Swedish Sentiment Analysis with SVM and handlers for language specific traits

Michelle Ludovici

Department of Computer and Systems Sciences

Degree project 30 HE credits
Degree subject: Computer and Systems Sciences
Degree project at the Master level
Spring 2016
Advisor: Hercules Dalianis
Assistant advisor: Rebecka Weegar
Reviewer: Paul Johannesson

Stockholm University

Swedish Sentiment Analysis with SVM and handlers for language specific traits

Michelle Ludovici

Abstract

Sentiment analysis is a field in the area of text mining, concerned with the classification of subjective expressions from written text. For the English language sentiment lexicons and models exist, but in Swedish, there is a lack of labelled training data, and of open-source sentiment classification models and tools for Swedish. This thesis addresses the lack of knowledge on how to feasibly implement a Swedish sentiment model. Additionally it addresses the lack of knowledge if Swedish language specific handlers could improve model accuracy. All handler implementations aggregate information which semantically belongs together in Swedish, but is separated within a sentence. The research goal is to design, instantiate and evaluate a Swedish sentiment classification model using a Support Vector Machine algorithm (SVM). The main research question is the following: "What is the performance of the designed Swedish sentiment classification model, measured with accuracy, precision and recall?" Domain specific semi-supervised sentiment classification of article newspaper texts using support vector machines is applied. The Design Science Research Methodology (DSRM) and a Framework for Evaluation in Design Science Research (FEDS) are employed in the thesis. The labelled training data for Swedish sentences was created with the help of two algorithms which gave sentiment labels to Swedish sentences. Sentences with matching labels by the two algorithms were selected for the training data set. The SVM was trained on the training data and SVM parameters were tuned to get best accuracy. Language specific handlers were implemented to handle Swedish idiomatic expressions, phrasal verbs and negations. The result of the model evaluation shows that the SVM achieves a best accuracy of 83.71%, which is comparable to similar previous research. This accuracy was achieved with the radial basis function kernel, without scaling, and binary weighting. The inclusion of language specific handlers did not improve the classifier. Each logically combined handler feature-value pair is rare and gets negligible weight by the SVM; in addition, those combined features-value pairs may enhance the total feature set, which might have detrimental effects on accuracy. Future studies could explicitly concentrate on researching the effects of language specific feature reduction and enhancement on SVM prediction results.

Keywords

Sentiment analysis, SVM, semi-supervised, classification, language specific

Contents

Acknowledgements

I would first like to give special thanks to my thesis advisor PhD Rebecka Weegar and Professor Hercules Dalianis of the department of Computer and Systems Sciences, DSV, at Stockholm University. I appreciate the great advice and fast responses which made studying over distance effortless. Both guided me by giving me feedback, while at the same time letting me create my own work.

In addition I would like to thank my three volunteers who patiently labelled so many sentences and thereby helped me greatly in validating the results of my program!

Furthermore I owe many thanks to my mentor, Magnus Engström for his feedback and general wisdom and finally I would like to thank the great team of MittMedia, DMU, for helping me with getting the data - thank you Frasse! - and for their genuine interest in my thesis work.

1. Introduction

Sentiment analysis is a field in the area of text mining, concerned with the extraction and classification of subjective expressions from written text. The usefulness of sentiment analysis stems from the premise that lexical items carry affective and attitudinal information (Karlgren et al., 2012, p.426). An example of sentiment classification is the classification of texts in 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, 2013). In research there are many names for the 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. For example, Tweets have been employed to foresee stock market trends (Bollen et al., 2011). Sentiment analysis can also be used for political party opinion monitoring in order to calculate the preference of one candidate or party over others (Conover et al., 2011).

In this research, the possibility for sentiment analysis on newspaper articles is explored. The practical application of sentiment analysis of newspaper articles would be to display article sentiment in order to visualize long-lasting problems as well as positive trends on important societal topics, such as the education system. The topics could be compared over time, by cities and keywords. The results of the sentiment analysis might be used by governmental agencies for decision making based on historical facts and as trend indicator. At the same time, for the public media domain, sentiment analysis highlights one of the most important functions of newspapers: to enable the democratic process by giving a voice to society, describing existing problems, working solutions and surveying power.

1.1 Previous Research

The article "Sentiment Analysis and Opinion Mining: A Survey" by Vinodhini and Chandrasekaran (2012) shows that comprehensive research has been done in the field of sentiment analysis. There exists many sentiment classifier programs which are trained on words, sentences or entire texts. Such classifications may be done in the categories 'positive', 'negative' and 'neutral'. The term 'sentiment model' refers in this thesis to the design and creation of a sentiment analysis concept, as well as the physical program to execute sentiment analysis.

1.1.1 Domain specific or general

Supervised and unsupervised machine learning techniques were applied to sentiment analysis with both domain specific and general data[1]. It has been shown that classification programs trained on unspecific, i.e. general domains, usually perform worse than programs trained on specified domains, such as product reviews (Andreevskaia and Bergler,

[1] Domain specific data can for example be movie reviews, where the domain would be movies. General data is found in articles, which can be about anything and can not be categorised

2008; Aue and Gamon, 2005). Also classifiers trained on one domain perform worse when used on another domain. This has been called the domain-transfer problem and seems intuitively understandable since people express sentiments different depending on context (Pan et al., 2010). However, classifiers trained on a general domain have the advantage of being able to be adapted to a certain domain in hindsight, whereas classifiers trained on a specific domain can not be generalized as easily. Examples of sentiment classification on general domains are the work of (Turney, 2002; Read and Carroll, 2009; Lambov et al., 2011) with accuracies between 65.8% to 86%.

1.1.2 Unsupervised learning

In unsupervised learning no desired example output datasets are provided - instead the algorithm defines the output based on input patterns (Ghahramani, 2004). Unsupervised learning techniques have the advantage of needing little labelled training data for sentiment classification. Unsupervised learning has been employed by Turney to classify reviews into "thumbs up" or "thumbs down". Turney used just two seed words[2], "excellent" and "poor", and calculated the polarity of a phrase with an algorithm[3] that measured the distance of pairs of words to both of the two seed words (Turney, 2002). His method achieved 66% accuracy. Rothfels and Tibshirani attempted unsupervised learning without a lexicon. They used the linguistic intuition that positive items occur frequently after negation, for instance in the example "not good", and they occur even more frequently without negation. Their method statistically classified items according to this assumption, but the authors find that this type of classification is "consistently unimpressive" (Rothfels and Tibshirani, 2010). Other unsupervised methods include automatic clustering of adjectives into two groups via linguistic constraints on adjectives semantic orientation. Examples of such constraints are negations and conjunctions[4], where a negation reverses an adjectives semantic orientation (Hatzivassiloglou et al., 1997). After clustering, positive orientation is assigned to the class with the highest average frequency. This is based on previous research which shows that unmarked adjectives almost always have positive orientation (Hatzivassiloglou et al., 1997). Hatzivassiloglou et al. (1997) achieve an accuracy between 78% and 82% for different linguistic constraints. However precision and recall are not mentioned.

1.1.3 Supervised and semi-supervised learning

Supervised learning algorithms learn by training on a validated and classified set of data, for example words, sentences or texts and their polarities. Supervised sentiment analysis is often combined with feature extraction by means of POS-tagging and heuristic linguistic rules. Supervised and semi-supervised methods frequently start by building a sentiment lexicon or dictionary automatically and then attempt to classify sentiments with the lexicon as the basis (Turchi and Balahur, 2013; Taboada et al., 2011; Wei et al., 2012; Neviarouskaya et al., 2009). Using sentiment lexicons has the disadvantage that opinion words are being classified without context. Yet, 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 (Pang and Lee, 2008).

[2]seed words are words which are at the core either 'positive' or 'negative' independently of context
[3]Pointwise Mutual Information and Information Retrieval
[4]example of conjunctions are "and", "but"

Statistical approaches to sentiment analysis use syntactic patterns and word frequencies in order to infer context (Medhat et al., 2014; Pang and Lee, 2002; Psomakelis et al., 2014; Santos and Gatti, 2014). Probabilistic classifiers, such as Naive Bayes, Bayesian Networks and Maximum Entropy classifiers use statistical approaches to find co-occurrence frequencies of unlabelled words and labelled words by training on large amounts of texts (Psomakelis et al., 2014; Rice and Zorn, 2013; Cao and Zukerman, 2012). Another example of a statistical classifiers are support vector machines (SVM). They make decisions based on a linear combination of feature-value pairs. SVM have consistently delivered best results for sentiment classification in previous research (Vinodhini and Chandrasekaran, 2012; Vohra and Teraiya, 2013; Patil et al., 2014). Vinodhini and Chandrasekaran (2012) for example reports different studies with SVM which resulted in accuracies of 86.4% to up to 93% for movie and restaurant reviews. Vohra and Teraiya (2013) informs on SVM results from 82.9% for IMDB reviews up to 95.55% on forum postings.

Vohra and Teraiya (2013) 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". Reasons for this conclusion are not presented, but with the information given on previous research, it could be inferred that supervised classification is better adapted to context than unsupervised classification.

1.1.4 Sentiment analysis in different languages

Years of ongoing research in sentiment analysis for the English language has produced a lot of openly available training data such as the SentiWordNet lexicon (Esuli et al., 2006), the Bing Liu Opinion Lexicon, the MPGQ Subjectivity Lexicon, the Harvard General Inquirer and many more (Potts, 2011). While enough data is available to inspire research for classifier improvement in English (Lambov et al., 2011; Kaufmann, 2012), there seems to be a lack of resources and research in sentiment analysis for other languages as Medhat et al. (2014) concludes.

Previous research about sentiment analysis in other languages than English has attempted to translate English sentiment lexicons into other languages or to translate texts to English and then analyse them for sentiments by applying trained English sentiment models[5] (Denecke, 2008; Kim and Hovy, 2006). Balahur and Turchi (2012) use machine translation to French, German and Spanish on a 20 topics question corpus and achieved a highest F-score of approximately 66% with SVM. Wan (2009) uses co-training for Chinese-English sentiment analysis and achieves accuracies up to 81% on product reviews. It has though been shown that the translation method may induce errors due to the differences in linguistic features in different languages (Hedlund et al., 2001). For instance word disambiguation by homonymy is frequent. Also other language-specific traits such as negation, idiomatic expressions and phrasal verbs[6], are difficult to capture with the translation method (Sumanth and Inkpen, 2015; Liu, 2012). In addition sentiment analysis programs often use English specific part-of-speech taggers or parsers, which means that sentiment models are not readily transferable to other languages. Sentiment analysis research for other languages frequently include to automatically create sentiment lexicons. This method is faster, but less precise than a manual creation of lexicons. The validation or cleaning of such lexicons is time consuming (Liu, 2012). Additionally, auto-

[5]For a definition of 'sentiment model', see Appendix A

[6]An explanation of idiomatic expression and phrasal verbs is given in Appendix A

matically collected words are domain independent(Liu, 2012), and as mentioned before, domain independent sentiment analysis generally performs worse than domain specific models.

Abdul-Mageed et al. (2011) takes account of language specific morphological features for Arabic and achieves F-scores up to 95% with the SVM light package on the PATB dataset. Jang and Shin (2010) does the same for Korean and claims it approves the accuracy. Nonetheless, other research using translation, parallel corpora and similar methods with high accuracy (Kim and Hovy, 2006; Mihalcea et al., 2007) suggest that language specific traits might not play an important role in sentiment analysis. Although many researchers agree that sentiment analysis is influenced by language specific dimensions (Hogenboom et al., 2014), the impact of those on the results of different sentiment analysis models it is still not quite understood.

The literature review shows that domain specific models often result in a higher sentiment classification accuracy than general models. Supervised learning seems to perform worse than semi-supervised learning on comparable domains in sentiment analysis. Amongst the classifiers for sentiment polarity, SVM seems to outperform other algorithms. Based on the outcome of the literature review, it seems that good sentiment classification accuracy can be expected when applying domain specific semi-supervised sentiment classification of texts using support vector machines. Therefore this approach is taken in the thesis which continues and contributes to previous research firstly by instantiating a Swedish SVM model, and secondly by implementing language specific handlers in order to understand their impact on the classification results of the SVM model.

1.1.5 Feature selection and weighting

Sentiment analysis is text classification and machine learning in text classification usually entails handling a large feature space. Features in text classifications are usually words. They can be single words or they can consist of several words together. The bag of word model considers a document to be a bag of independent words, i.e. each feature is one word, whereas n-grams can make use of dependencies of words, by considering sequences of words as being one feature. Since the features space is large, a relevant feature selection can improve the performance of sentiment analysis in terms of accuracy and classifier training time.

According to Tahir et al. (2007) feature selection lays aside irrelevant features while the remaining selection retains their original physical interpretation. Feature weighting is a technique in which the original features are multiplied by a weight value that is used to approximate the optimal degree of influence of the individual features (Keyvanpour and Tavoli, 2012).

Wettschereck et al. (1997) found that "feature selection algorithms perform best when the features used to describe instances are either highly correlated with the class label or completely irrelevant. Feature weighting is more appropriate for tasks where features vary in their relevance [...]".

The feature selection and weighting in this study is based on the findings from the literature review.

1.2 Research problem

As described above, labelled data is often the basis for (semi-) supervised sentiment classification models. For the English language many sentiment lexicons and sentiment mod-

els exist (Potts, 2011). However, there is a lack of resources and researches in sentiment analysis for other languages (Medhat et al., 2014). For instance, there is no ready-to-use Swedish data for sentiment analysis. One rather crude method for Swedish word sentiment classification based on a bootstrapped sentiment lexicon of verbs, adjectives and adverbs has been attempted, but only on word, not sentence level (Ludovici and Bignon, 2015). For sentence and text level sentiment classification in Swedish, there is a lack of labelled training data. The difficulty to create labelled training data might also explain why there are no open-source sentiment classification models and tools for the Swedish language.

This thesis firstly addresses the lack of knowledge on how to feasibly implement a sentiment model[7] for Swedish text sentiment analysis and secondly, the lack of knowledge if Swedish language specific handlers could improve the overall results.

1.3 Research goal and questions

The research goal is to design, instantiate and evaluate a Swedish sentiment classification model. This research study attempts to answer the following main research question:

What is the performance of the designed Swedish sentiment classification model, measured with accuracy, precision and recall? [8]

The sub-questions are as follows:

1. Which variation of the parameters kernel type, feature weighting, scaling, grid search and number of sentences increases the accuracy, precision and recall of the created Swedish sentiment model?
2. Does the inclusion of language specific handlers improve the classifier?

1.4 Contribution

This research contributes with a prototype of a functional open source Swedish sentiment classification model. It takes account of typical Swedish idiomatic expressions, negations and phrasal verbs.

It is expected that this research will increase knowledge on which factors are important in refining a Swedish sentiment model and lead to insights on how to conduct more advanced sentiment analysis of Swedish documents for specific domains.

1.5 Limitations of the study

Sentiment analysis will be done for the genre 'newspaper article' only and for the two classes 'positive' and 'negative'.

[7] For a definition of 'sentiment model' see Appendix A

[8] For explanation of the terms, see Appendix A

2. Method

In this chapter the Design Science Research strategy and methodology is presented. In addition the methodologies stepwise applications on this study are explained. Furthermore, the FEDS evaluation framework is used for evaluation of the created artifact. The theoretical FEDS model and its practical application are described in this chapter as well.

2.1 Design Science Research (DSR) and the alternative Action Research (AR)

Design Science Research is also known and discussed in research under the names "Design Research" and "Design Science". It has been called a research paradigm (Baskerville, 2008), a research orientation (Iivari and Venable, 2009), or a research approach (Järrvinen, 2007; Cleven et al., 2009; Goldkuhl, 2013; Alturki et al., 2013). Design science has developed from the area of information systems and there is no consensus for a standard definition. One example definition is the following by Johannesson and Perjons (2012): "Design science is the scientific study and creation of artefacts as they are developed and used by people with the goal of solving practical problems of general interest". In general, the researchers on design science mentioned above seem to agree that tasks for DSR include to the design, instantiation and evaluation of an artifact with the intent to solve a problem and that DSR is a validated and accepted research activity.

Similar to Design Science Research, Action Research (AR) has been called a research activity (Iivari and Venable, 2009), a research approach (Goldkuhl, 2013; Järrvinen, 2007; Cleven et al., 2009; Alturki et al., 2013), a methodology (Farooq and O'Brien, 2015), or a research strategy (Johannesson and Perjons, 2012; Denscombe, 2003). In contrast to DSR, AR originates from the field of social sciences (Goldkuhl, 2013). Again, there is no consensus on a definition. An example of a definition for AR given by Hult and Lennung (1978), includes many broad aspects, such as solving a practical problem, contribution to research and the researchers knowledge, in a collaborative and ethical process.

It must be assumed that DSR and AR are not fundamentally different in their nature, since much research is available that compares and contrasts DSR and AR. Nonetheless, the titles of two research papers which compare DSR and AR, "Action Research is Similar to Design Science" (Järrvinen, 2007) and "Action research and design science research - Seemingly similar but decisively dissimilar" (Iivari and Venable, 2009), are good examples of the two opposite poles within the scientific discourse surrounding DSR and AR.

Researchers often find dissimilarities between the two research strategies when examining different cases of AR and DSR applications, since not all cases of applied AR require the creation of an artifact. This is for example true if the AR problem is a research problem rather than a practical problem (Iivari and Venable, 2009). In contrast, DSR always involves the creation of an artifact.

However, in cases where AR involves the creation of an artifact, then the two research strategies seem to become almost interchangeable. The main difference between AR and DSR in that case, seems to be that the researcher in AR develops artifacts in collaboration with, and for a specific client. This means that although AR can contribute with knowledge to research, the problem that is being solved is often context specific. DSR on

the other hand, is not as context dependent and aims to solve a general problem with the artifact that does not have to involve external people in the process. (Iivari and Venable, 2009; Goldkuhl, 2013; Cleven et al., 2009; Johannesson and Perjons, 2012)

In conclusion, both the AR and DSR may accommodate the design and development of artifacts and seem to be state-of-the-art and well established research strategies in academia.

This study is of quantitative and inductive nature since large amounts of data are analysed in order to arrive to a conclusion on which parameters increase the overall SVM accuracy, precision and recall. As formulated in section 1.2, the thesis addresses the lack of knowledge on how to construct a sentiment model for Swedish texts. It is therefore highly focused on design and creation of artifacts, i.e. training and test data, as well as a conceptual sentiment analysis design from which the physical program to execute sentiment analysis has been developed. Therefore both AR and DSR could be applicable, but it seems as if DSR is the research strategy that is better suited for this thesis. DSR is the better choice of research strategy for this thesis since the thesis result should contribute to a general field rather than to a client specific problem and this research does not necessitate a participatory design and development process with external people. A short literature review confirms that researchers have previously adopted the design science approach for sentiment analysis (Lau et al., 2010; Eickhoff, 2015; Park et al., 2013).

2.2 Design Science Research Methodology (DSRM)

A research methodology, according to Crotty (1998), is "the strategy, plan of action, process, or design lying behind the choice and use of particular methods and linking the choice and use of methods to the desired outcomes". Howell (2013), the author of "The Philosophy of Methodology", defines methodology as a strategy that describes a way to structure research. In the light of those two definitions, the Design Science Research Methodology (DSRM) is the strategy used to describe and structure the approach and methods used in this research project.

The DSRM process model can be seen in figure 1.

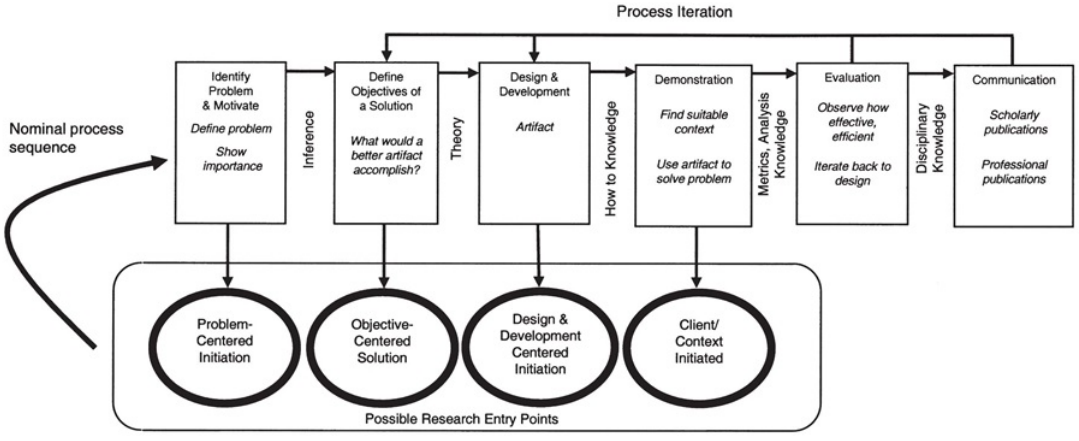

Figure 1: DSRM Process Model

The DSRM process methodology consists of six steps:

1. Problem identification and motivation: A research problem is defined and the value of a solution is motivated

2. Define the objectives for a solution: Through knowledge of the research area, define and argue the expectations of a solution
3. Design and development: Artifact creation either by conceptual design or physical instantiations
4. Demonstration: Usage of the artifact to solve a defined problem, for example by experimentation
5. Evaluation: Observations and measurements of the capability of the artifact to solve a defined problem.
6. Communication: Overarching short discussion of the the before five DSRM steps in order to structure and summarize the research.

2.3 Application of DSRM

2.3.1 Problem identification and motivation

The DSRM process starts with problem identification and motivation. An example of the use for sentiment analysis is to identify if texts are positive, negative or neutral, in order to leverage this knowledge for improvements to customer relation or trends prediction. In Swedish there is no almost no available training data to train classification programs for sentiment analysis of Swedish texts. An exception is a short research study by Ludovici and Bignon (2015). Without language specific training data, the practical applications of Swedish sentiment analysis are not available and research within sentiment analysis can not be conducted in Swedish. Furthermore there is no available sentiment analysis tool to transform Swedish text input through training to useful sentiment information. The mentioned facts are motivation for this research to create labelled sentiment training data[1] in Swedish and furthermore to build and evaluate a tool for sentiment analysis of Swedish texts.

2.3.2 Objectives for a solution

The second step of the DSRM requires to define the objectives for a solution. The FEDS evaluation framework is used for evaluation of the objectives described in this section. It has to be noted here that the FEDS evaluation framework uses the term 'requirement' which describes the same concept as the term 'objective for a solution' within the DSRM framework. This is why both 'requirements' and 'objectives' for a solution are used synonymously in this thesis. In order to define objectives for a solution, firstly comparisons to outcomes of similar research are drawn, after which the objectives are defined.

Comparison to similar research

Previous research about sentiment analysis in other languages than English has attempted to translate English sentiment lexicons into other languages or to translate texts to English and then analyse them for sentiments by applying trained English sentiment models (Denecke, 2008; Kim and Hovy, 2006). Accuracies between 62% and 68% were achieved with precision between 55% and 68% and recall between 40% and 80% for the domain of movie reviews and emails with SVM. Balahur and Turchi (2012) use machine translation to French, German and Spanish on a 20 topics question corpus and achieved a highest F-score of approximately 66% with SVM. Wan (2009) uses co-training for Chinese-English

[1] In this study, 'labelled' means the classification of newspaper articles into the categories positive and negative

sentiment analysis and achieves accuracies up to 81% on product reviews. Abdul-Mageed et al. (2011) takes account of language specific morphological features for Arabic and achieves F-scores up to 95% with the SVM light package on the PATB dataset. Research like Abdul-Mageed et al. (2011) which take account of language specific features is scarce in sentiment analysis. The hypothesis in this research project is that a special handler for Swedish language specific features, including idiomatic expressions, phrasal verbs, and negation handlers[2] will lead to a higher classification performance than without such a handler. A comparison to similar research in sentiment analysis with SVM shows that sentiment classification for domains which are rather broad can result in lower accuracies between 60% and 70%, whereas domains with less features or specifically tuned sentiment analysis tools as in Abdul-Mageed et al. (2011) may achieve higher accuracies.

Objectives/Requirements for a solution

Since the domain of news articles can be considered large with over 40 000 observed different features. Therefore a comparison with results to research on general domains and results on the lower end of specific domains could mean realistic objectives for this thesis. A realistic objective for the Swedish sentiment classification prototype (without handlers) would therefore be to reach a sentiment classification accuracy between 65% and up to 85% by varying the critical parameters described in 2.3.3. All accuracies over 75% for the news article domain could be considered good and even higher than previous research on large feature domains 2.3.2. The gold standard for accuracy is given by human evaluation of the training data set, and by human evaluation of the test set.

The requirements of a machine learning accuracy between 65% and about up to 85% for the evaluation of both validation and training data set, and the final test set, are measured in regard to the human gold standard.

Furthermore, a hypothesis was that the implementation of language specific handlers increases the overall measured accuracy of the Swedish sentiment classification model. Given that all other parameters are hold constant, a requirement that would prove this hypothesis to be true, is that the inclusion of language specific handlers increases the total accuracy of the model.

Furthermore requirements of an acceptable Swedish classification prototype includes that the classification process can be finished within a reasonable timeframe, i.e. not more than 12 hours, and within reasonable memory use, i.e. not more than 8 GB.

2.3.3 Design and development

The third step in the DSRM process model is the design and development of the artifact. In this section, an overview is given on the design and development of the Swedish classification model, see figure 2. The arguments for methodological choices and design choices are brought forward in this chapter, while the technical details of the design and in-depth explanations are described in chapter 3.

Design choices and rationale for the classification model

Figure 2 has four main sections: Preparation, creation of training data, preprocessing and training and test.

The preparation step did not have to include topic selection, but then the domain for sentiment classification would have been "news", which is very broad. This would have

[2] An explanation of idiomatic expression and phrasal verbs is given in Appendix A

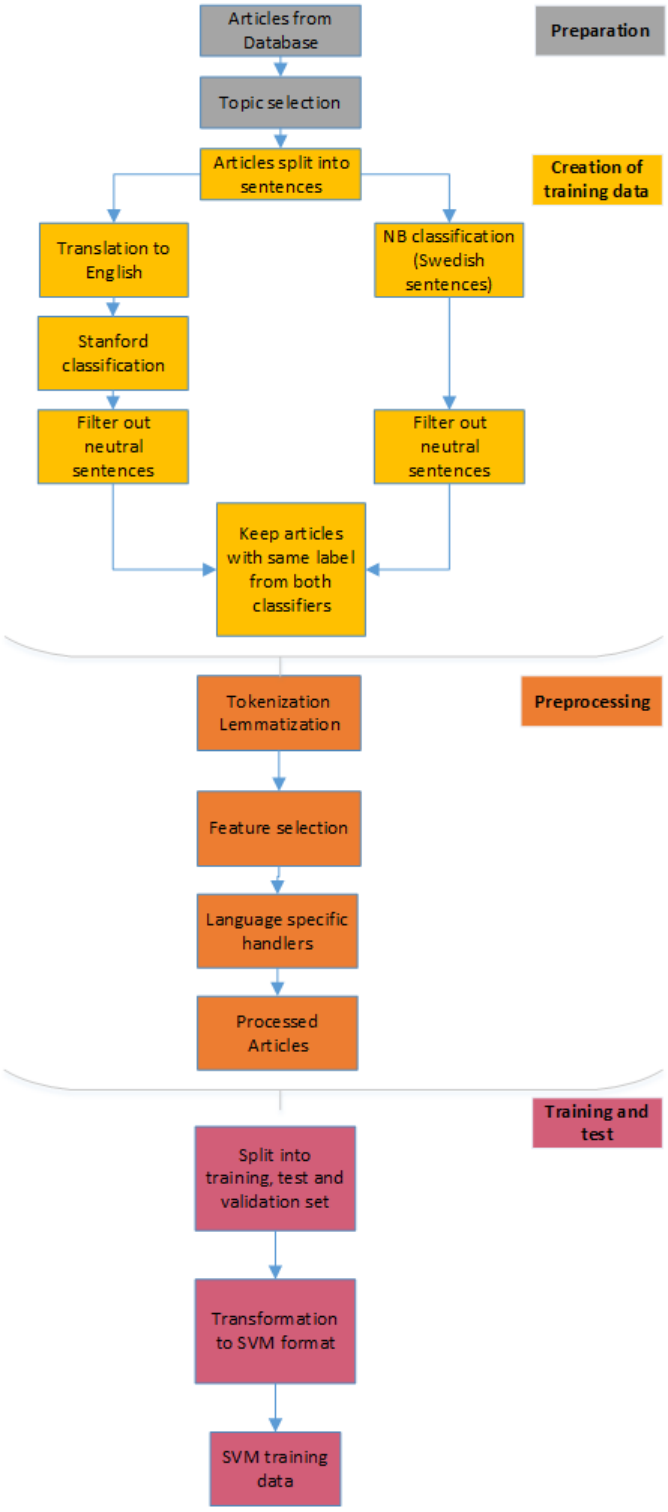

Figure 2: Structure of the Swedish classification model

had two distinctive disadvantages: Firstly it would have been more time consuming to create the necessary training data and to finally train and test the model. Secondly, the results of training on such non-contextual data would be questionable, since news is seldom categorical and even humans might have difficulty to judge positive from negative without any context.

In the next step, the topic-filtered selection of articles is passed in parallel to two classifiers, in order to create sentiment labelled training data.

One of the classifiers, the Stanford classifier requires translation to English before being able to give sentiment labels. Since translation to another language works better when sentences are given in their whole context, it is decided that the preprocessing step should begin after the creation of labelled training data instead of before.

Regarding methods for the creation of labelled training data, the choices are limited. Manually created training data tends to be the most correctly labelled data, but the process is very time consuming. Therefore automatic creation of labelled data is desirable, while trying to keep labelling errors at a minimum. The method to automatically create labelled training data in this research is the use of two classification algorithms: A Swedish classifier based on Naive Bayes and the Stanford RNN sentiment classifier.

The Swedish classifier functions on word level. It uses a lexicon with positive and negative verbs, adjectives and adverbs at its basis. The lexicon has been bootstrapped in two iterations, by searching synonyms for a defined list of positive and negative words, which then were added to the original list. It was measured to reach an average accuracy of 89% and an F1-score of 0.92. It has therefore been deemed reliable when it was evaluated on word level sentiment classification (Ludovici and Bignon, 2015), while the Stanford classifier has an accuracy level of text classification up to 85.4% (Socher et al., 2013) on sentence level classification. Although the Swedish classifier has not been evaluated on sentences, it uses the only existing and validated Swedish sentiment word lexicon for classification and was therefore an obvious choice for sentiment classification of Swedish sentences.

On the contrary, the Stanford RNN has been validated many times with good results on English sentiment data (Erik et al., 2013; Quoc V and Mikolov, 2014; Li and Dong, 2014), which is why it has been chosen. Of course any other state-of-the-art classifier would have been a good choice. One problem that would occur for any classifier trained on English texts, is that translation from Swedish to English is needed which might induce language specific errors due to differences in linguistic features of two languages (Hedlund et al., 2001).

The design imperfections of creating the training data were addressed with the assumption that the risks for incorrect labels could be reduced by double confirmation of the labels: Only non-neutral sentences where both classifiers, which are very different in their underlying logic, agreed on the same label were kept.

The next step is the preprocessing step. It includes the methods of tokenization, lemmatization, feature selection and language specific handlers, as described in chapter 3.4. Preprocessing is often necessary, but the SVM is a classifier which can handle much information. Therefore, during the design of the model, the model performance was iteratively tested on the validation set with and without different preprocessing steps. The final order displayed in figure 2 is the result of the best accuracy achieved with different preprocessing steps and orders of preprocessing steps. It could for example be shown that feature selection improves the training results. Furthermore the language specific handlers could not be placed before the feature selection step, since the filtering of words with frequency 1 would have eliminated some of the created handlers.

The processed articles are then split into training, test and validation set[3]. The training and validation set are transformed into the libSVM format (section 3.8). Finally the SVM model is trained.

[3]For explanation of training, test and validation set, see Appendix A

Choice of classifier

The SVM classifier is chosen for this sentiment analysis project since it has repeatedly been proven to be the best performing algorithm for sentiment analysis tasks (Vinodhini and Chandrasekaran, 2012; Vohra and Teraiya, 2013; Patil et al., 2014).

Choice of weighting method

In sentiment analysis, words can be weighted by many methods, such as word count, term frequency or term-frequencey - inverse document frequency (TF-IDF)[4]. The weighting of relevant features in a text defines the impact of those features in a classification problem. A comparison of feature weighting methods for text classification with SVM (Sahlgren and Cöster, 2004; Salton and Buckley, 1988),have shown that TF-IDF performs better than word count or term frequency. Different weighting schemes can be used to calculate both TF and IDF.

According to Paltoglou and Thelwall (2010), the best results for sentiment analysis with SVM were delivered with a weighting scheme combination of o(BM25) TF and Δ(k), i.e. Delta BM 25 IDF [5].

Salton and Buckley (1988) research compares different TF-IDF weighting schemes and concludes that the tfc-nfx scheme is one of the best performing weighting schemes.

Pang and Lee (2002) achieve an accuracy of 87.15% with the SVM classifier and unigrams, using the simplest TF-IDF weighting scheme, the binary TF-IDF

Therefore this research will experiment with the three weighting schemes o(BM25) TF and Δ(k)IDF, section 3.7.1, the TF-IDF as described in section 3.7.2, and the binary TF-IDF in section 3.7.3.

Choice for variation of parameters

In the structure of the Swedish classification model, many parameters can be potentially varied.

Firstly it was chosen that classification in the creation of training data should be done sentence-wise, since it is expected to result in a higher accuracy than an article-wise classification. The underlying reasoning is that every article can have both negative and positive sentences, and an aggregation of those sentence-sentiments in an article results in a less precise representation of sentiments and consequently training an SVM on article-wise sentiments will be less precise as well. The choice of article versus sentence-wise classification is not varied.

Tokenization, lemmatization and feature selection methods do not affect the sentiment of sentences, since they affect non-sentiment bearing words and words, which, because of their frequencies, do not bear a significant weight for the classification process. In addition, it was deemed necessary to preprocess the data with above methods in order to reduce the SVM feature space. A large feature space is problematic for the performance of most SVM (Joachims, 1998), therefore including feature filtering steps is necessary. These three preprocessing methods are not varied.

It has not conclusively been shown if language specific handlers affect the sentiment analysis results. Therefore inclusion versus exclusion of handlers is tested in order to determine the impact on sentiment classification results.

[4] An explanation and formula is given in section 3.7

[5] according to the SMART notation for term frequency and inverse document frequency variants

Chih-Wei et al. (2008) hints that there are no rule-of-thumbs for the best parameters of a SVM. Therefore, the beneath parameters are important and have to be determined for each individual dataset. The following parameters[6] are varied iteratively in order to find the best SVM model:

1. kernel type
2. feature weighting
3. scaling
4. C (explained in section 3.8)
5. gamma (explained in section 3.8)
6. number of sentences

2.3.4 Demonstration

Stage four of the DSRM is to demonstrate the use of the created artifact for problem solving.

The best classification SVM model and how it was found by varying different parameters is demonstrated by visualising the results in tables and figures, see chapter 4.

2.3.5 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 described objectives for a solution are compared to the actual results 2.3.2. A Framework for Evaluation in Design Science Research (FEDS) is used for evaluation of the created artifacts.

2.3.6 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 classification model are discussed. This part of the work corresponds to the discussion section (5) of this thesis.

2.4 The FEDS evaluation framework

For evaluation, help is drawn from the FEDS evaluation framework of Venable et al. (2016). The evaluation framework contains four steps:

1. Explicate the goals
2. Choose a strategy or strategies for the evaluation
3. Determine the properties to evaluate
4. Design the individual evaluation episode(s)

2.5 Application of the FEDS evaluation framework

2.5.1 Explicate the goals

Venable et al. (2016) explains four competing goals for designing the evaluation of an artifact.

[6] the parameters are explained in chapter 3

The first goal is called 'rigour'. The first purpose of rigour is thus to "determine how well a designed artefact or ensemble of artefacts achieves its expected environmental utility (an artefact's main purpose)" (Venable et al., 2016). This can be achieved by measuring the performance of the artifact in comparison to its purpose. The second purpose of rigour is to evaluate if the "substantiation of design theory leads to some developed artifact that will be useful for solving some problem or making some improvement." (Venable et al., 2016) Rigour and has two aspects: Firstly, rigour ensures that the artifact instantiation and not other variables or circumstances cause an observed outcome. Secondly, 'rigour' means to check that the artifact instantiation functions in a real life situation.

A second goal of designing an evaluation can be 'uncertainty and risk reduction'. By that Venable et al. (2016) means to identify design uncertainties and thus to reduce risks due to design and to identify bias that could reduce the reliability and validity of the evaluation. One way to evaluate uncertainty and risks.

A third goal of designing an evaluation is called 'ethics' and concerns the evaluation of safety and risks to the environment, for example people and organisations, but also future generations.

The fourth goal of designing an evaluation concerns efficiency, i.e. the goals of the evaluation are held against the resources needed.

The goals in this research mostly concern the two aspects of rigour, which are aligned with the research question and subquestions. Nonetheless, induced bias in the evaluation process and uncertainties in the design, research ethics and the efficiency of the resources needed will be discussed in the discussion chapter of this thesis.

2.5.2 Choose a strategy or strategies for the evaluation

According to Venable et al. (2016), the purpose of formative evaluation is to "help improve the outcomes of the process under evaluation". Summative evaluation means to "judge the extent that the outcomes match expectation". As the name describes, summative evaluations arrive at a final evaluative interpretation by contrasting consequences and meanings from all project phases.(Venable et al., 2016) Formative and summative evaluation help to meet the chosen design goal 'rigour' and will be employed in this thesis.

The paradigms which the evaluation in this research follows are both artificial evaluation and naturalistic evaluation. The outcome of an artificial evaluation is used to measure the first aspect of rigour in the design goals of an evaluation: to determine what causes an observed outcome, i.e. by laboratory experiments and simulations. The outcome of a naturalistic evaluation measures the second aspect of rigour: check if the artifact may be used for solving real life problems. Naturalistic evaluation is often empirical and interpretivistic and can for instance employ surveys or comparisons to similar research to verify the second aspect of 'rigour'.

Considering the above mentioned aspects, of the four evaluation strategies proposed by Venable et al. (2016), a hybrid between the " Technical Risk & Efficacy evaluation strategy" and the "Quick & Simple strategy" are employed.

The "Technical Risk & Efficacy evaluation strategy" focuses on artificial formative and iterative evaluations early in the design process, to help improve the outcome under the design process. This strategy is used to determine the best parameters for the SVM through iterative experiments, where one parameter will be changed at a time while the others are held constant. This evaluation is done on the training and validation set. Fur-

thermore, in a second round of iterations, combinations of the best individual parameters will be tested regarding their impact on the classification results.

The "Quick & Simple strategy" also seems appropriate for this research since it focuses on summative and naturalistic evaluations, i.e. evaluations including humans and other resources, such as literature. For the summative and naturalistic evaluation of the Swedish sentiment analysis classification model, a sample of sentences from the test set is classified by three native Swedish speakers. Kim and Hovy (Kim and Hovy, 2004) uses three 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 (Denecke, 2008). Three annotators were therefore deemed appropriate for evaluating a sample of 399 article sentences. The human classified sample of sentences is evaluated once by each annotator and the total of sentences manually classified corresponds to a confidence level of 95% with a confidence interval of 5%. Although the process of sentence labelling 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 (2002, pp.35).
Comparison to the results of similar research are also drawn. Differences, similarities and expectations are discussed and the results are compared to the objectives of this research.

The "Human Risk & Effectiveness evaluation strategy" can be sorted out since it has an emphasis on formative evaluation which corresponds to different evaluation goals than the ones chosen in this project. The "Purely Technical strategy" can also be excluded since it excludes naturalistic evaluation and therefore it excludes humans in the evaluation process. However, human evaluation is necessary in sentiment analysis, since there is a lack of a gold standard for Swedish sentiment analysis and therefore human classifications will be the gold-standard of sentiment analysis result comparisons.

2.5.3 Determine the properties to evaluate

The properties to evaluate include the "set of features, goals, and requirements of the artifact" (Venable et al., 2016). The features which are varied are summarized in section 2.3.3 and described in detail in section 3. The goals are of the artifact are described in section 1.3. The requirements of the artifact are described in section 2.3.2. It has to be noted here that the FEDS framework uses the term ’requirement’ synonymously to how the DSRM framework uses "objective for a solution".

2.5.4 Design the individual evaluation episode(s) with validity and reliability in mind

Designing the individual evaluation episodes includes to consider constraints in the environment, to prioritise, and to determine the exact steps in the evaluation process.

The evaluation process can be seen in figure 3 and is explained in detail in the chapter ’The Artifact’, see 3.10.

As can be seen in the evaluation model, both formative artificial and summative naturalistic evaluation is done.

The validity of having chosen the best parameters, i.e. the relevance of the chosen parameters, is assured by the iterative search process described in section 4.1. The validity of the training data, as well as of the results of the SVM model is strengthened through comparison to human classification of sentence samples.

The reliability of the humans classification of Swedish sentences is strengthened by using adult Swedish native speakers for the classification tasks. The three volunteers had

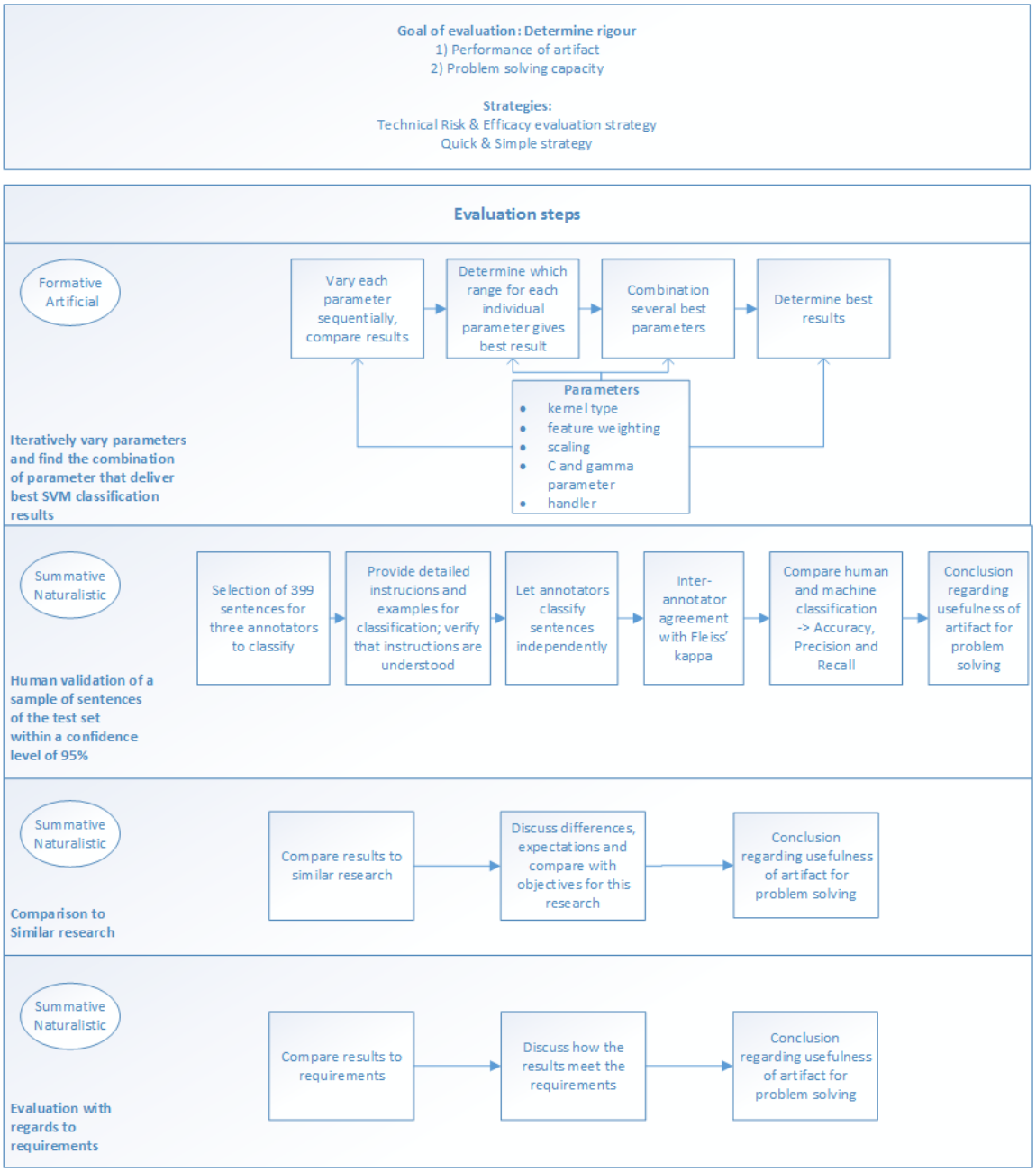

Figure 3: Structure of the evaluation flow

a month time to anonymously classify both the training data and the validation set of sentences through a website.

The validity of the results in regards to the requirements is shown through measurements and finally, the validity of the overall results of the SVM model can be seen through comparison with similar research.

The nature of this thesis, including design and instantiation, entails a strong reliability, since the exact same training data and SVM results will be achieved, if the model is used with exact the same code and parameters. For the same reasons, transferability is high.

2.6 Research Ethics

Research should be ethical, i.e. the researcher needs to firstly respect the rights and dignity of participants, secondly avoid harm, and thirdly operate with honesty and integrity (Denscombe, 2003). All above three principles are observed. No private data or other personal information has been used for this study, since the annotators are anonymous. Sentences for the validation set are classified out of context, i.e. no personal statements are selectively classified. According to Denscombe (2003), the impersonal nature of the article classification task is grounds for not seeking informed consent. No incentives for any participation have been offered and the three volunteers who classified articles by hand are treated confidentially during and after the study. All participants had enough time to classify the given data. No software licensing agreement have 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'.

3. The Artifact

In this chapter the creation of the artifacts, i.e. training data and the Swedish sentiment classification model are described in detail.

3.1 The Material

The SVM model is built in Java in the IntelliJ IDEA(JetBrains, 2015), and the libSVM package was used (Chang and Lin, 2015). The dataset used comes from the MittMedia PostgreSQL article database. The database contains articles from 2002 up to 2015. The articles for this research project were taken from the five regional newspapers with the most articles in the database:

1. Östersunds Posten (128 692 articles)
2. Dalarna Tidning (416 149 articles)
3. Sundsvalls Tidning (347 122 articles)
4. Gefle Dagblad (170 388 articles)
5. Hela Hälsingland (163 945 articles)

It should be mentioned that the five biggest newspapers to which the author had access to, are all liberal newspapers which certainly has an influence on how negative or positive certain topics are mentioned in the articles. However, this research is not concerned with the political stance that the newspapers take towards certain topics, but only with a correct classification of negative and positive sentences, which can be found in any newspaper about the topic 'school'.

3.2 Topic filtering

The original dataset of articles from five newspapers was vast with a total of $1,226,296$ articles. To reduce the amount of data and to narrow the broad domain of newspaper articles, the method of topic filtering was applied. Articles from the specific topic 'school' were filtered out[1] with the help of keywords and keyword combinations, see Appendix B. The keywords were selected from a frequency list of unique word occurrences for all articles. From this list, the most frequent 23 nouns in relation to the topic 'school' were chosen manually. The choice of keywords has a direct influence on the choice of sentences in the sentence filtering process, which means that more negative than positive sentences were filtered out since more negative than positive keywords were found. However the same argument as above for the choice of newspapers applies: This research is not concerned with the amount of negative or positive labelled sentences, but only with a correct sentiment classification. In addition articles filtered with a negative keyword might still have the same amount of negative as positive sentences.

The process of topic selection left $6,865$ articles after topic selection (about $252,975$ sentences with an average of 15 sentences per article).

[1] Any topic that is mentioned in newspapers could have been used as filter. The topic 'school' was chosen as an example.

3.3 Process for creation of labelled training data

After topic filtering, labelled[2] training data was created. For creating labelled training data, a sentiment classifier by Ludovici and Bignon (2015) and the Stanford RNN sentiment classification algorithm created by Socher et al. (2013) were applied.

The classifier by Ludovici and Bignon (2015) uses a Swedish sentiment lexicon with labelled verbs, adjectives and adverbs at the basis in order to classify individual Swedish words.

The Stanford sentiment classifier is trained on English sentences (Socher et al., 2013) and therefore the Swedish newspaper articles had to be translated[3] before classification.

Sentiment classification with both classifiers was done sentence-wise. After all articles have been labelled by both classifiers, neutral articles were sorted out. Of the set of positive and negative sentences, only the sentences where both classifiers produced the same sentiment labels, were deemed reliable enough to be used as training, test and validation data for the SVM.

3.4 Preprocessing

In the preprocessing stage tokenization, lemmatization, feature selection and the use of language specific handlers were employed, which are described in this section.

3.4.1 Tokenization and Lemmatization

Tokenization includes breaking up a text in smaller units such as words or sentences. In this study, whitespace is removed and words separated with dashes are split into two words. Every character that is not letter or numbers is removed as well. Lemmatization is the process of reducing a word into its base form, such as 'house' is the base form of 'houses'. Lemmatization was done with the Swedish maltparser (Johan et al., 2012).

3.4.2 Feature selection

Feature selection might help to improve accuracy and to reduce the feature space. Feature selection is done on word level where a text is seen as a 'bag of words' and the frequency of each word is used for feature selection. Tahir et al. (2007) found that feature selection works well on features that are irrelevant to the semantic or structural elements of sentences. In the sentiment model for this thesis, word frequencies are calculated and all words with frequency of 1 are removed, since their weight is insignificant for the overall sentence sentiment. Words with frequency of 1 may even represent word misspellings, place or people names that are not in the list.

Furthermore, if a word is all too frequent, that indicates that it is a stop word. Stop words are fill-words that are irrelevant to a texts meaning or sentiment, such as the Swedish word 'liksom' (translates to the comparative use of 'like'). Words with high frequency are found and matched to the stop word list provided by Doyle (2016).

Name filtering is also employed. People are often not the focus of a directed sentiment, which is why their impact on the sentiment analysis is considered negligible in this research. Names were removed with help of name lists.

[2] 'labelled' is used synonymously with 'classified'

[3] by Google Translate

3.4.3 Creating handlers for language specific traits

Language specific handlers were constructed as methods to adapt sentiment analysis to Swedish specific language traits in the hope to thereby elevate accuracy of the SVM. Idiomatic expressions and phrasal verbs[4] are specific to one language and not likely to appear with the same form or meaning in another language (Sjöholm, 1993). In comparison to the English language, the Swedish negation may precede or follow a verb depending on its appearance in a main or sub clause. Common to idiomatic expressions, phrasal verbs and negation is that they involve several words which provide meaning only when considered together. Therefore the main idea with creating Swedish handlers is to aggregate group of words into one logical unit and pass them as one word to the SVM, in order to treat them as connected in relation to the rest of a clause.

Idiomatic expressions

Ludovici and Bignon (2015) use a list of idiomatic expressions from the largest online list of Swedish idiomatic expressions on Wikipedia (*Lista över svenska idiomatiska uttryck*, 2015) whereof 273 idiomatic expressions contained sentiment and were subsequently classified as positive or negative by three Swedish natives and added to a lists idiomatic expressions. The idiomatic expression handler checks if expressions in the sentences are matching an expression in the list. If that is the case, and the idiomatic expression is classified as positive, the whole expression is replaced with the Swedish word for 'good' (bra), if the expression is negative, it is replaced with the Swedish word for 'bad' (dålig).

Negation

Negations are used in order to contradict or deny a statement. Commonly used Swedish negations are introduced with 'inte' and 'ej'.

The model built for this thesis recognizes those negations and prepares for the introduction of negation rules.

The negation rules only affects sentiment bearing verbs and adverbs. Those verbs and adverbs in combination with 'inte', then reverse their original sentiment. An example with a sentiment bearing verb:
"Hon lider" (she suffers) "Hon lider inte" (she does not suffer) An example with a sentiment bearing adverb:
"Detta är bra" (this is good)
"Detta är inte bra" (this is not good)

Modal verbs and the verb 'to be' are excluded from the 'inte'-rule, since they bear no sentiment.

The negations 'inte' and 'ej' can be placed after a verb in a main clause and in front of a verb in a subordinate clause:
"Han tycker inte om resan" (main clause)
"Han sa att han inte tycker om resan" (subordinate clause)

The negation rule therefore encompasses a window of 3 words in front of and after inte. Sentiment bearing verbs, adjectives and adverbs are detected with help of a sentiment lexicon (Ludovici and Bignon, 2015). The word 'inte' and the affected sentiment bearing word are then replaced by only one word of the opposite sentiment class than the affected

[4] An explanation of idiomatic expression and phrasal verbs is given in Appendix A

sentiment bearing word. The words replacing the negation and sentiment bearing verb, adjective or adverb are 'bra' (good), respectively 'dålig' (bad).

In order to recognize verbs, adverbs and adjectives, sentences are passed to Stagger, a part-of-speech tagger developed by the Stockholm University (Östling 2016). The Stagger marks each word according to its grammatical property.

Phrasal verbs

Phrasal verbs are common in the Swedish language (Mattisson et al., 2000). They are verbs which are composed by one verb in combination with one or more particles, for example "göra bort sig" (make a fool of oneself). Each part needs to be taken account of for the expression to make sense (Mattisson et al., 2000). A list of phrasal verbs has been found in (*Swedish phrasal verbs*, 2015) and the program matches word combinations in the sentences to the list. Whenever phrasal verbs are encountered in an article, they are combined to one word by an undeline: göra_bort_sig. This way, the SVM is able to weight the phrasal verbs' importance as a total, instead of seeing one phrasal verb as two or more separate words without connection.

3.5 Training, validation and test sets

After the creation of labelled sentences, potential training, test and validation data[5] has been reduced to 57,936 sentences. Of those sentences, a subset of 400 sentences were set aside into a test set. The test set has not been used as training or validation data. 400 sentences were chosen since their classification by humans would represent the whole set of sentences with a 95% confidence level and 5% confidence interval. The test set stands for 0.69% of all sentences, the training set stands for 79.45% and the validation set stands for 19.86% of all sentences. The results of running the trained SVM model on the test set is finally evaluated in comparison to human classification of the same set, see chapter 2.3.5.

3.6 The SVM classifier

In the following section, a short explanation of the intuition behind the SVM classifier is given, after which a mathematical model of the SVM classifier is presented.

3.6.1 Description of SVM

The goal for the SVM in this research project is to classify topic-filtered news articles[6] into the sentiment categories positive and negative. Since the SVM is a supervised algorithm, it first needs to be trained and validated in order to produce a classification model that is fitted to the classification task at hand. For each topic category a new SVM classification model is trained with the same underlying process. Each article sentence constitutes a data point and this data point is expressed as a vector. These vectors represent the coordinates of the data points. The SVM algorithm tries to mathematically find a hyperplane H_0 that separates the given data points into the two categories. If a hyperplane is close to training data points of any category, the found separation of data might not apply any more if more data points are given. Therefore the SVM algorithm attempts to maximize the margin between the separating hyperplane and the data points from each category. To

[5]For explanation of the training, test and validation set, see Appendix A

[6]In this project, articles are filtered into the topic category 'school system'

this end, other possible planes H_1 and H_2 are calculated with the constraints that there should not be any data points in-between those two planes. Furthermore, amongst the solution space of possible planes that fulfil the first constraint, another constraint dictates that the margin between H_1 and H_2 should be maximized.

3.6.2 SVM mathematical model

Given a training set of input and output data pairs (x_i, y_i), where $i = 1, ..., l$, $x_i \in \mathbb{R}^n$ and $y \in \{-1, 1\}^l$, the problem is to find hyperplanes that separate the two categories $y_{i,j}$. The hyperplanes can be written as the dot product of two vectors, $w\langle -b, -a, 1\rangle$ and $x\langle 1, x, y\rangle$, where $\overline{\mathrm{w}}$ is always normal to the hyperplane. Therefore $\overline{\mathrm{w}} \bullet \overline{\mathrm{x}}$ is equal to $y - ax - b$. A constraint is imposed which separates the two categories:

$$\begin{aligned} \overline{\mathrm{w}} \bullet \overline{\mathrm{x}}_+ + b \geq 1 \\ \overline{\mathrm{w}} \bullet \overline{\mathrm{x}}_- + b \leq 1 \end{aligned} \tag{3.1}$$

When the above equations are multiplied with y_i, the two equations result in the combined constraint:

$$y_i[(\overline{\mathrm{w}} \bullet \overline{\mathrm{x}}_i) + b] - 1 \geq 0, \quad i = 1, 2, ...l \tag{3.2}$$

Sometimes, if the problem is not linearly separable, the margin that separates categories is not enforced, i.e. a slack variable ξ can be introduced, where C represents the penalty for wrongly classified variables:

$$\begin{aligned} y_i[(\overline{\mathrm{w}} \bullet \overline{\mathrm{x}}_i) + b] \geq 1 - C\sum_i \xi_i, \\ \forall i \xi_i \geq 0, \quad i = 1, 2, ...l \end{aligned} \tag{3.3}$$

To allow slack is optional and for ease of understanding, the explanation of the mathematical model behind the SVM algorithm will continue without the slack variable.

Equation 3.2 defines two hyperplanes separating the categories. The width between those hyperplanes is calculated by taking the difference between the vector $\overline{\mathrm{x}}_+$ and the vector $\overline{\mathrm{x}}_-$ and projecting it onto the unit normal to the hyperplane, i.e. $\frac{\overline{\mathrm{w}}}{\|w\|}$:

$$\left(\frac{1-b}{\|w\|}\right) - \left(\frac{-1-b}{\|w\|}\right) = \frac{2}{\|w\|} \tag{3.4}$$

As shown in fig. 4, the margin between H_1 and H_2 is of length $\frac{2}{\|w\|}$.

To maximize the margin between the hyperplanes equals maximizing $\frac{2}{\|w\|}$ which is the same as minimizing $\frac{\|w\|^2}{2}$.

Finding the largest margin hyperplane is now a quadratic optimization problem.

$$min\left(\frac{\|w\|^2}{2}\right) \quad s.t. \quad y_i[(\overline{\mathrm{w}} \bullet \overline{\mathrm{x}}_i) + b] \geq 1 - C\sum_i \xi_i \tag{3.5}$$

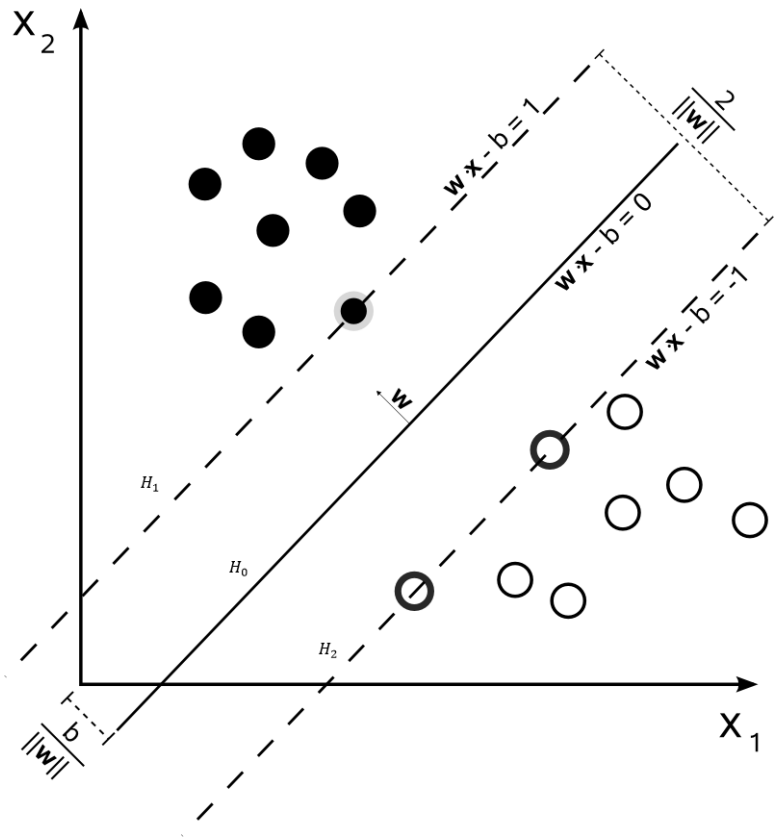

Figure 4: Maximum margin hyperplanes

The solution to the quadratic optimization problem can be found through the Lagrange multiplier, with a multiplier a_i for each constraint. The quadratic problem can thus be rewritten as:

$$L(w,b,a) = \left(\frac{\|w\|^2}{2}\right) - \sum_i^l a_i[y_i[(\overline{\mathrm{w}} \bullet \overline{\mathrm{x}}_i) + b] - 1] + \sum_i^l a_i, \quad a_i \geq 0 \quad \forall_i \tag{3.6}$$

This equation is referred to as the dual form, since w and b should be minimized while a should be maximized (Fletcher, 2009).

The equation 3.6 is derived with regard to $\overline{\mathrm{w}}$ and to b and set equal to zero to find an extremum. The results of both derivations are substituted in 3.6. These operations result in the following equation:

$$\begin{aligned} L_D &= \sum_i^l a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j (x_i \bullet x_j) \\ s.t. \quad &\sum_i^l a_i y_i = 0 \quad \& \quad a_i \geq 0 \end{aligned} \tag{3.7}$$

Those kind of problems are a convex quadratic optimization problem and solutions can be found with programs called "quadratic programming solvers" (QP). As equation 3.7 indicates, the dependence on w and b is removed by using Lagrangian multipliers, and the problem can now be solved by computing the dot product of x_i and x_j. The goal for maximization of L is to find x_i and x_j which are similar, but do not predict the same class.

If the dot product is 0, then the two vectors x_i and x_j are not similar (perpendicular) and do not contribute to the maximization problem. If the dot product is 1, then x_i and x_j are completely alike. If the two vectors predict the same class, for example y_j, then xy_i is always 1 and the value of $a_i a_j y_i y_j x_i \bullet x_j$ is positive which decreases L. If x_i and x_j predict different classes of y, then the dot product becomes negative, which means the term is added to L and maximizes the sum.

The SVM is especially useful for solving non-linear problems. This is done with the kernel trick. The kernel trick is a technical term used to describe the use of kernel functions, which are functions that map the sample space to a higher dimensional space in which a non-linear solution to a problem might be found. Each kernel function can be expressed as a dot product, which means that each SVM feature vector can be transformed by a kernel function. Kernel functions are therefore applicable for SVM text classification problems. The general equation of a kernel function is given in equation 3.8 follows:

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \tag{3.8}$$

libSVM offers four basic kernels, the linear kernel, the polynomial kernel, the radial basis function and the sigmoid kernel.

3.7 TF-IDF Weighting methods

The intuition behind the TF is the assumption that a term which appears often in one given document must be important for that document. TF thus assigns a higher weight to a term that appears often in a given document in relation to other words in that document and to document length (TF normalization) (He and Ounis, 2005). Zipf's law states that word frequencies follow a power law distribution, which means that their frequencies are inversely proportional to their ranks, i.e. terms that are common in all documents probably do not bear importance, for example the word 'the' (Powers, 1998). According to Zipf's law, the IDF reshifts the TF weighting by assuming that a term which is frequent in all considered documents is not a good discriminator (Robertson, 2004). It is therefore weighted less than terms which occur in few documents. To weight the relative importance of terms between documents, the IDF formula includes document counts - with and without the term under consideration - and in the IDF formula used in this research, even the class distribution for a given term within documents is considered.

Given a query Q for a document D, the general formula for the TF-IDF score is as follows:

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) * TF(q_i) \tag{3.9}$$

3.7.1 O(BM25) TF and Δ(k)IDF weighting scheme

Paltoglou and Thelwall (2010) showed that best results for sentiment analysis with SVM were delivered with a weighting scheme combination of o(BM25) TF and Δ(k) IDF (Delta BM 25 IDF). In this weighting scheme, each document D is considered being a bag of

words, i.e. each word is an own feature. N is the number of unique words and w_i is the weight of term i in document D, where $i = 1, ..., N$.

$$o(BM25)TF = \frac{(k_1+1)*tf}{k_1\left((1-b)+b*\frac{dl}{avg_{dl}}\right)+tf} \tag{3.10}$$

$k_1 = 1.2$ and $b = 0.95$ are default parameters of BM25 (Sparck Jones et al., 2000), tf is the term frequency of each term, dl is the document length and avg_dl is the average document length.

The following equation describes the calculation of IDF:

$$\Delta(k)(DeltaBM25IDF) = log\frac{(N_1 - df_1 + 0.5)*df_2 + 0.5}{(N_2 - df_2 + 0.5)*df_1 + 0.5} \tag{3.11}$$

N_j is the total number of training documents in class c_j. $df_{i,j}$ is the number of training documents in class c_j that contain term i.

3.7.2 Tfc-nfx weighting

The calculation for the tfc-nfx weighting scheme is described below.

$$tfc\_nfx = \frac{tf*log(N/n)}{sum((tf_i*log(N/n_i))^2)} \tag{3.12}$$

N is the total number of training sentences n is the total number of training sentences containing the word to be weighted. tf is the term frequency of each term tf_i is the term frequency of each term withing a class n is the total number of training sentences containing the word to be weighted of a class.

3.7.3 Binary TF-IDF

The binary TF-IDF is calculated as follows:

$$\begin{aligned} &tf \in \{0,1\} \\ &idf = \frac{log((N1+N2)}{(n1+n2)}) \end{aligned} \tag{3.13}$$

TF is 1 if the term exists in the sentence and 0 if it does not. $N1$ is the number of sentences of one class and $N2$ is the sumber of sentences of the other class. $n1$ is the number of sentences of one class containing the term. $n1$ is the number of sentences of the other class class containing the term.

3.8 LibSVM format and parameter choice

The C parameter of the SVM controls the cost of misclassification. It is a cost parameter for the soft margin function and describes the "hardness" or "strictness" of the margin. A high C will put a high penalty on non-separable data points (low bias), which is a risk for overfitting and a low C may result in a low variance and underfitting instead. To adjust

C is a tradeoff between the size of the training error and stability. The gamma parameter of the SVM describes how data points are transformed, by controlling the shape when raising the data points into a higher dimension. A high gamma may result in high bias and low variance, while the opposite might be the case with a low gamma.

Chih-Wei et al. (2008) recommend to randomly test a few kernels and parameters in order to choose the best solution by trial and error. This is the case since the SVM performs differently based on the nature of the input data. The SVM performance is tested with all four kernels: the linear, the rbf, the sigmoid and the polynomial kernel. Furthermore the kernels often contains the parameters gamma and C. The best combination of both parameters is found with the java package 'grid search', where a grid for C between 2^{-5} and 2^{15} and a gamma between 2^{-15} and 2^{3} is recommended (Chih-Wei et al., 2008). An interval of 10 steps is used for the grid search.

The input data for training and testing of the libSVM package (i.e. articles) needs first to be transformed into a libSVM format. The SVM takes as input data an attribute-value representation of text in the following format:
$label \quad index1 : value1 \quad index2 : value2 \quad \ldots$ as can be seen in figure 5.

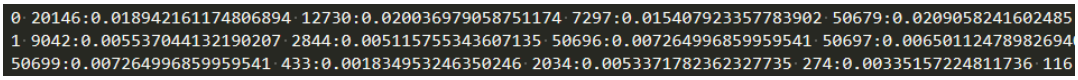

Figure 5: libSVM format

'Label' refers to the class (for example positive or negative), Index is an index to a unique feature (word) *w* in a list of unique words and value is the weighting scheme selected for approximating an optimal degree of influence of each feature.
The value in is calculated by o(BM25)TF and Δ(k) IDF, see chapter 3.7. The feature values are scaled to an interval between 0 and 1 with the libSVM scale method. This is done since the linear and the RBF kernel may have numerical problems with large feature value (Chih-Wei et al., 2008).

Furthermore the libsvm function svm_scale is used which scales the SVM feature values between the interval 0 and 1. Some kernels require scaling since they can not handle high feature values. In addition, when feature values are spread over a great spectrum, scaling prevents attributes in "higher numeric ranges to dominate those in smaller numeric ranges" (Chih-Wei et al., 2008).

3.9 Accuracy, precision and recall

The methods of accuracy, precision and recall calculations are used to evaluate the final results of the created sentiment classification model. The formulas for the calculation of accuracy, precision and recall are given beneath.

Precision in the context of this thesis is the total number of a label type[7], that has been assigned correctly, in relation to the total number of assigned labels of that type. Recall is the number of sentences that have been assigned the correct label type in relation to all sentences that actually possess that label type. Accuracy is the total number of correctly assigned sentences of any label type over the total number of assigned sentences.

$$Precision = \frac{tp}{tp + fp} \tag{3.14}$$

[7] for example the label type "positive"

$$Recall = \frac{tp}{tp+fn} \tag{3.15}$$

$$Accuracy = \frac{tp+tn}{tp+tn+fn+fp} \tag{3.16}$$

In the equations for accuracy, precision and recall, tp means "true positive", tn means "true negative", fp stands for "false positive" and fn stands for "false negative". These abbreviations are used to compare the outcome of the predicted labels by the machine, to the actual labels determined by human annotations, as in table 3.1. In a "false negative" for example, the machine predicted falsely a negative label, while the true outcome is a positive label as determined by the human annotators.

		Machine (Predicted condition)	
	Outcome	Positive	Negative
Human annotators (True condition)	Positive	tp	fn
	Negative	fp	tn

Table 3.1: *Predicted condition and true condition*

3.10 Evaluation

Evaluation was done for four aspects: evaluation of parameter combination for best accuracy, evaluation of machine classification results by comparison to human classification, evaluation of results by comparison to similar research, evaluation with regards to the requirements, defined in section 2.3.2.

3.10.1 Parameter combination for best accuracy

In total all possible combination of two kernels, three weighting methods, scaling, no scaling and handlers versus no handlers have been combined in program runs for a grid of C and gamma parameters. The grid of C and gamma parameters combined 23 different C's and 10 different gammas to include the whole recommended C and gamma range, but also to narrow in on the C's and gammas that resulted in best accuracy.

For variations of C and gamma, first the libsvm grid search function was used with 10 steps and the intervals recommended by (Chih-Wei et al., 2008). No more than 10 steps could be used because heap space memory problems would occur otherwise. In some cases, grid search did not result in the best parameter combination. In those cases, C's and gammas where tuned manually, starting with variation of the C's and gammas recommended by the grid search function. Where best accuracies were found for a particular C and gamma combination, both parameters were varied in proximity of this C and gamma combination in the hope for another local accuracy maximum nearby.

3.10.2 Inter annotator agreement (IAA)

The volunteers classified the article sentences on the scale negative, neutral and positive. They did this for a sample of the training data, but also for a separate set called the test

set. Neutral was included in the human classification scale, since the machine filtering of neutral sentences in the training data creation phase could still include sentences in the set, which humans would judge to be neutral.

The calculation of the inter-annotator agreement value, called Fleiss' kappa measures by how much the human annotators agree on their annotations of sentence sentiment. It is used to evaluate the degree of difficulty of sentence classification by humans, so that conclusions can be drawn for the difficulty of machine classifications. The kappa inter annotator agreement calculations are done to evaluate the human classifications of the test and training set against machine classification. The IAA is also calculated for the human evaluation of the test set in comparison to the machine classification. The calculations of the IAA are explained here with the example of the test set for the sentiment model.

The test set contains 399 sentences which are classified by three Swedish volunteers. Fleiss' kappa is a method to measure inter-annotator reliability or agreement which extends Cohen's Kappa in order to be applied for more than two annotators (Schaer, 2012). The kappa value measures by how much the annotators ratings exceed a chance rating. Landis and Koch Landis and Koch (1977) recommend to interpret the kappa values according to an interval from 0 to 1, where

- $k \leq 0$ is poor agreement,
- $0 \leq k \leq 0.2$ is slight agreement,
- $0.2 \leq k \leq 0.4$ is fair agreement,
- $0.4 \leq k \leq 0.6$ is moderate agreement
- $0.6 \leq k \leq 0.8$ is substantial agreement
- $0.8 \leq k \leq 1$ is almost perfect agreement

The kappa k is defined as follows:

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{3.17}$$

The denominator describes the percentage of agreement that is possible to obtain above chance, while the nominator describes the actual percentage of agreement above chance.

In this thesis, the total number of sentences rated in the test set is $N = 399$, the number of ratings per sentence is $n = 3$ and the number of categories of ratings is $m = 3$ (positive, neutral and negative). The sentences are indexed from $i = 1, ..., N$ and the categories are indexed from $j = 1, ..., m$.

The number of ratings per sentence and category can be visualized in a $399x3$ matrice as in table 3.2.

p_j is the sum of all ratings for one category (for example for the category 'negative'), divided by the number of total ratings for all categories, which is $3x399$ in the matrice above.
$\bar{P}_e$ is the sum of all p_j^2:

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2 \tag{3.18}$$

$n_{i,j}$	positive	neutral	negative	P_i
Sentence 1	0	0	3	1
Sentence 2	2	1	0	0.33
...				
...				
...				
Sentence 399				
p_j				

Table 3.2: *Fleiss' kappa matrice*

P_i represents the inter-rater agreement for the i-th sentence:

$$P_i = \frac{1}{n(n-1)}\left((\sum_{j=1}^{k} n_{ij}^2) - n\right) \tag{3.19}$$

$\bar{P}$ is the mean of P_i's.

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} P_i \tag{3.20}$$

If the annotators disagreed on labelling for a sentence, for example two annotators label a sentence as positive and one annotator as neutral, then the majority of "votes" becomes the sentence label, i.e. positive in the example. If all three annotators disagree, the label for a sentence is neutral. The IAA was calculated for the human evaluation of the training data set and also for human evaluation of the test set.

3.10.3 Validation by comparison of machine and human classification results

Accuracy, precision, recall and F1-score are calculated for both the evaluation of the machine classification performance on the test set and also for evaluation of the training set.

While the validation set is used to adjust parameters and improve the results of the SVM, the test set serves the only purpose to evaluate the final algorithm performance on unseen data that has been validated by humans. During the comparison of the machine classification of the test set and the human classification of the test set, sentences that were classified as neutral by humans were included. Since machine classification only takes account of the classes positive versus negative, neutral sentences are by default classified as false negatives or false positives.

A preliminary evaluation of the training set accuracy through a comparison to human labelling, revealed that the Stanford and Naive Bayes classifier, which automatically labelled the training data for the libSVM, included faultily many neutral sentences. Since the set of $57,936$ sentences was too large in order to sort out neutral sentences manually, it was instead decided that neutral sentences would be sorted out of the final test set by the author to evaluate the libSVM model only on what it was supposed to accomplish: rate positive and negative sentences. The test set was then given to three independent Swedish annotators.

3.10.4 Comparison to similar research

The evaluation of the thesis results include placing the results in context with similar research results in order to show that the they do not deviate abnormally from what can be expected. This is done by comparison to the research presented in the introduction chapter.

3.10.5 Comparison to requirements

Finally the evaluation of the artifact includes a comparison to the requirements presented in the method chapter in order to show the artifacts practical use for further research regarding the accuracy obtained, the use of handlers and regarding time and memory for program execution.

4. Results

The result chapter first demonstrates the results of parameters which deliver the best accuracy for the SVM model. Additionally, the calculations for the inter annontator agreement of the test set and the training data set are presented. The results for the evaluation of the SVM model and the training data by comparison of machine versus human classifications are shown. Lastly, comparisons of the test set results to similar research, and a comparison of results to the artifact requirements are brought forward.

4.1 Parameter combination for best accuracy of the SVM model

During SVM trials with parameter variation, it became clear that less data would give lower accuracy, no matter which other parameters were varied. Therefore, in all later trials, all training and validation data was used, which is a total of $57,536$ sentences. Furthermore, it became apparent that higher accuracies for the RBF kernel were achieved without scaling, while results varied for the Linear kernel, see Appendix C.

In the SVM runs, all possible combination of different parameters were executed. This resulted in a total of 24 runs with grid of C and gamma combinations that comprised 23 different C's and 10 different gammas.

The following parameters were varied and combined:

- kernel type: RBF, Linear
- feature weighting: Okapi, TFIDF, Binary
- C
- gamma
- scaling vs no scaling
- handlers vs no handlers

The results show that the best training accuracy of 93,82% was achieved with the RBF kernel, without scaling and with Binary weighting (see figure 6).

For the RBF kernel, the parameters C and gamma had to be varied to find the best accuracy, while only C had to be varied with the Linear kernel.

The following image shows an example of how the SVM model accuracy varies with different C and gamma:

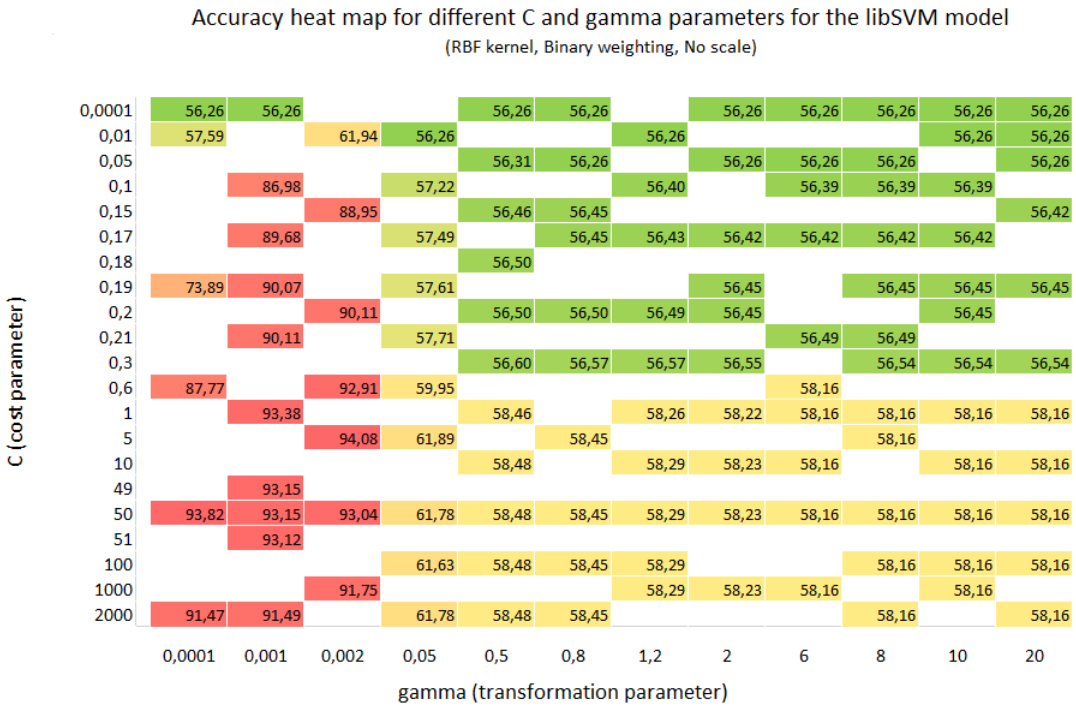

Figure 6: Heat map of SVM model accuracy with different C and gamma parameters

4.1.1 Feature filtering

Trials with and without feature filtering, i.e. stop word filtering, frequency filtering, name filtering, tokenization and lemmatization show that the accuracy seems to be better for most parameter variations, when feature filtering is used (even if the increase in accuracy only influences the first decimal after the comma for certain C and gamma parameter combinations). For the best C and gamma parameters found for example, no feature filtering resulted in 65 303 unique features and an accuracy of 93.49 %, but with feature filtering the features were reduced to 18 737 unique features and the accuracy was slightly better at 93.82 %. A small table with the number of unique features after broad filtering is shown in Appendix G.

4.1.2 Inclusion of language specific handlers

The best accuracy achieved remained the same whether language specific handlers were used or not, see figure 8.

When using the linear kernel, the use of handlers with scaling results in 0.1% higher accuracy for TFIDF and in 0.08% higher accuracy for Binary weighting, but in 1% lower accuracy for Okapi weighting. The use of handlers without scaling for the linear kernel results in up to 0.3% higher accuracy without language specific handlers (figure 8).

For the RBF kernel with scaling, the accuracy improved by 8% without handlers for the Binary weighting scheme in comparison to using handlers, and by up to 1% for TFIDF and Okapi weighting. The accuracy is 0.4% higher with handlers for TFIDF, but it improves by 0.2% without handlers for Okapi and stay the same for binary weighting, independently of scaling or handlers (figure 8).

In summary, the use of handlers mostly decreases accuracy and in the best case has minimal or no impact on accuracy independently of kernel, weighting or scaling.

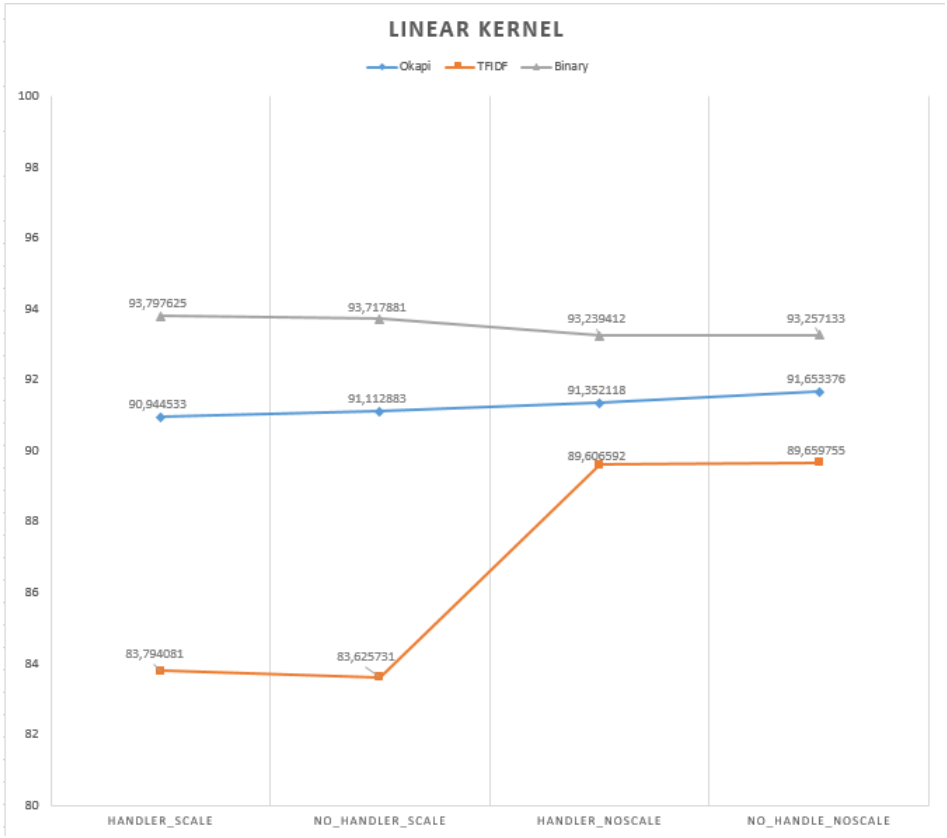

Figure 7: Comparison of accuracy with and without handlers - Linear kernel

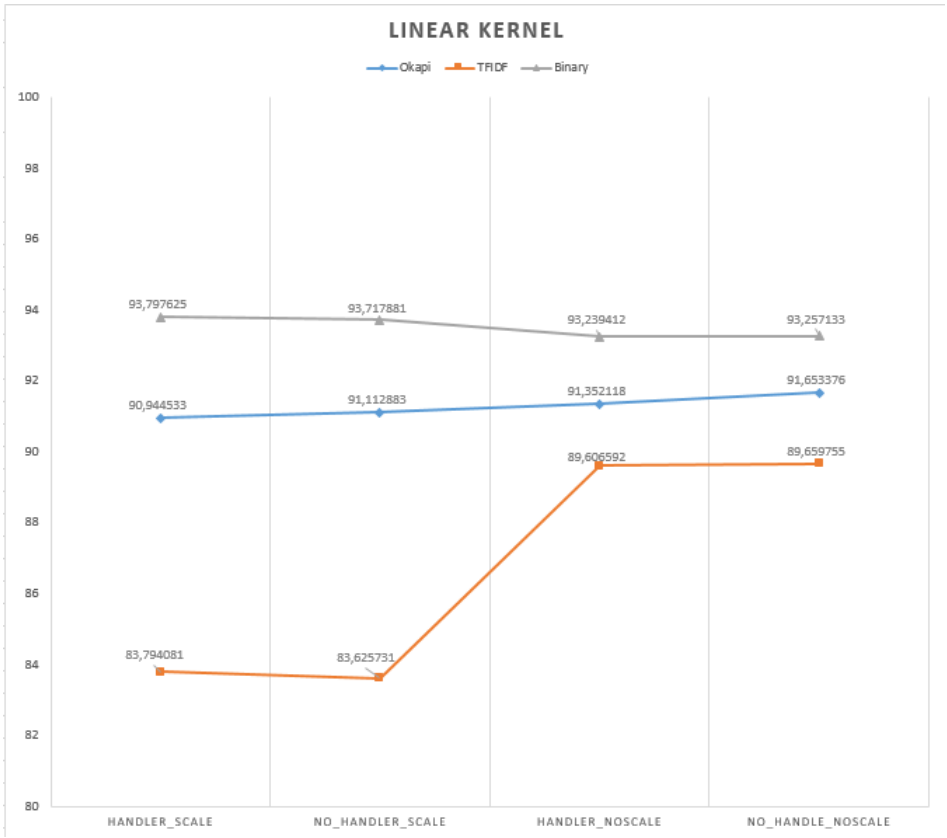

Figure 8: Comparison of accuracy with and without handlers - RBF kernel

Since the use of language specific handlers did not seem to yield any satisfactory results, further tests were run to understand why.

The first test attempted to see how many sentences in the training set were affected by each handler separately. Results show that about 17% of all sentences were affected by the compound words handler, 7% were affected by the negation handler and 5% by the idiomatic expression handler.

In the second test, all handlers were run together on the training set: circa 17% of all sentences were affected by compound words, 7% by negation handlers and 2% by idiomatic expression handlers. The difference in percentage between running each handler itself and running all handlers together, suggests that between 17% and up to 26 % of all sentences are affected by the handlers.

The results show furthermore that the handlers only minimally influence each other.

For a third test, all sentences that are affected by handlers were gathered in a handler set. The SVM model was trained with all three handlers, and the prediction was done on the handler set. Although the SVM model was trained with handlers, prediction accuracies on the handler set were approximately 1% better without handlers (96,48%) than with handlers (95,82%). Also predictions with different combinations of handlers did not result in better accuracy than without handlers.

Finally, the author manually inspected the handler set after preprocessing with the handlers and found only few errors which do not explain the ineffectiveness of handlers.

The handlers do only minimally interfere with each other and neither interference, or the number of sentences affected by handlers, or handler errors could be reasons for the insignificant influence on the accuracy of the SVM model.

4.2 Inter annotator agreement for test data

The test set of 399 sentences was displayed sentence by sentence on a website on which 3 Swedish volunteers validated each sentence on a nominal scale with the values positive, neutral and negative. In total the whole test set has been validated three times. The volunteers classified the sentences on the scale negative, neutral and positive. Neutral was included in the human validation scale, since the machine filtering of neutral sentences in the training data creation phase could still include sentences in the set, which humans would judge to be neutral. The inter-annotator agreement was measured with Fleiss' kappa, see section 3.10.

The kappa value for the sentiment model is:

$$
\begin{aligned}
k &= \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \\
&= \frac{0.823726 - 0.439945}{1 - 0.439945} \\
&\simeq 0.69
\end{aligned}
\tag{4.1}
$$

According to Landis and Koch (1977), this kappa value can be interpreted as 'substantial agreement'.

4.3 Inter annotator agreement for training data

A sample of the training data, 400 sentences, has been validated by three humans the same way as described in section 4.2. The kappa value for the annotations of the training data is 64.15 which corresponds to the Landis and Koch (1977) label 'substantial agreement'.

4.4 Evaluation of the sentiment model

The parameters which gave the best accuracy were used for machine classification of the test set, i.e. RBF kernel, without scaling, Binary weighting and no handlers. A comparison of the machine classification of the test set and the human classification of the test set resulted in an accuracy of 83.71% . Precision for the test set is 89.5% and recall 82.42%, which corresponds to an F1-score of 85.8%, (see Appendix D).

4.5 Evaluation of training data

The Stanford classifier and the Swedish classifier attibuted the same label to 57 935 sentences out of a total of 261 416 unfiltered sentences, i.e. the classifiers agreed on 22 % of all sentences. The evaluation of the training data was calculated the same way as described in 3.10.3. Since the humans validated almost half of all sentences as neutral and the class neutral is not included in the machine classification, neutral sentences are by default counted as false positives or false negatives in the comparison. The accuracy of the human-machine comparison of classifications of training data was accordingly low at about 60%. When neutral sentences were excluded from the comparison, accuracy was 78.8%. For all values, including precision, recall and F1-score, see Appendix E.

4.6 Comparison to similar research

Narrow domain sentiment analysis studies, such as movie and restaurant review sentiment classification by SVM, as described in the introduction chapter, resulted in accuracies between 82.9% to up to 95.55%. The results of this study with an accuracy of 83.71% thus fall into the lower range of the interval, but are similar to previous research. Research within sentiment analysis on general/broad domains report between 65% and 85% accuracy (see introduction chapter, 1.1.1). Comparisons of the results of this study to research on sentiment analysis with SVM on those domains, place the thesis results on the upper range of the interval.

4.7 Comparison to artifact requirements

The requirements of a machine learning accuracy between 65% and about up to 85% for the evaluation of both test and training data set, and the final validation set, were measured in regard to the human gold standard and are fulfilled.

Furthermore, a hypothesis was that the implementation of language specific handlers increases the overall measured accuracy of the Swedish sentiment classification model. The requirement that would prove this hypothesis to be true - given all parameters are hold constant, the total accuracy should increase with the use of language specific handlers - has not been met.

Lastly, requirements for an acceptable Swedish classification prototype include that the classification process can be finished within a reasonable timeframe, i.e. not more than 12 hours, and within reasonable memory use, i.e. not more than 8 GB. Measurements show

that training runs with most parameters (RBF, scaling, handlers) take up to 40 minutes until a result is produced and when the libsvm grid search is used with no more than 10 steps, the memory requirements are met. Classification took not more than a maximum of 30 minutes per run.

5. Discussion and Conclusions

In the discussion and conclusion section, the answer to the research question and sub-questions are provided. A discussion of the model design and the results of the parameter tuning, the feature filtering step and the language specific handlers is held. Continued future studies are proposed and the ethical and societal impact of the research is contemplated. Finally, examples of the practical use of sentiment analysis for newspaper articles are given.

5.1 Answer to the research question

The research goal in this thesis was to design, instantiate and evaluate a Swedish sentiment classification model, which has been done.

The evaluation of the model provides an answer to the main research question: "What is the performance of the designed Swedish sentiment classification model, measured with accuracy, precision and recall?"

The best performance of the designed Swedish sentiment classification model was measured to be an accuracy of 83.71%, a precision of 89.5% and recall of 82.42%, which corresponds to an F1-score of 85.8% (see Appendix D).

Furthermore, to gain an understanding of the factors that influence the performance of the Swedish sentiment classification model, two sub questions were asked:
1) which variation of parameters results in best accuracy and
2) does the inclusion of language specific handlers improve the classifier?

The RBF kernel, without scaling, and Binary weighting resulted in best accuracy of the model. The inclusion of language specific handlers did not improve the classifier. Reasons for the answers obtained in this research will be discussed.

5.2 The model design

The five biggest newspapers to which the author had access to were all liberal newspapers which certainly has an influence on how negative or positive certain topics are mentioned in the articles. Furthermore topic filtering employed more negative than positive keywords which has a direct influence on articles that were added to the training set. However both biases mentioned above are not of concern since this research is not concerned with the political stance or the number of negative and positive articles , but only with a correct classification of negative and positive sentences. Positive and negative sentences can be found in any newspaper and any article about the topic 'school'.

The results of the libSVM training and testing accuracy would suggest that the design for creating sentiment training data through two very different algorithms[1] has not been too successful, since many neutral sentences were included despite efforts to sort them out. However, an evaluation of the training data without neutral sentences showed that the accuracy for human-machine comparison for labels on positive and neutral sentences was high at 78% with a high precision of 85% and a recall of 72% (see Appendix E). In conclusion, the training data included many neutral labels, but had overall the right labelling for positive and negative sentences. This meant that the libSVM model could be

[1] the Stanford RNN through Google Translate, and through a Swedish sentiment classifier based on Naive Bayes

trained with the training data at hand for the task to label positive and negative sentences. A final validation of the libSVM model on its performance to classify positive and negative sentences[2] showed that the approach of training the model was successful with a 83.71% accuracy.

It should be noted that the IAA for the training data was a kappa of 64, which means substantial agreement, but is rather on the lower end of this agreement interval. This indicates that the general task of sentiment categorization was difficult for humans. Therefore it can be expected that classification errors were introduced even in the human labelling process.

Of course there are inherent uncertainties in this new design for the creation of Swedish labelled training data. Both, the Stanford RNN classifier with Google Translate and the Swedish classifier based on Naive Bayes, are not perfect and even when selecting sentences with the same labels, errors are introduced to the labelled training data. Nonetheless, it could be shown that precision and recall was high for the labels positive and negative, which were the focus of this research.

The efficiency of the classification model is good enough for real life applications, since 40 minutes for training and a maximum of 30 minutes for classification of approximately 57000 sentences is usable.

5.3 Parameter tuning and feature filtering impact on model accuracy

The parameter tuning showed that the C and gamma parameters which result in optimal accuracy were a C of 50 and a gamma of 0.0001 for the RBF kernel with Binary weighting and no scaling. The C parameter is not large on the suggested scale for C which can reach 2^{15}. Therefore the margin allows for underfitting and training error, but ensures more stability. The gamma parameter is low as well, which means relatively high bias, but low variance. The parameter combination of C and gamma for the SVM allows therefore for training errors; on the other hand the SVM was able to classify all sentences and the validation shows that the SVM does so quite accurately nonetheless. The fact that highest accuracy was achieved with the RBF kernel rather than the Linear kernel might indicate that the features space might not have been entirely linearly separable.

Besides parameter tuning, even feature filtering influenced the accuracy of the SVM. Trials showed that accuracy improved when feature filtering was used. This is likely because this step eliminates unnecessary features, for example by filtering out words that only occur once in the data set. Many of those words turned out to be spelling mistakes or names which do not contribute to the overall sentiment.

5.4 Language specific handler impact on model accuracy

Unfortunately the language specific handlers did not have any significant impact on the accuracy results of the SVM model. Tests were done to determine the reason why, and several factors could be eliminated.

Firstly it was assumed that not enough sentences were affected by the handlers, but tests showed that this is not the case, since between 17% and 26% of all sentences are affected by handler modifications. A second assumption was that the handlers affect each other

[2] via a presorted test set, see explanation in section 3.10.3

and eliminate the intended results, but tests showed again that this is not the case. Thirdly, it was assumed that the handler implementation was faulty, but a routine check showed that although some errors were detected[3], nothing would indicate that the implementation would account for a lack of increase in accuracy.

After elimination of the above assumptions, another step was to gather all sentences that are affected by handlers in a separate handler set and run the libSVM prediction on the handler set, once with and once without handlers. A comparison of prediction accuracies with and without handlers on the handler set shows that the total accuracy does not improve with any combination of handlers. Prediction accuracies of the handlers set were always 1% better without any handlers.

All handler implementations aggregate information which semantically belongs together, but is separated within a sentence. An example of such information is the sentence "Det är inte så pass tydligt" (translates to: this is not really so clear), where the "inte" ("not") and the "tydligt" (clear) are separated with two other words in between, but belong logically together. Translated into libSVM format, each handler transformation will result in one new feature-value pair which logically combines two words. If handlers were not used, logically connected words would be transformed separately into feature-value pairs.

One reason for handlers not improving the accuracy of the libSVM predictions, could be because each logically combined features-value pair is a relatively rare presence in the whole set of features, which means that the features would get negligible weight. On the one hand, those combined features-value pairs may enhance the total feature set with rare features, which might have detrimental effects on accuracy. On the other hand, previous studies have shown that the SVM can handle a great number of features in text categorization (Joachims, 1998). In fact, the SVM might need all information that could be lost by aggregating several words within one sentence. However, it can not be concluded that enhancing the number of features by providing more information, as is the case with handlers, results in better accuracy.

5.5 Thesis results in relation to previous findings

The results of this study seem to be in line with the conclusions of Andreevskaia and Bergler (2008) and Aue and Gamon (2005) who found that sentiment analysis on narrow domains is often more precise than on broad domains. Vinodhini and Chandrasekaran (2012) for example report SVM results from 86.4% to up to 93% for movie and restaurant reviews(specific domain), whereas Lambov et al. (2011) did a study for domain independent sentiment analysis and achieved performances between 60% and up to 80% with different SVM classifiers. The performance of the model created in this thesis falls on the higher interval of general domain sentiment analysis and on the lower interval of specific domain sentiment analysis. This seems to be a logical outcome, since newspaper articles are a broad domain, but this domain has been narrowed by filtering articles according to the topic 'school'. However the topic 'school' is per se a broad domain, including articles from school building renovations to school theatre performances.

The unsignificant impact of the created language specific handlers on the sentiment model performance agree with the findings of (Kim and Hovy, 2006; Mihalcea et al., 2007) which seem to suggest that language specific traits do not play an important role in sentiment analysis. However, the results in this study may only give a further indice

[3] due to word occurrence combinations in sentences that the implementation does not catch

towards such a conclusion, and much more variations for language specific handlers with variations of different classifiers would have to be tested for an affirmation.

5.6 Reproducibility, validity, reliability, generalisability and further applicability

This results of this study can be reproduced, using the same data, code and parameters as in this study. The code will be made freely available upon request. The validity and reliability of this study is ensured through external validation of the thesis results, and through the structured use of extensive design and evaluation frameworks, which have been followed step for step. The application of the frameworks for this study has been described in detail. All necessary information, data analysis methods and tools for building the artifact has been explained and documented. The results of this study allow to a certain degree the estimation of results for similar studies and can certainly be used to compare to similar research on broad domains and on the domain of newspaper sentiment classification. Of course, it is in inherent in the logic of different algorithms that results for one classifier can not be generalised to other classifiers. The results of this study show that the sentiment model may be applied to real world problems or for further studies in the same domain of text mining.

5.7 Conclusion

This thesis contributed to the text mining community in the area of sentiment analysis. The research goal has been reached and the research question and subquestions have been answered. Methodological frameworks have been followed and the validity and reliability of this work has been ensured. Labelled training data has been created. A Swedish sentiment analysis model has been designed, instantiated and evaluated with good results.

5.8 Future studies

Future studies could explicitly concentrate on researching the effects of language specific feature reduction and feature additions on SVM prediction results. It would be interesting to see which, if any, language specific information should be provided to improve accuracy for SVM's. This reasoning can be generalized to any classifier, since there is a lack of knowledge regarding the balance between language specific feature selection and accuracy results for any classifier. Relevant research on dimension reduction in text classification for SVM has been done by Kim et al. (2005). Future studies could continue Kim et al. (2005)'s research, but with a tighter focus on language adapted feature reduction of enhancements, since there is still a discrepancy between the fact that each language possesses language specific features and the fact that no conclusive evidence has been given that those features could be used to an advantage in machine learning.

In addition it could be studied if language specific handlers have a better effect on the accuracy of other classification algorithms, such as the Naive Bayes. It could be that the inherent logic of the SVM requires as much information as possible which makes gathering of information unnecessary, while such aggreation will benefit the accuracy of other classifiers.

Furthermore, sentiment analysis in Swedish would profit from research about new and improved automatic methods to generate labelled sentiment training data in other languages than English. The problem that there is almost no labelled training data in Swedish

applies certainly also to other minor languages. In order to advance the field of sentiment analysis even in those countries necessitates correct training data in the respective language. Since manual labelling is time consuming, further studies regarding the creation of training data should concentrate on supervised or semi-supervised methods. In more detail, such future studies could try to eliminate error sources when translating labelled training data from English to other languages or to find accurate methods that can directly be applied on native texts.

5.9 Ethical and societal consequences

All ethical guidelines have been observed in this research. Newspaper articles are public information that have to be archived and accessible over years and therefore no sensitive material, but only publicly available information has been handled. The thesis results are sentence labels without connection to organisations or people and can only be used to help understand and visualize the sentiment in public data. In a stable democracy like Sweden, where this study has been written, this is very unlikely to entail any societal consequences. The only human participants were three annotators for which all ethical guidelines have been observed. Therefore this thesis does not entail any negative ethical consequences either.

5.10 Societal and academic contributions

Sentiment analysis of newspaper articles, sorted after certain topics, may help visualize important public debates and thus contribute to the societal democratic process. An example of the practical use of sentiment analysis on newspaper articles can be found in Appendix F.

This research contributed to the domain of text classification with a concept for the creation of labelled Swedish sentiment training data, with the design concept and instantiation of a Swedish sentiment classification model with good accuracy results, and lastly with further insights into the impact of language specific handlers on SVM.

Bibliography

Abdul-Mageed, M., Diab, M. T. and Korayem, M. (2011), 'Subjectivity and sentiment analysis of modern standard Arabi0063', *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 2(June), 587–591.
URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84859045256&partnerID=tZOtx3y1

Alturki, A., Gable, G. G. and Bandara, W. (2013), 'The design science research roadmap: In progress evaluation', *Proceedings of the 17th Pacific Asia Conference on Information Systems* p. 160.

Andreevskaia, A. and Bergler, S. (2008), 'When specialists and generalists work together: overcoming domain dependence in sentiment tagging', *Proceedings of ACL-08: HLT* (June), 290–298.

Aue, A. and Gamon, M. (2005), 'Customizing Sentiment Classifiers to New Domains : a Case Study', *Proceedings of Recent Advances in Natural Language Processing RANLP* 49(2), 207–18.

Balahur, A. and Turchi, M. (2012), 'Comparative Experiments for Multilingual Sentiment Analysis Using Machine Translation', *1st International Workshop on Sentiment Discovery from Affective Data (SDAD 2012)* pp. 75–86.

Baskerville, R. (2008), 'What design science is not', *European Journal of Information Systems* 17, 441–443.

Bollen, J., Mao, H. and Zeng, X.-J. (2011), 'Twitter mood predicts the stock market', *Journal of Computational Science* 2.1, 1–8.
URL: http://arxiv.org/abs/1010.3003http://dx.doi.org/10.1016/j.jocs.2010.12.007

Cao, M. D. and Zukerman, I. (2012), '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.
URL: http://www.aclweb.org/anthology/U/U12/

Chang, C.-C. and Lin, C.-J. (2015), 'LIBSVM - A Library for Support Vector Machines'.
URL: https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Chih-Wei, H., Chih-Chung, C. and Chih-Jen, L. (2008), 'A Practical Guide to Support Vector Classification', *BJU international* 101(1), 1396–400.
URL: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Cleven, A., Gubler, P. and Hüner, K. M. (2009), 'Design alternatives for the evaluation of design science research artifacts', *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09* p. 19.
URL: http://portal.acm.org/citation.cfm?doid=1555619.1555645

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A. and Menczer, F. (2011), '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.

Crotty, M. (1998), *The foundations of social research: Meaning and perspective in the research process*, SAGE Publications Limited.

Denecke, K. (2008), 'Using SentiWordNet for multilingual sentiment analysis', *Proceedings - International Conference on Data Engineering* pp. 507–512.

Denscombe, M. (2003), *The good research guide for small-scale social research projects*, 2nd edn, Open University Press, Maidenhead: Philadelphia.

Doyle, D. (2016), 'Swedish Stopwords'.
URL: http://www.ranks.nl/stopwords/swedish

Eickhoff, M. (2015), Enabling reproducible sentiment analysis: A hybrid domain-portable framework for sentiment classification, *in* 'New Horizons in Design Science: Broadening the Research Agenda - 10th International Conference, DESRIST 2015, Dublin, Ireland, May 20-22, 2015, Proceedings', pp. 215–229.

Erik, C., Schuller, B., Xia, Y. and Havasi, C. (2013), 'New Avenues in Opinion Mining and Sentiment Analysis', *IEEE Intelligent Systems* 28(2), 15–21.

Esuli, A., Sebastiani, F. and Moruzzi, V. G. (2006), 'S ENTI W ORD N ET : A Publicly Available Lexical Resource for Opinion Mining', *Proceedings of LREC 2006* 0, 417–422.

Farooq, S. and O'Brien, C. (2015), 'An action research methodology for manufacturing technology selection: a supply chain perspective', *Production Planning & Control* 0(0), 1–22.
URL: http://dx.doi.org/10.1080/09537287.2014.924599

Fletcher, T. (2009), 'Support Vector Machines Explained', *Unpublished* pp. 1–19.
URL: http://sutikno.blog.undip.ac.id/files/2011/11/SVM-Explained.pdf

Ghahramani, Z. (2004), 'Unsupervised Learning BT - Advanced Lectures on Machine Learning', *Advanced Lectures on Machine Learning* 3176(Chapter 5), 72–112.

Goldkuhl, G. (2013), 'Action Research Vs . Design Research : Using Practice Research As a Lens for Comparison and Integration', *Workshop on IT Artefact Design & Workpractice Improvement* pp. 1–23.

Hatzivassiloglou, V., Hatzivassiloglou, V., McKeown, K. and McKeown, K. (1997), '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* p. 181.
URL: http://portal.acm.org/citation.cfm?id=976909.979640

He, B. and Ounis, I. (2005), Term frequency normalisation tuning for BM25 and DFR models, *in* 'Proceedings of the 27th European Conference on Information Retrieval', number 1, Santiago de Compostela, pp. 200–214.
URL: http://dx.doi.org/10.1007/b107096

Hedlund, T., Pirkola, A. and Järvelin, K. (2001), 'Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language information retrieval', *Information Processing and Management* 37, 147–161.

Hogenboom, A., Heerschop, B., Frasincar, F., Kaymak, U. and De Jong, F. (2014), 'Multi-lingual support for lexicon-based sentiment analysis guided by semantics', *Decision Support Systems* 62, 43–53.

Howell, K. E. (2013), 'Book launch - 'The Philosophy of Methodology''.
URL: http://www1.plymouth.ac.uk/research/slgp/SLGPevents/Pages/Book-launch-Professor-Kerry-E.-Howell.aspx

Hult, M. and Lennung, S. (1978), 'Towards a definition of action research: a note and bibliography', *Journal of Management Studies* 17(2), 241–250.

Iivari, J. and Venable, J. (2009), 'Action Research and Design Science Research - Seemingly similar but decisively sissimilar', *17th European Conference on Information Systems* pp. 1–13.

Jang, H. and Shin, H. (2010), 'Language-specific sentiment analysis in morphologically rich languages', *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (August), 498–506.
URL: http://portal.acm.org/citation.cfm?id=1944623

Järrvinen, P. (2007), 'Action research is similar to design science', *Quality and Quantity* 41(1), 37–54.

JetBrains (2015), 'IntelliJ IDEA'.
URL: https://www.jetbrains.com/idea/

Joachims, T. (1998), 'Text Categorization with Support Vector Machines: Learning with Many Relevant Features', *Machine Learning* 1398(LS-8 Report 23), 137–142.
URL: http://www.springerlink.com/index/drhq581108850171.pdf

Johan, H., Nilsson, J. and Nivre, J. (2012), 'Swedish MaltParser model'.
URL: http://www.maltparser.org/mco/swedish{_}parser/swemalt.html

Johannesson, P. and Perjons, E. (2012), *A Design Science Primer*.
URL: https://docs.google.com/file/d/0B95tYlhL8SrlMHE5czkyUF9BVG8/edit

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

Kaufmann, J. M. (2012), 'JMaxAlign : A Maximum Entropy Parallel Sentence Alignment Tool', *Coling-2012* (December 2012), 277–288.

Keyvanpour, M. and Tavoli, R. (2012), 'Feature Weighting for Improving Document Image Retrieval System Performance', *IJCSI International Journal of Computer Science Issues* 9(3), 125–130.

Kim, H., Howland, P. and Park, H. (2005), 'Dimension Reduction in Text Classification with Support Vector Machines', *Journal ofMachine Learning Research* 6, 37–53.

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

Kim, S.-M. and Hovy, E. (2006), 'Identifying and analyzing judgment opinions', *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (2003), 200–207.
URL: http://portal.acm.org/citation.cfm?doid=1220835.1220861

Lambov, D., Pais, S. and Dias, G. (2011), 'Merged agreement algorithms for domain independent sentiment analysis', *Procedia - Social and Behavioral Sciences* 27(Pacling), 248–257.
URL: http://linkinghub.elsevier.com/retrieve/pii/S1877042811024323

Landis, R. and Koch, G. (1977), 'The measurement of observer agreement for categorical data', *Biometrics* 33(1), 159–174.

Lau, R., Liao, S. and Xu, K. (2010), 'An Empirical Study of Online Consumer Review Spam: A Design Science Approach', *ICIS* 2010, 103–123.

Li, D. and Dong, Y. (2014), 'Deep learning: Methods and applications', *Foundations and Trends in Signal Processing* 7(3-4), 197–387.

Lista över svenska idiomatiska uttryck (2015).
URL: https://sv.wikipedia.org/wiki/Lista_%C3%B6ver_svenska_idiomatiska_uttryck

Liu, B. (2012), 'Sentiment analysis and opinion mining', *Synthesis lectures on human language technologies* (May), 1–108.
URL: http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10570911

Ludovici, M. and Bignon, G. (2015), *Sentiment analysis of Swedish social media*, Stockholm University, Student report.

Mattisson, B., Borin, L., Berglund, C. and Berthelsen, H. (2000), 'Automatisk detektering av partikelverb', pp. 1–62.

Medhat, W., Hassan, A. and Korashy, H. (2014), 'Sentiment analysis algorithms and applications: A survey', *Ain Shams Engineering Journal* 5(4), 1093–1113.
URL: http://dx.doi.org/10.1016/j.asej.2014.04.011

Mihalcea, R., Wiebe, J. and Banea, C. (2007), 'Learning multilingual subjective language via cross-lingual projections', *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (June), 976–983.
URL: http://acl.ldc.upenn.edu/p/p07/p07-1132.pdf

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

OECD (2015), Improving Schools in Sweden: An OECD Perspective, Technical report.

Okasha, S. (2002), *Philosophy of Science: A Very Short Introduction*, Press, Oxford University.

Paltoglou, G. and Thelwall, M. (2010), A study of Information Retrieval weighting schemes for sentiment analysis, *in* 'Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics', number July, pp. 1386–1395.
URL: http://www.aclweb.org/anthology/P10-1141

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q. and Chen, Z. (2010), 'Cross-domain sentiment classification via spectral feature alignment', *Proceedings of the 19th international conference on World wide web - WWW '10* p. 751.
URL: http://portal.acm.org/citation.cfm?doid=1772690.1772767

Pang, B. and Lee, L. (2002), 'A Sentimental Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts', *Framework* cs.CL, 271–278.
URL: http://arxiv.org/abs/cs/0409058v1

Pang, B. and Lee, L. (2008), 'Opinion Mining and Sentiment Analysis', *Foundations and Trends in Information Retrieval* 2(1-2), 1–135.
URL: http://www.nowpublishers.com/product.aspx?product=INR{&}doi=1500000001

Park, E., Storey, V. and Givens, S. (2013), 'An ontology artifact for information systems sentiment analysis'.

Patil, M. G., Galande, M. V., Kekan, V. and Dange, M. K. (2014), 'Sentiment Analysis Using Support Vector Machine', *International Journal of Innovative Research in Computer and Communication Engineering* 2(1), 2607–2612.

Potts, C. (2011), 'Sentiment Symposium Tutorial: Lexicons'.
URL: http://sentiment.christopherpotts.net/lexicons.html

Powers, D. M. W. (1998), Applications and Explanations of Zipf's Law, *in* 'NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning, ACL', pp. 151–160.

Psomakelis, E., Tserpes, K., Anagnostopoulos, D. and Varvarigou, T. (2014), 'Comparing methods for Twitter Sentiment Analysis', *6th Conference on Knowledge Discovery and Information Retrieval 2014* .

Quoc V, L. and Mikolov, T. (2014), 'Distributed representations of sentences and documents', *arXiv preprint arXiv:1405.4053* .

Read, J. and Carroll, J. (2009), '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.
URL: http://dx.doi.org/10.1145/1651461.1651470

Rice, D. R. and Zorn, C. (2013), 'Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies', *Proceedings of NDATAD* .

Robertson, S. (2004), 'Understanding inverse document frequency: on theoretical arguments for IDF', *Journal of Documentation* 60(5), 503–520.

Rothfels, J. and Tibshirani, J. (2010), Unsupervised sentiment classification of English movie reviews using automatic selection of positive and negative sentiment items, Technical report, Stanford University.

Sahlgren, M. and Cöster, R. (2004), 'Using bag-of-concepts to improve the performance of support vector machines in text categorization', *Proceedings of the 20th international conference on Computational Linguistics* .
URL: http://dl.acm.org/citation.cfm?id=1220355.1220425

Salton, G. and Buckley, C. (1988), *Term-weighting approaches in automatic text retrieval*, Vol. 24.

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

Schaer, P. (2012), 'Better than their reputation? On the reliability of relevance assessments with students', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7488 LNCS(1), 124–135.

Sjöholm, K. (1993), Patterns fo transferability among fixed expressions in L2 acquisition, *in* B. Ketteman and W. Wieden, eds, 'Current Issues in European Second Language Acquisition Research', Gunter Narr Verlag, Tubingen.

Skolverket (2010), Rustad att möta framtiden?, Technical report.

Skolverket (2015), 'PISA i korthet'.
URL: http://www.skolverket.se/statistik-och-utvardering/internationella-studier/pisa

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. and Potts, C. (2013), 'Recursive deep models for semantic compositionality over a sentiment treebank', *Empirical Methods in Natural Language Processing* pp. 1631–1642.

Sparck Jones, K., Walker, S. and Robertson, S. (2000), 'A Probabilistic Model of Onformation Retrieval: development and comparative experiments', *Information Processing & Management* 36, 809–840.

Sumanth, C. and Inkpen, D. (2015), 'How much does word sense disambiguation help in sentiment analysis of micropost data?', *6Th Workshop on Computational Approaches To Subjectivity, Sentiment and Social Media Analysis Wassa 2015* (Wassa), 115.

Swedish phrasal verbs (2015).
URL: https://en.wiktionary.org/wiki/Appendix:Swedish_phrasal_verbs

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011), 'Lexicon-Based Methods for Sentiment Analysis', *Computational Linguistics* 37(2), 267–307.

Tahir, M. A., Bouridane, A. and Kurugollu, F. (2007), 'Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier', *Pattern Recognition Letters* 28(4), 438–446.

Turchi, M. and Balahur, A. (2013), Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data, *in* 'Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP', number September, Hissar, Bulgaria, pp. 49–55.

Turney, P. (2002), Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews, *in* 'Proceedings of the ACL', pp. 417–424.

Venable, J., Pries-Heje, J. and Baskerville, R. (2016), 'A Framework for Evaluation in Design Science Research', *Eur J Inf Syst* 25(1), 77–89.
URL: http://dx.doi.org/10.1057/ejis.2014.36

Vinodhini, G. and Chandrasekaran, R. (2012), 'Sentiment Analysis and Opinion Mining: A Survey', *International Journal of Advanced Research in Computer Science and Software Engineering* 2(6), 282–292.
URL: http://www.dmi.unict.it/~faro/tesi/sentiment_analysis/SA2.pdf

Vohra, S. M. and Teraiya, J. B. (2013), 'A comparative study of sentiment analysis techniques', *Journal of Information, Knowledge and research in Computer Engineering* 2(2), 313–317.

Wan, X. (2009), 'Co-training for cross-lingual sentiment classification', *Proceedings of the Joint Conference of the 47th Annual ...* (August), 1–9.
URL: http://dl.acm.org/citation.cfm?id=1687913$\delimiter"026E30F$npapers2://publication/uuid/B6AA09E1-D4EA-497A-B0EE-5A3F68648302$\delimiter"026E30F$nhttp://www.aclweb.org/anthology/P/P09/P09-1027.pdf$\delimiter"026E30F$npapers2://publication/uuid/A45CADD0-CB44-4277-8279-4320F1C52239$\delimiter"026E30F$npapers2://publication/uuid/3FB

Wei, F., Ge, X., Zhang, L., Liu, X. and Zhou, M. (2012), 'Lost in Translations ? Building Sentiment Lexicons Using Context Based Machine Translation', *Coling 2012* (December 2012), 829–838.

Wettschereck, D., Aha, D. W. and Mohri, T. (1997), 'A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms', *Artificial Intelligence Review* 11, 273–314.

Appendices

Appendix A - Glossary of Terms and Abbreviations

Features	"Feature" in this research refers to "word" or "term". All three expressions are used interchangeably in this paper.	Training set	The training set contains labelled data with which to train the model by combining input data with the correct output data
Label, labelling	Where ever the word 'label/labelling' is used in the further proceedings of this thesis, giving a label/labelling is the classification of newspaper articles into the categories positive and negative.	Validation set	Previously unseen data for the model to verify training results and to tune machine learning parameters for optimal output. Also used to avoid overfitting.
IAA	Inter annotater agreement; the measure of how much human sentiment annotaters agree.	Test set	An unseen data set which is only used to test the final results of the model (often on real life data), to evaluate the final performance of the model.
Domain specific vs general data	Domain specific data is can for example be movie reviews, where the domain would be movies. General data is found in articles, which can be about anything and is can not be categorised	FEDS	A Framework for Evaluation in Design Science Research
Precision	Measures the exactness of the classifier. Percentage of correctly classified items	Idiomatic expressions	Expressions that are figurative, i.e. they mean something else than what they describe. Ex: "The ball is in your court"
Recall	Measures the completeness of a classifier. Percentage of correct items of one class	Phrasal verbs	A short expression composed of a verb and a particle that belong together for the expression to be meaningful. Ex: "break out"
F1-score	F1-score is also called F1-measure and is the weighted harmonic mean of precision and recall.		

Appendix B - Filter articles by topic with keywords

The word ’skola’ had to be in every article, with the following variations: skol*,gymnasie*,högstadie*. The three words above needed to appear in combination with any or several of the following words, in order for the article to be part of the topic selection:
lärare, elever, vikarie, problem, lärar*, *lärare, undervisning, *läxa, fritid, uppvisning, betyg*, tenta*, invandrare, immigrant, migrant, barn, stöd, funktionsnedsättning, särskilda behov, mobbning, näthat

Appendix C - Visualization of SVM accuracy with and without scaling

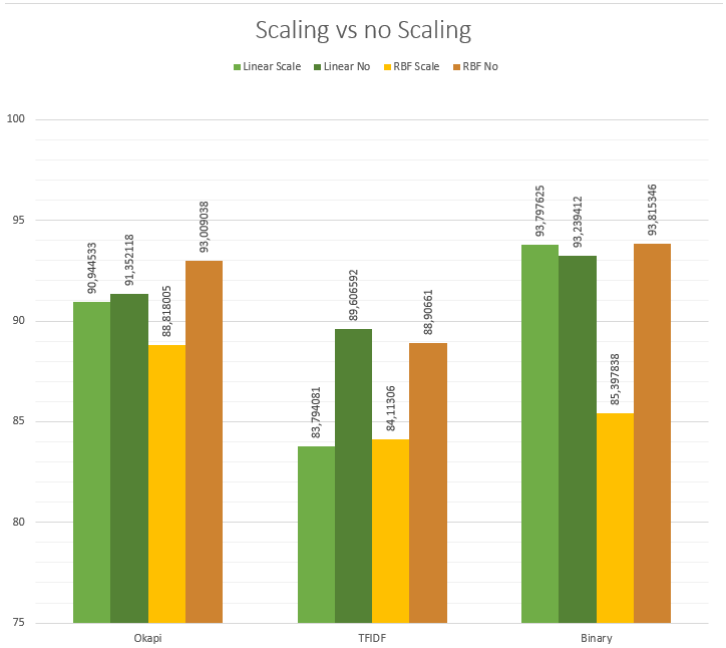

Figure 9: Comparison of SVM training results with Scaling versus no Scaling

Appendix D - Accuracy, Precision, Recall and F1-score calculation results of the Test set

These are the results for accuracy, precision, recall and F1-score for the validation set, see figure 10

```
"C:\Program ...
PrecisionRecall{
precision=0.8954545,
      recall=0.8242678,
      accuracy=0.83709276,
      f1_score=0.8583878,
      true_positives=197,
      false_positives=23,
      true_negatives=137,
      false_negatives=42,
      __total_rows=399
}

Process finished with exit code 0
```

Figure 10: Results for accuracy, precision, recall and F1-score for the test set

Appendix E - Accuracy, Precision, Recall and F1-score calculation results of the training data

These are the results for accuracy, precision, recall and F1-score for the training data, including neutral sentences, see figure 11.

```
"C:\Program ...
PrecisionRecall{
precision=0.70625,
     recall=0.49779737,
     accuracy=0.5975,
     f1_score=0.5839793,
     true_positives=113,
     false_positives=47,
     true_negatives=126,
     false_negatives=114,
     __total_rows=400
}
```

Figure 11: Accuracy, precision, recall and F1-score for the training data, including neutral sentences

The following is the result for accuracy, precision, recall and F1-score for the training data, without neutral sentences, see figure 12.

```
"C:\Program ...
PrecisionRecall{
precision=0.8560606,
     recall=0.7151899,
     accuracy=0.7887789,
     f1_score=0.7793104,
     true_positives=113,
     false_positives=19,
     true_negatives=126,
     false_negatives=45,
     __total_rows=303
}
```

Figure 12: Accuracy, precision, recall and F1-score for the training data, excluding neutral sentences

Appendix F - The practical use of sentiment analysis on newspaper articles

The results of machine classification of sentences in the newspaper articles taken from the database, shows that all five newspapers have written more negatively than positively about the topic 'school' between 2001 and 2016 (see figure 13).

Figure 14 shows that every year, consistently negative reports about the topic school have been written by all five newspapers (representative for five regions in Sweden), with peaks of negative reports between 2009 and 2010, between 2012 and 2013, and finally around 2015. These peaks coincide with the execution and report of the results of the

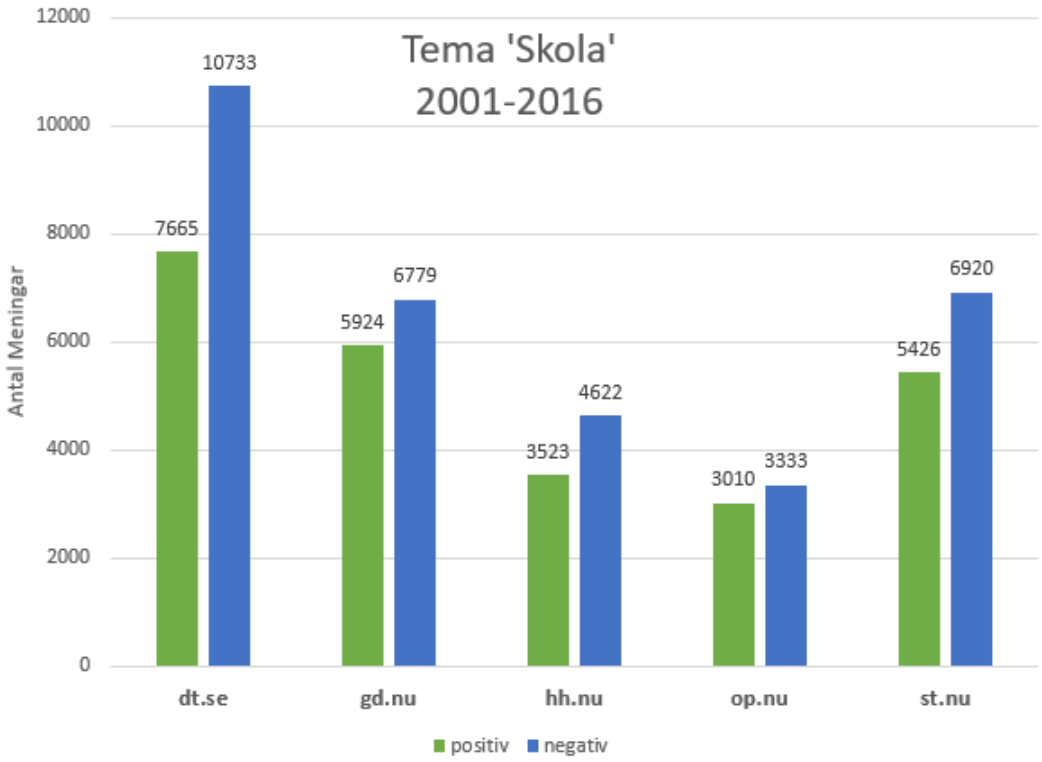

Figure 13: Positive versus negative sentences written about the topic 'school' sorted by newspaper

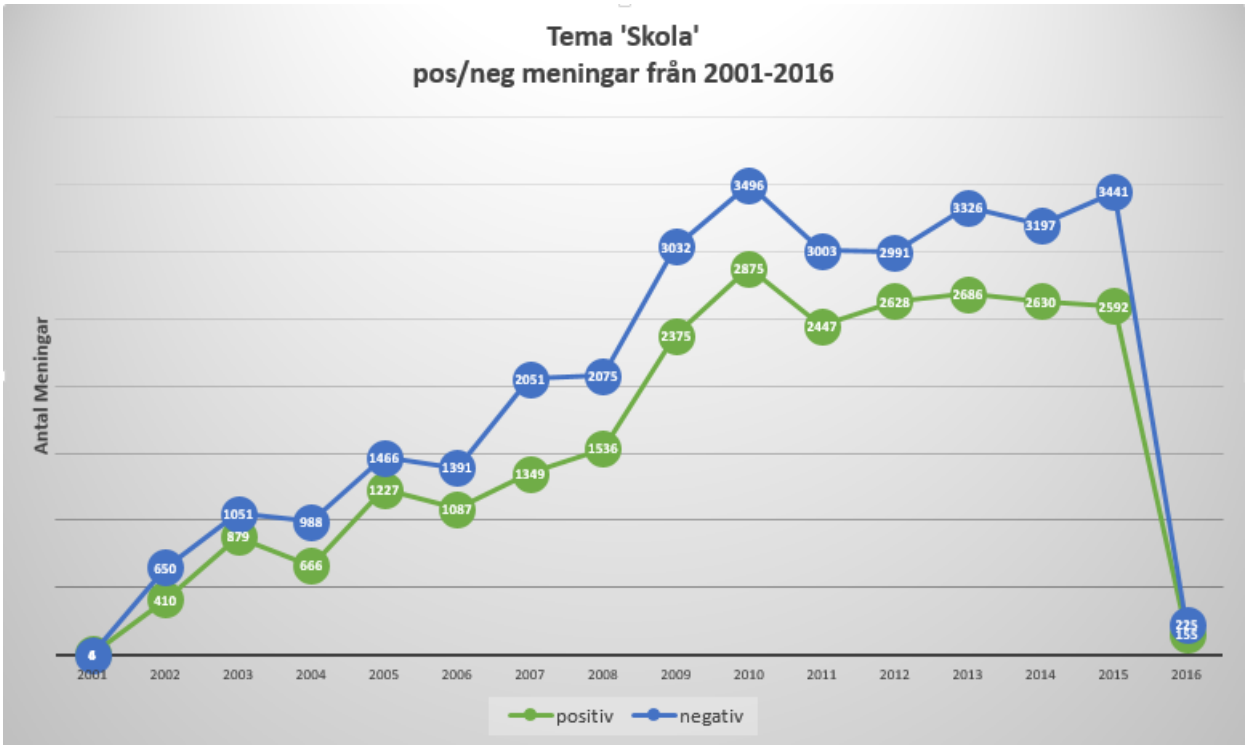

Figure 14: Positive versus negative sentences written about the topic 'school' sorted by year

international Pisa[4] study, which is carried out by the OECD[5] organisation every three years, beginning in the year 2000 (Skolverket, 2015).

[4]Pisa stands for "Programme for International Student Assessment"
[5]OECD is the acronym for "Organisation for Economic Co-operation and Development"

The Pisa study results from 2009, published in 2010, obviously sparked a debate in Swedish newspapers, since the study shows a significant downward trend of Swedish school childrens' knowledge with Sweden falling from the top of the list in 2000 to the OECD average in 2009 (Skolverket, 2010). Again in 2012 and 2013 newspapers lift the topic "school" since the results of the Pisa study 2012 show that the downward trend continues with Swedish maths and science skills below the OECD average (OECD, 2015). Finally, the results of the Pisa study 2015 will be published in December 2016, but the report "Improving Schools in Sweden: An OECD Perspective" from the OECD published in 2015, might already spark public debates in newspapers, as it warns Sweden urgently to improve its school system: "[...] the country's performance in the OECD's PISA survey has declined over the past decade from around average to significantly below average. No other country taking part in PISA has seen a steeper fall." (OECD, 2015).

Appendix G - Summary table for feature reduction following the structure of the model development

The following is a table that shows the number of unique features after some broad filtering steps.

	Original data (approx.)	Topic filtering ("school")	Feature filtering, Handlers	No Feature filtering and Handlers
Number of unique features	18 394 440	47099	18737	65303
Accuracy (%)			93,82	93,49

Figure 15: Summary table for feature reduction following the structure of the model development

Stockholm University
Department of Computer and Systems Sciences
Forum 100
SE-164 40 Kista
Phone: +46 8 16 20 00
www.dsv.su.se

Stockholm
University