic expressions and negation, but could be refined to take sentence connectors into consideration amongst other. The sentence classification model also needs to be tested on more sentences to ensure its reliability. The entire code can be found on 'https://github.com/michelleludovici'.

Building a first open-source model for Swedish sentiment analysis could be a stepping stone towards improved sentiment analysis in Swedish. It might therefore incite further research in this field in order to develop more sophisticated approaches for sentiment analysis on sentence-, paragraph- and document-level and applications for that purpose. The process would give insights into the efficiency of the lexicon compared to using an English lexicon as well as into improvements needed for next generation lexicons.

References

[1] J. Karlgren, M. Sahlgren, F. Olsson, and F. Espinoza, *Usefulness of Sentiment Analysis*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012.

[2] D. R. Rice and C. Zorn, "Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies ∗," *Proceedings of NDATAD*, 2013.

[3] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2.1, pp. 1–8, 2011. [Online]. Available: http://arxiv.org/abs/1010.3003http://dx.doi.org/10.1016/j.jocs.2010.12.007

[4] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, pp. 192–199, 2011.

[5] G. Vinodhini and R. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 6, pp. 282–292, 2012. [Online]. Available: http://www.dmi.unict.it/{~}faro/tesi/sentiment{_}analysis/SA2.pdf

[6] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.asej.2014.04.011

[7] B. Pang and L. Lee, "A Sentimental Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Framework*, vol. cs.CL, pp. 271–278, 2002. [Online]. Available: http://arxiv.org/abs/cs/0409058v1

[8] E. Psomakelis, K. Tserpes, D. Anagnostopoulos, and T. Varvarigou, "Comparing methods for Twitter Sentiment Analysis," *6th Conference on Knowledge Discovery and Information Retrieval 2014*, 2014.

[9] C. N. D. Santos and M. Gatti, "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts," *Proceedings of the 25th International Conference on Computational Linguistics*, pp. 69–78, 2014.

[10] M. Fuchs, "Sentiment Analysis: Extracting Decision-Relevant Knowledge from UGC," *Information and Communication Technologies in Tourism 2014*, no. JANUARY, 2013. [Online]. Available: http://link.springer.com/10.1007/978-3-319-03973-2

[11] P. Turney, "Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews," in *Proceedings of the ACL*, 2002, pp. 417–424.

[12] M. Turchi and A. Balahur, "Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data," in *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP*, no. September, Hissar, Bulgaria, 2013, pp. 49–55.

[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[14] F. Wei, X. Ge, L. Zhang, X. Liu, and M. Zhou, "Lost in Translations ? Building Sentiment Lexicons Using Context Based Machine Translation," *Coling 2012*, no. December 2012, pp. 829–838, 2012.

[15] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "SentiFul: Generating a reliable lexicon for sentiment analysis," *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–6, 2009. [Online]. Available: http://ieeexplore.ieee.org/xpl/freeabs{_}all.jsp?arnumber=5349575

[16] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Infomration Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[Online]. Available: http://www.nowpublishers.com/product.aspx?product=INR{&}doi=1500000001

[17] M. D. Cao and I. Zukerman, "Experimental Evaluation of a Lexicon- and Corpus-based Ensemble for Multi-way Sentiment Analysis," *Proceedings of the Australasian Language Technology Association Workshop 2012*, pp. 52–60, 2012. [Online]. Available: http://www.aclweb.org/anthology/U/U12/

[18] J. Read and J. Carroll, "Weakly supervised techniques for domain-independent sentiment classification," *Proceeding of the 1st international CIKM workshop on Topicsentiment analysis for mass opinion TSA 09*, p. 45, 2009. [Online]. Available: http://dx.doi.org/10.1145/1651461.1651470

[19] D. Lambov, S. Pais, and G. Dias, "Merged agreement algorithms for domain independent sentiment analysis," *Procedia - Social and Behavioral Sciences*, vol. 27, no. Pacling, pp. 248–257, 2011. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1877042811024323

[20] J. Rothfels and J. Tibshirani, "Unsupervised sentiment classification of English movie reviews using automatic selection of positive and negative sentiment items," Stanford University, Tech. Rep., 2010.

[21] V. Hatzivassiloglou, V. Hatzivassiloglou, K. McKeown, and K. McKeown, "Predicting the semantic orientation of adjectives," *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, vol. pages, p. 181, 1997. [Online]. Available: http://portal.acm.org/citation.cfm?id=976909.979640

[22] S. M. Vohra and J. B. Teraiya, "A comparative study of sentiment analysis techniques," *Journal of Information, Knowledge and research in Computer Engineering*, vol. 2, no. 2, pp. 313–317, 2013.

[23] A. Balahur and J. M. Perea-Ortega, "Sentiment analysis system adaptation for multilingual processing: The case of tweets," *Information Processing & Management*, vol. 51, no. 4, pp. 547–556, 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0306457314000934

[24] H. Kang, S. J. Yoo, and D. Han, "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews," *Expert Systems with Applications*, vol. 39, no. 5, pp. 6000–6010, 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0957417411016538

[25] P. Chikersal, S. Poria, and E. Cambria, "SeNTU : Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning," *SemEval2015*, no. SemEval, pp. 647–651, 2015.

[26] J. A. Øye, "Sentiment Analysis of Norwegian Twitter Messages," Ph.D. dissertation, Norwegian University of Science and Technology, 2015. [Online]. Available: http://brage.bibsys.no/xmlui/bitstream/handle/11250/2352299/12125{_}FULLTEXT.pdf?sequence=1{\protect\T1\textbraceleft}{&}{\protect\T1\textbraceright}isAllowed=y

[27] N. Vincze and Y. Bestgen, "Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée," *18ème Conférence sur le ...*, vol. 2, no. 1, 2011. [Online]. Available: http://dial.academielouvain.be/handle/boreal:74775

[28] J. Steinberger, M. Ebrahim, M. Ehrmann, A. Hurriyetoglu, M. Kabadjov, P. Lenkova, R. Steinberger, H. Tanev, S. Vázquez, and V. Zavarella, "Creating sentiment dictionaries via triangulation," *Decision Support Systems*, vol. 53, no. 4, pp. 689–694, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2012.05.029

[29] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*, pp. 1367–es, 2004. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1220355.1220555

[30] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2008.

[31] T. Hedlund, A. Pirkola, and K. Järvelin, "Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval," *Information Processing and Management*, vol. 37, pp. 147–161, 2001.

[32] JetBrains, "IntelliJ IDEA," 2015. [Online]. Available: https://www.jetbrains.com/idea/

[33] M. Denscombe, *The good research guide for small-scale social research projects*, 2nd ed. Maidenhead: Philadelphia: Open University Press, 2003.

[34] Z. Kozareva, E. Riloff, and E. Hovy, "Semantic class learning from the web with hyponym pattern linkage graphs," in *roceedings of ACL-08: HLT*, 2008, pp. 1048–1056.

[35] S. Igo and E. Riloff, "Corpus-based semantic lexicon induction with web-based corroboration," in *Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics.*, 2009.

[36] Z. Kozareva and E. Hovy, "Not All Seeds Are Equal : Measuring the Quality of Text Mining Seeds," *Proceeding HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, no. June, pp. 618–626, 2010. [Online]. Available: http://portal.acm.org/citation.cfm?id=1858086

[37] P. D. Turney and M. L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," *ACM Transactions on Information Systems*, vol. 21, no. 4, pp. 315–346, 2003. [Online]. Available: papers3://publication/uuid/0F0B6624-89AC-4619-9BF5-0CBE95FB6C64

[38] M. Hu and B. Liu, "Mining and summarizing customer reviews," *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining KDD 04*, vol. 04, p. 168, 2004. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1014052.1014073

[39] S. Okasha, *Philosophy of Science: A Very Short Introduction*, O. U. Press, Ed., 2002.

[40] Sinovum Media, "Om Synonymer.se," 2015. [Online]. Available: http://www.synonymer.se/om{_}synonymer.php

[41] H. Johan, J. Nilsson, and J. Nivre, "Swedish MaltParser model," 2012. [Online]. Available: http://www.maltparser.org/mco/swedish{_}parser/swemalt.html

[42] "Lista över svenska idiomatiska uttryck," 2015. [Online]. Available: https://sv.wikipedia.org/wiki/Lista{_}{%}C3{%}B6ver{_}svenska{_}idiomatiska{_}uttryck

[43] R. Östling, "Stagger – The Stockholm Tagger." [Online]. Available: http://www.ling.su.se/english/nlp/tools/stagger

[44] M. McHugh, "Interrater reliability: the kappa statistic." *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.

[45] W. J. Hall, "Welcome t the General Inquirer Home Page." 2002. [Online]. Available: http://www.wjh.harvard.edu/{~}inquirer/

Appendix A
Glossary of Terms and Abbreviations

IAA	Inter annotater agreement; the measure of how much human sentiment annotaters agree
NB	Naive Bayes
AM	Averaging Method, see section V-B2
Precision	Measures the exactness of the classifier. Percentage of correctly classified items
Recall	Measures the completeness of a classifier. Percentage of correct items of one class
F1-score	F1-score is also called F1-measure and is the weighted harmonic mean of precision and recall.

Appendix B
DSRM Process Model by [30]

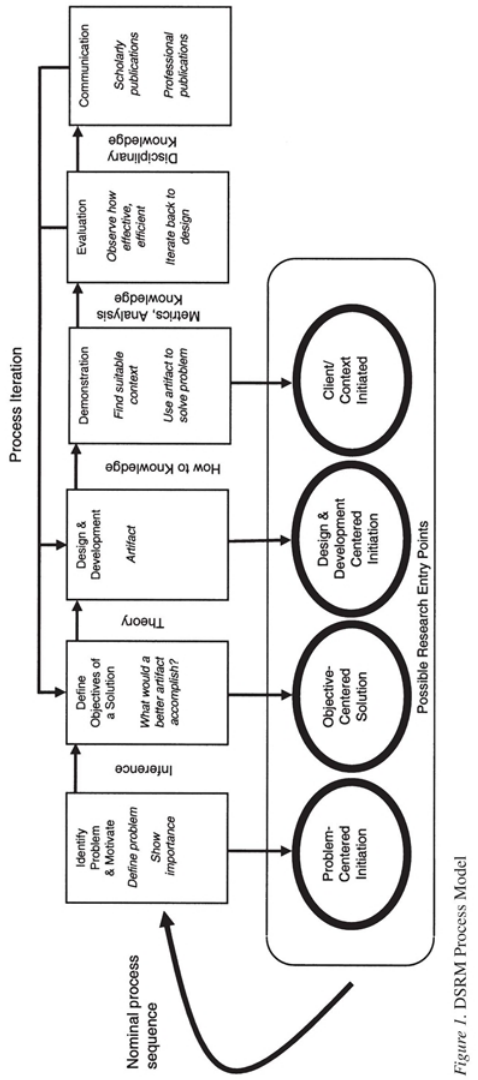


Figure 1. DSRM Process Model

Fig. 6: DSRM Process Model by [30]

Appendix C
Demonstration of word classification

SynonymAPI.java × ShufflingWords.java × NaiveBayes.java × MaltParser.java ×

```
package se.milu.maltparser;

import org.junit.Test;

import java.io.*;
import java.util.HashSet;
import java.util.Scanner;
import java.util.Set;

/**
 * Created by Michelle on 2015-12-01.
 */
public class NaiveBayesTest {

    public static final String WORD = "bra";

    @Test
    public void test_naive_bayes() throws IOException {

        File pos = new File ("posAdjAdv.txt");
        File neg= new File ("negAdjAdv.txt");

        NaiveBayes nb = new NaiveBayes(pos,neg);
        float classification = nb.naiveBayesClassify(WORD);

        System.out.println("Classification NB = " + classification);

    }
```

ne: 4 of 4 (in 0,482 s)

```
"C:\Program ...
{}
Synonymer: [god, fin, finfin, prima, fräsch, cool, schysst, just, läcker, vass, utmärkt, f
Word: bra, P(+) = 0,973244  P(-) = 0,026756
Classification AM = 0.9732442
{}
Synonymer: [god, fin, finfin, prima, fräsch, cool, schysst, just, läcker, vass, utmärkt, f
Word: bra, P(+) = 0,959598  P(-) = 0,040402
Classification NB = 0.9595982

Process finished with exit code 0
```

Fig. 7: Demonstration of word classification algorithms

Appendix D
Classification Model

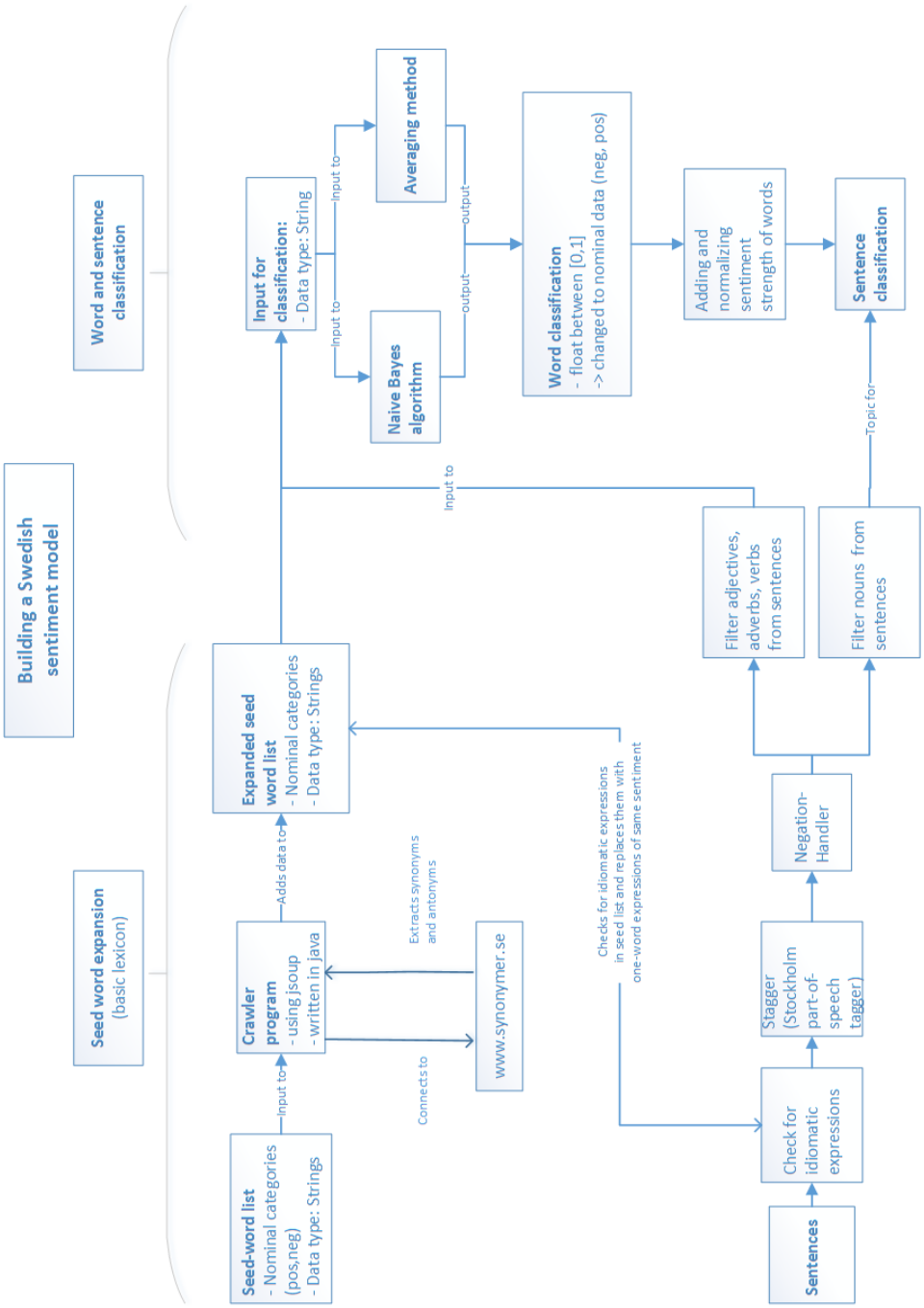
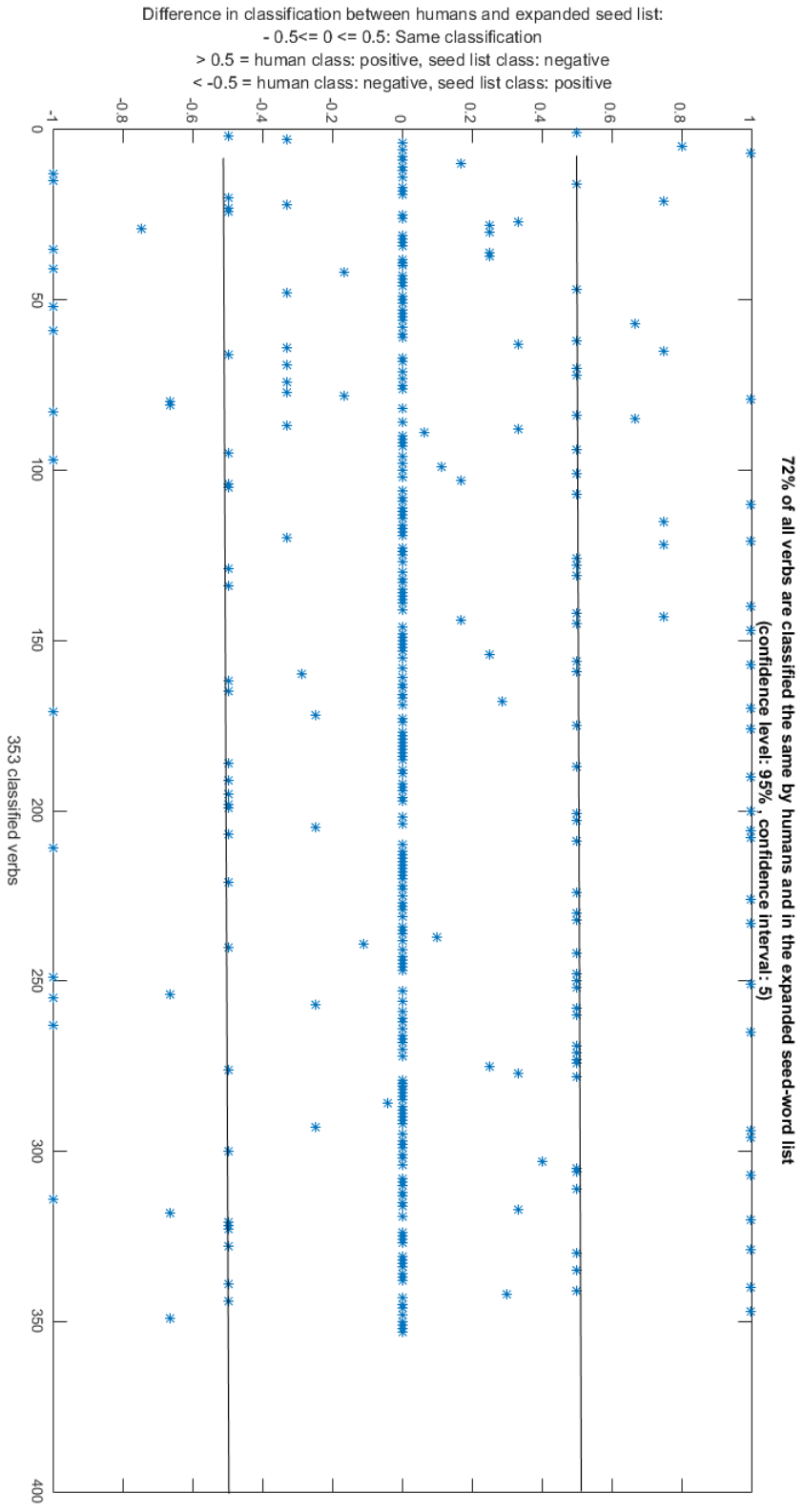


Fig. 8: Classification Model

Appendix E
Percentage of classification agreement for expanded seed word list



Appendix F
Sentence for sentiment classification

The sentences for sentiment classification and the corpus are the following:

- Miljöorganisationen WWF tycker inte att länderna gör något bra jobb för att skydda miljön. (8sidor.se)
- Konferensen fokuserade kring att tolerans och förståelse behövs för att förebygga konflikter och krig. (Blog)
- Jag skäms för dem eftersom jag ju vet hur fruktansvärt mycket lidande jag skulle åsamka mina nära och kära. (Blog)
- Enligt Naturvårdsverkets klassificering är föroreningen "mycket omfattande" och ämnenas farlighet "mycket hög". (Göteborgs Posten)
- Vi råder dig att kontakta vår ombordspersonal så kan de hjälpa dig att kika på detta. (Twitter)
- Men han har efter slutad klostervisitation varit ovanligt upprörd, nästan vild. (August Strindbergs samlade verk)
- Själv skulle jag tjäna på högerstyrd politik , därmed inte sagt att moderaterna är mitt favoritparti. (Familjeliv: Allmänna rubriker – Ekonomi & juridik)
- Sa även vad jag tyckte om det och det blev inskrivet i journalen senare , något i stil med det mest menigslösa 13 minutrarna i mitt liv. (Flashback forum: Vetenskap & humaniora)

Appendix G
Rightly and wrongly classified sentences

First Try	sentence1	sentence2	sentence3	sentence4	sentence5	sentence6	sentence7	sentence8
	< 0,5 is negative ; > = 0,5 is positive							
Human	0,23	0,68	0,31	0,3	0,61	0,37	0,57	0,25
NB	0,572	0,532	0,643	0,494	0,68	0,505	0,52	0,467
AM	0,571	0,508	0,725	0,549	0,721	0,583	0,609	0,446
AfterSmoothing	sentence1	sentence2	sentence3	sentence4	sentence5	sentence6	sentence7	sentence8
	< 0,5 is negative ; > = 0,5 is positive							
Human	0,23	0,68	0,31	0,3	0,61	0,37	0,57	0,25
NB	0,572	0,533	0,644	0,483	0,68	0,502	0,524	0,454
AM	0,565	0,508	0,722	0,489	0,715	0,55	0,583	0,433
3Iterations	sentence1	sentence2	sentence3	sentence4	sentence5	sentence6	sentence7	sentence8
	< 0,5 is negative ; > = 0,5 is positive							
Human	0,23	0,68	0,31	0,3	0,61	0,37	0,57	0,25
NB	0,492	0,537	0,474	0,425	0,604	0,471	0,502	0,409
AM	0,566	0,479	0,681	0,56	0,754	0,539	0,629	0,416

TABLE III: Overview of rightly and wrongly classified sentences

Appendix H
Sentence accuracy measurement after smoothing and after three iterations

	Improvement (+) and decrease (-) in accuracy by smoothing								Total
	Sentence1	Sentence2	Sentence3	Sentence4	Sentence5	Sentence6	Sentence7	Sentence8	
NB	0	1,238599982	-0,00155521	0,022267206	0	0,005940594	0,007692308	0,027837259	**0,162598**
AM	0,973091704	0	0,004137931	0,109289617	0,008321775	0,056603774	0,042692939	0,029147982	**0,152911**

	Improvement (+) and decrease (-) in accuracy of 3 iterations relative to two iterations								Total
	Sentence1	Sentence2	Sentence3	Sentence4	Sentence5	Sentence6	Sentence7	Sentence8	
NB	0,13986014	0,009398496	0,262830482	0,139676113	0,111764706	0,067326733	-0,034615385	0,124197002	**0,102555**
AM	0,008833922	-0,060542797	0,064610866	-0,019642857	-0,043766578	0,081632653	0,031796502	0,072115385	**0,01688**

TABLE IV: Improvement of NB and AM sentence classification after smoothing and after three iterations

Appendix I
Sentence accuracy after increasing negation window

Neg.Window:5	Sentence 3
NB	0,153965785
AM	0

TABLE V: Improvement of sentence classification after increasing the negation window to 5