English title: Time to Open the Black Box: Explaining the Predictions of Text Classification

Author: Helena Löfström

Completed: March 2018

Abstract: The purpose of this thesis has been to evaluate if a new instance based explanation method, called Automatic Instance Text Classification Explanator (AITCE), could provide researchers with insights about the predictions of automatic text classification and decision support about documents requiring human classification. Making it possible for researchers, that normally use manual classification, to cut time and money in their research, with the maintained quality. In the study, AITCE was implemented and applied to the predictions of a black box classifier. The evaluation was performed at two levels: at instance level, where a group of 3 senior researchers, that use human classification in their research, evaluated the results from AITCE from an expert view; and at model level, where a group of 24 non experts evaluated the characteristics of the classes. The evaluations indicate that AITCE produces insights about which words that most strongly affect the prediction. The research also suggests that the quality of an automatic text classification may increase through an interaction between the user and the classifier in situations with unsure predictions.

Keywords: Text Classification, Explanation Methods, Machine Learning
To Tuve, for always believing in me.
Contents

1 Introduction .................................. 1
   1.1 Problem Formulation .......................... 2
       1.1.1 Formulation of the Objectives ............... 2
   1.2 Basic Concepts .............................. 4
   1.3 Outline ..................................... 4

2 Literature Review .............................. 5
   2.1 The Research Area of Library and Information Science ................. 5
   2.2 Machine Learning - Introduction ............................. 5
   2.3 Automatic Text Classification - Introduction ......................... 6
       2.3.1 Document representation ......................... 7
       2.3.2 Characteristics of Document Classification .......... 8
       2.3.3 Supervised Classification ......................... 8
       2.3.4 Random Forest and Support Vector Machine .......... 9
   2.4 Automatic Text Classification - Review .......................... 10
   2.5 Explanation Methods ............................ 11
   2.6 Suggested Solution, the Automatic Instance Text Classification Explanator 12
   2.7 Related Work ................................ 13

3 Method ...................................... 15

4 Construction of the Data Set ................. 18
   4.1 Adaptation of Method ........................... 18
       4.1.1 Class Selection .............................. 19
   4.2 Results ..................................... 20

5 Implementation of the AITCE .................. 21
   5.1 Adaptation of Method ........................... 21
       5.1.1 Experimental Design ........................... 21
       5.1.2 Defining the Setup ......................... 21
   5.2 Results of the Experimental Study ......................... 22
       5.2.1 Results from the Classification ............... 22
5.2.2 Article Selection Criteria ........................................... 23

6 Evaluation of the AITCE ...................................................... 30
   6.1 Adaptation of method ................................................... 30
      6.1.1 Formulation of questionnaire ................................... 30
      6.1.2 Design of Presentation ........................................... 31
      6.1.3 The Questions ..................................................... 32
      6.1.4 Construction of the Web Survey ................................... 34
   6.2 Results of the self completion questionnaire ......................... 34
      6.2.1 Correctly predicted articles ......................................... 35
      6.2.2 Incorrectly predicted articles ....................................... 36
      6.2.3 Uncertain articles .................................................. 36
      6.2.4 Presentation ........................................................ 38
   6.3 Results of the web survey ............................................. 38

7 Concluding Discussion ..................................................... 41
   7.1 Analysis and reflections ............................................... 41
      7.1.1 The data set ........................................................ 41
      7.1.2 The implementation ............................................... 42
      7.1.3 The Evaluation ..................................................... 43
   7.2 Conclusion ............................................................. 44
   7.3 Repeatability ............................................................ 45
   7.4 Sources of errors ....................................................... 46
      7.4.1 The Data set ........................................................ 46
      7.4.2 The number of respondents ......................................... 46
      7.4.3 Choice of research method ......................................... 46
   7.5 Ethical aspects .......................................................... 47
      7.5.1 The respondents .................................................... 47
      7.5.2 Sustainability ....................................................... 47

8 Future Work ................................................................. 48

Bibliography ................................................................. 48

A The Self-completion Questionnaire ....................................... 51
List of Tables

4.1 Class and number ................................................................. 20
4.2 Document length (number of words/features) within the classes in the
data set ................................................................. 20
4.3 Number of features before and after stop word removal ............. 20

5.1 Distribution of PE among the predictions ................................. 22
5.2 Selection of articles for evaluation ........................................ 24
5.3 Feature EC, Article 1 ............................................................ 25
5.4 Feature EC, Article 2 ............................................................ 26
5.5 Feature EC, Article 3 ............................................................ 27
5.6 Feature EC, Article 4 ............................................................ 27
5.7 Feature EC, Article 5 ............................................................ 28
5.8 Feature EC, Article 6 ............................................................ 29
List of Figures

2.1 Machine learning and the task of classification (from Flach (2012, p. 11)) 6

5.1 Distribution of documents within the interval 0.4-0.6 among correct and incorrect predictions. 23

6.1 The two types of classification. 31
6.2 The first type of presentation 32
6.3 The second type of presentation 33
6.4 Results of class Politics 39
6.5 Results of class Social issues 40
Chapter 1

Introduction

We are living in a world of constant change, where every change produces information, and by this an increasing amount of new documents; documents to store and classify (or categorize). Already 50 years ago people within the Library and Information Science (LIS) tried to find ways to keep up with this development, by automatizing the procedure with the help of computers. Library Science (LS) has since used the advantages that resulted from the collaboration between Information Science (IS) and Computer Science (CS) to make automatic text classification more efficient (see e.g. Sebastiani, 2002, pp. 1, 5; Eklund, 2016, pp. 2-3; van Rijsbergen, 1975, p. 1). Machine learning (ML) became popular in the 1990’s, and through the advantages of high levels of efficiency, high accuracy and the ability to easily add rational structure to the increasing amount of documents, it became the dominant solution for text classification (see e.g. Sebastiani, 2002, pp. 8,9; Eklund, 2016, pp. 2-3).

Making automatic text classification more efficient using ML has been attracting more attention of researchers (see e.g. Ribeiro et al., 2016; Dzindolet et al., 2003; Martens and Foster, 2014), through the development of new text classifiers. Many of these text classifiers are behaving more or less like a black box, as it is expressed by Ribeiro et al. (2016, p 1135), where data is given the classifier and predictions are presented to the users. When the users do not understand the classifiers behaviour, it could according to Dzindolet et al. (2003) make the users unwilling to use them, although the classifiers provide good results. Dzindolet et al. (2003) also showed that if the users are given a reason why the automatic classification might give an erroneous answer, the users did not become unwilling to use the classifiers. In other words, if users of automatic text classification would have access to a method where the predictions are explained they would probably, as is seen in Martens and Foster (2014), become more inclined to use automatic text classification in their work.
1.1 Problem Formulation

For decades, the focus within ML has been how to train a classifier to get the most accurate predictions (Baeza-Yates et al., 2011, p. 282). This has resulted in many well functioning classifiers. But as Ribeiro et al. (2016) write, when the users of automatic text classification don’t agree or understand the predictions of the model, they require some sort of explanation aid. In their research Dzindolet et al. (2003, p. 704) show that users tend to become unwilling to use classification models when they realize that the classifier sometimes may be erroneous, but if the users get the reasons explained why and how the classifier made the errors, they tend to be more willing to use them. Martens and Foster (2014) show in their work that it is very important to explain the predictions of document classification for the users, not only the prediction of the entire model, and points out that not much is done, encouraging more research in the area.

When using manual text classification within research, large quantities of documents are used. To have to classify thousands of documents is not unusual, which is done by training several human coders, often for weeks. These coders are those who classify the documents. The costs are huge and it is very time consuming, but it is seen as the best and most accurate way (Respondents, personal communications, Spring, 2016).

Automatic classification would give the researchers the possibility to decrease the cost of classification, making it possible to use a larger part of the funds to research. Although, there are always documents that are difficult to classify automatically and could become incorrectly classified. With an explanation aid that not only presents insights to why the documents have been predicted to belong to a certain class, but also pinpoints the documents that have are more likely to be incorrectly predicted, the researchers can maintain quality to a reduced cost.

It is obvious that researchers that use automatic text classification, especially those who have not used it before and may feel uncertain when it comes to technical systems, would benefit from an explanation tool that could provide them with insights about the causes of the predictions as well as providing decision support by indicating which documents that may require human classification.

1.1.1 Formulation of the Objectives

From the discussion above, a main objective was formulated: Develop and evaluate an explanation method which can provide researchers with insights about causes of the predictions and provide decision support about documents requiring human classification.

From the main objective, three sub-objectives were formulated. The main objective will be accomplished by fulfilling the following sub-objectives:

1. Construct a suitable data set
   The classifier requires data to work with from the area of research, already classified by humans. The data will be used to train the chosen classifier and to test the
validity of the results.

2. *Implement an instance based explanation method*

   To be able to evaluate if the explanation method could provide researchers with insights about the causes of the predictions of an automatic text classification, by identifying the keywords in a document that most strongly affect the prediction, an explanation method that has the potential to provide insights at instance (document) level has to be implemented. Furthermore, the explanation method must also be able to provide decision support to help researchers decide which documents that may need human coding and which can be automatically classified.

3. *Evaluate the explanation method*

   To ensure that the explanation method can provide researchers with insights about the causes of the predictions, results from the first two objectives require evaluations by senior researchers within the domain, who are familiar with both the process of manual text classification and with the evaluated data and who may order automatic text classification in future research projects. For practical reasons, the senior researchers can only make an in-depth analysis of a handful of documents. In order to evaluate if the explanation method has the possibility to create explanations of the main characteristics of the classes and in that way evaluating the explanation method with the entire data set, a web based survey with a group of non experts will also be used.

Evaluations (sub-objective 3) could make it possible to investigate if an explanation method is able to provide researchers with insights about the causes of the predictions in an automatic classification. To be able to perform these evaluations, two tasks are fundamentally important: An instance based explanation method has to be available (sub-objective 2), as well as a data set that has been used in a project with manual classification (sub-objective 1). In the thesis, an instance based explanation method has been implemented and used, resulting in a classified data set from an existing research project, which will make it possible to evaluate if the chosen explanation method can provide researchers with insights about the causes of the predictions from the automatic text classification. Apart from the evaluation involving senior researchers, an analysis of the predictive results will also be carried out in order to show how the predictive performance varies depending on how certain the prediction is.

This thesis is written within the domain of LIS. A domain where, to the best knowledge of the author, no such research has been done before. The research is directly focusing on researchers working with manually classified documents, to help them when automatically classifying texts. The number of explanation methods for black box text classifiers at instance level are not many, and as late as 2014 Martens and Foster (2014) pointed out that this was an area which requires more interest from researchers. Since then only one method, to the knowledge of the author, has been developed: LIME from Ribeiro et al. (2016), which indicates that there is still much to do within the area. As
far as the author of this thesis has succeeded to find, there is still no research directed towards the requirements of researchers requiring text classification.

1.2 Basic Concepts

Some concepts used in the thesis are important to define, to avoid misunderstanding and increase readability:

- **Features**: Also called variables, are in the thesis the words within the data set.

- **Accuracy**: Measures the fraction of the documents in the test set that are assigned to their correct classes by the classifier, Baeza-Yates et al. (2011, p.326).

- **Instance**: The object of interest within the data set. In the thesis the instances are the individual documents.

- **Explanation word**: Words, presented to the user by the explanation method, to give insights/explain about the prediction of a document.

- **Estimate coefficient**: A statistical probability value, used in the thesis to express in what direction an explanation word, in a document, influences the prediction. Abbreviated as EC.

- **Probability estimate**: A statistical probability estimate value, used in the thesis to express an estimate of the probability for a document to belong to a certain class. Intimately connected to the results of the classifier. Abbreviated as PE.

1.3 Outline

The thesis will have the following outline: A literature chapter with an overview covering the research area, a short introduction to machine learning, text classification, explanation methods, as well as related work. Each of the objectives is addressed in chapters 4, 5 and 6, where both the methodological set up and the results are to be found. Chapter 7, Concluding discussion, covers both possible sources of errors, a discussion of the results of the objectives, conclusions, ethical aspects as well as an outline for future work.
Chapter 2

Literature Review

2.1 The Research Area of Library and Information Science

The research area of LIS is sometimes technologically driven and interdisciplinary, where some of the most frequently studied areas are information, web technologies and text mining (see e.g. Zhu et al., 2016, pp. 1464, 1474; Huang & Chang, 2012, p.1; Tuomaala et al., 2014, p. 1459). As Wildemuth (2009, p. 3) writes, it is a field that includes both “the profession of librarianship as well as a variety of other information professions”. Tuomaala et al. (2014, p. 1459) writes that the interdisciplinary tendency, that was strengthened in the 1990’s and the early 2000’s, occurred between LIS and CS. According to Wildemuth (2009, p. 3) the research within the field of LIS involve the improvement of practice of the information content, the people who will interact with the content as well as the technology used to support the creation, communication, storage, or transformation of the content. In light of the interdisciplinary tendency of LIS, it is natural that the thesis has an interdisciplinary tendency, with components from ML, a subfield of CS.

2.2 Machine Learning - Introduction

Machine learning (ML) is an area within Artificial Intelligence (AI), where algorithms are designed and developed for learning patterns (see figure 2.1\(^1\)) in data (Baeza-Yates et al., 2011, p. 282; Margaret, 2003, p. 43). It is within ML that many of the algorithms for text classification have been developed. The algorithms produced are said to be able to learn the patterns provided via the input of data (Baeza-Yates et al., 2011, p. 282). These patterns are used to make predictions about new and unseen data (Mar-

\(^1\)Publicized with approval from the author
Classification is one possible task in a broad range of domains where the applications of ML algorithms are implemented; from natural language processing, to stock market analysis. There are many different kinds of ML algorithms that can be used for classification, which has resulted in several different classifiers: linear classifiers, Bayesian classifiers, distance-based classifiers, etc. Document classification is only one possible task for these models.

2.3 Automatic Text Classification - Introduction

Text classification (TC) is a process where a set of documents is divided into subsets that have something in common. These subsets are called classes, categories, or clusters (Baeza-Yates et al., 2011, p. 281). When the documents are divided into groups they get one or more meaningful label. The labels are symbolic, and intended to characterize or in a way explain the documents in the class (Sebastiani, 2002, p. 1). The classes can describe different characteristics of the document such as genre, language, topic, etc. (Baeza-Yates et al., 2011, p. 281). According to Sebastiani (2002), some of the most important applications of TC are document organization, text filtering, and word sense disambiguation. The functional representation of text classification could be written as a binary function $F$ (Baeza-Yates et al., 2011, p. 283):

$$F : D \times C \rightarrow \{\text{True}, \text{False}\}$$

where $D = \{d_1, d_2, ..., d_j\}$ is a set of documents and $C = \{c_1, c_2, ..., c_l\}$ is a set of classes. If the function returns $\text{True}$, the document is assigned to the class and if the
function returns $\text{False}$ the document is not assigned to the class (Sebastiani, 2002, p. 3). This requires the system to assign the value $\text{True}$ or $\text{False}$ to every pair of possible class ($c_i$) and document ($d_n$), $<d_n, c_i>$ (Sebastiani, 2002, p. 3). There are both single-label text classification (where every document gets exactly one class) and multi-label text classification or overlapping categories (Sebastiani, 2002, p. 3). The single-label classification is more general and more frequently used than the multi-labelled. In the thesis, single label classification is used.

2.3.1 Document representation

Before the process of automatized text classification may start, the complete document, if not born digital, is converted into a digital, computable representation which computers can handle, called document representation (van Rijsbergen, 1975, pp. 12; Baeza-Yates et al., 2011, p. 58; Eklund, 2016, pp. 115). The process to create a document representation often involves a number of steps (Eklund, 2016, pp. 116):

1. Tokenization, which is the process when all the sentence delimiters (white-space, comma, punctuation marks etc.) are removed and the text is converted into a string of 'tokens' (or word units) (Eklund, 2016, pp. 116). For example: “Once upon a time” → (Once, upon, a, time)

2. Normalisation. In the next step the tokens are mapped onto the same lexeme or stem. The process results in a number of subsets that is called equivalence classes (Eklund, 2016, pp. 116). When removing a word’s affixes, you get the word’s ’stem’. Two tokens are in the same equivalence class if and only if they are equivalent to each other $^2$. Which, as Eklund writes, results in that they are treated as different forms of the same lexeme or stem, and are said to have the same characteristic. Stemming is not a compulsory part of the document representation process. In Forman (2003) it is argued that stemming is merely an engineer option and that a frequency count could be as effective. He also writes that this could be especially useful when documents vary a lot in length, arguing for a Boolean feature representation when it comes to short documents.

3. Feature or keyword selection. Automatic text classification of documents often results in a large feature space, since all the terms in the document representation is considered a separate feature (or variable) (Baeza-Yates et al., 2011, p. 320). To meet the problem when the classification algorithms cannot handle the complexity of the very high dimensional feature space, a subset of the features are chosen to represent the document (Baeza-Yates et al., 2011, p. 320; Eklund, 2016, p. 118). A method called stop word removal could be used to reduce the feature space. In this method words that are defined as stop words (unnecessary words that could

$^2$http://mathworld.wolfram.com/EquivalenceClass.html
be eliminated from feature space) are stored in a separate file which is used to identify the words and eliminate them. The stop word list could be based on words that have little semantic meaning or a high frequency in the texts (Eklund, 2016, p. 118).

When these steps are finished, each document representation has been reduced to a feature vector (Baeza-Yates et al., 2011, p. 320)

2.3.2 Characteristics of Document Classification

As Martens and Foster (2014) write, document classification is in several ways different from other traditional data mining applications e.g. medical analysis. Within document classification the data has less structure and the number of features is very high, which results in a high dimensional feature-vector space. Another characteristic of document classification is that the value of the features also indicates the frequency of occurrence in the documents (Martens and Foster, 2014, pp. 86-87). The two first characteristics makes explanation of document classification difficult, but the third characteristic also points at a possible starting point for building an explanation method as seen in Martens and Foster (2014).

2.3.3 Supervised Classification

When data that are used to train the models are documents already classified by humans, it is called supervised classification, or just classification which is the concept used in the thesis (Baeza-Yates et al., 2011, p. 291). The classifier is first trained on a smaller set of data, called training set, before being applied to the larger document collection (see figure 2.1 from Flach, p. 11). According to Baeza-Yates et al. (2011, p. 291) the training set function is defined as, given a sub-collection $D_t \subset D$ of documents:

$$T : D_t \times C \rightarrow \{True, False\}.$$  

In other words, $T$ assigns a value of True or False to each pair $\{d_j, c_p\}$, $\{d_j \in D_t\}$ and $\{c_p \in C\}$, according to the judgment of human specialists.

Evaluating the Classifier

The final phase of the training of the classifier is the evaluation, when the classifier is applied to the unseen data in the data set. It is done by comparing the results from the classifier with those produced by human specialists (Baeza-Yates et al., 2011, p. 292). The most common situation is that there is some difference between the human produced classification and the automatic classification. The human produced classification is seen as the most accurate and the more the classifier matches the human classification, the higher its effectiveness, also called accuracy (Baeza-Yates et al., 2011, p. 292).

When evaluating the classifier a pre-selected set of documents is used where the correct classes are known (Baeza-Yates et al., 2011, p. 291). This pre-selected set is called
the test set and as with the training set, the assignment of classes to documents is already done by human coders. It is important that the test and training sets are disjoint to guarantee that the evaluation is not done on documents known by the trained model, but that the documents are new and unseen. In situations when there is no pre-selected set of documents for testing, which is not unusual, the entire data set is split into two sets. One used for training and one used for testing.

When using a small data set, a suitable model of evaluating to use is one called Leave One Out Cross Validation. In this model, the evaluation is done by not dividing the data set into a predefined sets of training and test. Instead, all but one instance, which is used as test, is used to train the model. This is repeated n times. Each time one, new, instance is used as test and all other to train the model.

2.3.4 Random Forest and Support Vector Machine

Two state-of-the-art algorithms were considered as classifiers to use in the experiments: Support Vector Machine (SVM); Random Forest (RF).

The Random Forest Classifier

A predictive model that has been found to be very robust is the ensemble model. An ensemble is a model built by a set of other models, so called base classifiers. The ensemble is making its predictions by combining the predictions from the base classifiers and has been found to be more accurate than the individual base classifiers. Diversity is a property that together with accuracy are the most important aspects when succeeding in building a successful ensemble model. The base classifiers within the model must be different (diverse) from each other, but also accurate. Since a too high diversity between the base classifiers may result in less accurate base classifiers, it is important to create a balance between diversity and accuracy (Löfström, 2015, p. 6).

The Random Forest Classifier (RF) combines the predictions from a number of decision trees in an ensemble model (Breiman, 2001). A Decision tree is defined in Quinlan (1986) as a set of structures in a tree-like structure, beginning in the root and proceeding out to the leafs. Random forests creates diversity by using bagging, i.e., training each base classifier using a bootstrap aggregate sample, and by only using a random subset of all features when creating each split in the decision trees.

This classifier is one of the most accurate classifiers used within machine learning today (Foster et al., 2016, p. 171). It is efficient to run on large data sets and is capable to handle a large number of features.

The SVM classifier

The SVM (Support Vector Machine) classifier is a rather new and complex method, but one of the most popular and best performing (Foster et al., 2016, p. 164). According
to Baeza-Yates et al. (2011, p. 314), SVM is directly applicable to automatic text classification, when the documents are represented by weighted\(^3\) multidimensional vectors. Joachims (in Baeza-Yates et al. (2011, p. 314)) as well as Sebastiani (2005, p. 4) also point out that SVM is specially suitable for text classification, since it deals very well with highly dimensional documents. One drawback is that they generally need hyperparameter tuning.

The SVM is a vector space method for binary classification problems, where documents are represented as vectors (Baeza-Yates et al., 2011, p. 307). Given the vector representation of the documents in a data set and two classes \(\text{class}_1\) and \(\text{class}_2\), the idea is to find the line, the decision hyper plane, that best separates the elements in the two classes (Baeza-Yates et al., 2011, p. 307). To find the best line among all the lines that could separate the documents, the line that maximizes the distance between the closest documents of the two classes is sought (Baeza-Yates et al., 2011, p. 307).

### 2.4 Automatic Text Classification - Review

Humans have been working to classify texts and documents as long as there has been libraries. The task of labelling and in this way attaching meta-data to documents manually, is manageable as long as the size of the collections stays relatively small (Baeza-Yates et al., 2011, p. 281). When the collections’ growth accelerates, as the trend has looked like since the 1960’s, the task becomes increasingly difficult to perform by human experts. Not long after the introduction of automatized text classification, the project MYCIN started to study “the ability for intelligent systems to explain their decisions”, as it was seen necessary for effective use of such systems (Martens and Foster, 2014, p. 77).

Sebastiani (2002, p. 1) writes that since the 1990’s text classification has become a major research field within the information systems discipline, applied in many different contexts, with a numerous application domains, closely connected to IT. He also mentions that the manual alternative cannot compete with the short response time automatic text classification requires. Given the increasing number of documents born digital, automated text classification will continue to increase both in number and importance (Sebastiani, 2002, p. 1). Text classification can also improve the productivity of human classifiers, in applications in which no classification decision can be taken without a final human judgment.

Today the exponential growth of textual information online has made the hardware and software solutions of storing, organizing, and retrieving become an important and demanding issue (Sebastiani, 2002, p. 1; Pang and Jiang, 2013, p. 576). Text classification is a significant tool for handling this issue (Pang and Jiang, 2013, p. 576).

Although the document classification tools have a high level of accuracy, the humans

\(^3\)to give a statistical weight to. [http://www.dictionary.com/browse/weighted](http://www.dictionary.com/browse/weighted)
require understanding i.e. to have them explained, when making decisions based on them (Martens and Foster, 2014, p. 73). A research by Dzindolet et al. (2003) showed the importance of explaining the results of an automatic classification. It showed that when humans realized that the automatic classification tool could be erroneous, the trust dropped considerably. But the more the humans understood how the classifier worked, the more they trusted it. Dzindolet et al. (2003, p. 697) also showed that if humans understand how the classifier works or the reasons behind the classification, it makes the humans use them more effectively.

Martens and Foster (2014) wrote that many document classification tools need human understanding when making data-driven classification decisions. When classifying documents, there are problems with very high variable dimensionality, often with tens of thousands to millions of variables. Martens and Foster (2014) showed that the high dimensionality makes it very difficult for humans to understand the decisions made by the document classifiers. The necessity for understanding the reasons behind the predictions of the classifier has not had much attention, and Martens and Foster (2014) show in their work the reasons why explanations are needed; to detect anomalies, to grasp the inner workings of the intelligent system and for long-term learning. They also emphasized the requirement not just for a general explanation, but also for each of the documents in a data set, i.e. at instance level. They wrote, that to their knowledge, no method provides such explanation for the very high-dimensional models.

In Ribeiro et al. (2016) it was discussed how faithful and intelligible understanding of the relationship between the instance’s components, as words, and the prediction of a model was an important aspect to get humans to use machine learning effectively.

2.5 Explanation Methods

Ribeiro et al. (2016, p. 1135), define the term explanation method as a method to explain a prediction from a classifier, by some sort of textual representation that creates "a qualitative understanding of the relationship between the instance’s components (e.g. words in text, patches in an image) and the model’s prediction", which is similar to how the term is used in the thesis. In their article they wrote, that if the explanations are faithful and intelligible the explanation methods are important in getting humans to use machine learning more effectively. This is similar to what (Martens and Foster, 2014, p. 80) also write; if users do not understand the classifier, they will be skeptical and reluctant to use it.

An explanation method makes the decisions of the classifier transparent, and in that way independent of the accuracy (correctness) of the prediction. But as Robnik-Šikonja and Kononenko (2008, p. 590) write, we may assume that the quality of the explanations rises with higher accuracy of the predictions.

There are different levels of explanation methods in predictive models (classifiers). Martens and Foster (2014, p. 86) mention two levels; model (or global) explanation level
and instance explanation level. At model level the explanation method provides greater understanding of the entire classification model and its performance, while the instance level methods provide greater understanding of the model’s predictions of a specific instance (see e.g. Martens and Foster, 2014, p. 86; Robnik-Šikonja and Kononenko, 2008, p. 590).

In 2009 a group working within biochemistry at Astra Zeneca wrote an article about a new algorithm, called Alg in the thesis, where the molecular descriptor that made the largest change to the function was considered (Carlsson et al., 2009). They wrote that only the molecular descriptor that, when looking at a prediction of a specific class, contributes the most is the interesting one, i.e. explains the prediction.

2.6 Suggested Solution, the Automatic Instance Text Classification Explanator

The explanation method, called Automatic Instance Text Classification Explanator (AITCE), that is implemented in the thesis was inspired by the algorithm Alg, developed by Ahlberg et al. (2015). They wrote that the theory behind Alg is that the gradient of a function from a point indicates how the function will behave in the local neighbourhood to that point. Assuming that the gradient is possible to calculate, it yields a linear approximation, making it possible to explore the local neighbourhood to the point. In the original domain the local neighbourhood was the chemical space. In the thesis the expression local neighbourhood refers to the neighbourhood of the vector representation of an instance. When Alg was developed, only the descriptors (features) that made the largest change to the function was considered. In this research it will correspond to which features that most strongly affect the prediction. The idea behind AITCE is similar to those used by Martens and Foster (2014) and Ribeiro et al. (2016), where a minimal set of words that influences the prediction, is used as an explanation. These explanation methods are not, to the knowledge of the author, developed as a tool to support researchers in their work, but as a support for decisions (mostly within the domain of business management and medicine).

---

Robnik-Šikonja and Kononenko (2008, p. 590) also mention a third level; domain explanation. They write that this level is unknown, but should, if the accuracy is high enough, be similar to the model explanation level.
Algorithm 1 AITCE

Input: The text classifier: $M$, the document to get explanation for: $x$, the feature vector of $x$: $\text{vect}$

1: Calculate the probability estimate for the default class of $x$: $p = M(\text{vect})$
2: for $f = 1$ to the number of features do
3: Set value of feature $f$ in $\text{vect}$ to 0: $\text{vect}'_f = 0$
4: Calculate probability estimate for the default class: $p'_f = M(\text{vect}')$
5: Calculate the explanation coefficient: $ec_f = p'_f - p$
6: Reset $\text{vect}$ to its original state
7: end for

Output: $ec$, i.e., how much each excluded feature affects the default class
$p'$, i.e., probability estimates after setting the value of $f$ to 0.

In the thesis AITCE is implemented in the program Python to work with the results from a black box text classifier. The RF classifier was used, but AITCE is general and other classifiers could also be used, as long as they are able to produce some sort of probability estimate for the classes. The class Bunch that was used is from the library Scikit-learn (see Pedregosa et al. (2011)).

As seen in Algorithm 1, AITCE is implemented as a function which takes the text classifier $M$, the document $x$ that is going to get explanations and the document vector of $x$, $\text{vect}$, as input. AITCE steps through each of the active features $f$ in $\text{vect}$\(^5\) in the document $x$, and calculates how the PE is changed if the feature is excluded from the document. The default class is the class achieving highest PE. The EC gets a positive value if the excluding of the feature strengthens the prediction, and negative if it weakens the prediction. AITCE returns two lists; one list including how much each feature affects the default class and one with the PE if excluding each feature.

2.7 Related Work

Over the last 3-4 years, there has been an increase of interest in explanation methods at instance level. Martens and Foster (2014) published an article where they wrote that their main theoretical contribution was, to their knowledge, a totally new explanation method at instance level, for document data. Data that are characterized with a very high feature dimensionality, making it difficult to understand the decisions made by the classifier. In their article, they discuss the problem of people not understanding the classification process, particularly when using so called black box classifiers (e.g. SVM and RF). These classifiers have a low level of transparency and lack any sort of explanation of their results. The method Martens and Foster (2014) presented explains the predictions

\(^5\)The feature vector is binary, where the number ’1’ indicates an active feature for the current document, i.e. that the word represented by the feature exists in the document.
by a set of minimal words, such that if removing all of them the prediction for that document will change. Martens and Foster (2014) work within the domain of business management and it is within that domain the article is written and the AITCE is intended to work; as a help for decision making and providing better understanding of the reasons for the classification decisions, as well as making the business application better itself.

In August 2016, when the work with this thesis started, another article was published by Ribeiro et al. (2016). They introduced a method called LIME, which is similar to AITCE, although the evaluation is partly performed with simulated users and non-expert humans. Ribeiro et al. also write that other explanation methods that are similar to the one they present "do not address the problem of explaining individual predictions".
Chapter 3

Method

In the chapter of introduction, a main objective was formulated:

*Develop and evaluate an explanation method which can provide researchers with insights about causes of the predictions and provide decision support about documents requiring human classification.*

The main objective will be accomplished by fulfilling the following sub-objectives:

1. *Construct a suitable data set*
   
The classifier requires data to work with from the area of research, already classified by humans. The data will be used to train the chosen classifier and to test the validity of the results.

2. *Implement an instance based explanation method*
   
   To be able to evaluate if the explanation method (the AITCE) could provide researchers with insights about the causes of the predictions of an automatic text classification, by identifying the keywords in a document that most strongly affect the prediction, an explanation method that has the potential to provide insights at instance (document) level has to be implemented. Furthermore, the AITCE must also be able to provide decision support to help researchers decide which documents that may need human coding and which can be automatically classified.

3. *Evaluate the explanation method*
   
   To ensure that the AITCE can provide researchers with insights about the causes of the predictions, results from the first two objectives require evaluations by senior researchers within the domain, who are familiar with both the process of manual text classification and with the evaluated data and who are likely to order automatic text classification in future research projects. For practical reasons, the senior researchers can only make an in-depth analysis of a handful of documents. In order to evaluate if the AITCE has the possibility to create explanations of the main characteristics of the classes and in that way evaluating the AITCE with the entire data set, a web based survey with a group of non experts will also be used.
The purpose of the experimental study was to implement the explanation method AITCE, as mentioned in the literature chapter, to use with the results of an automated text classification. The purpose of the self-completion questionnaire, was to find if the implemented explanation method (AITCE) was able to provide researchers with insights about the causes of the predictions of an automatic text classification. The questionnaire was directed towards senior researchers, using manual classification in their work, which in this way made them possess a unique competence for evaluating such a questionnaire. Approaching researchers when evaluating an explanation method has, to the knowledge of the author, not been done before.

The purpose of the web survey was to find if the AITCE, has the possibility to explain the main characteristics of the classes and in that way evaluating the AITCE with the entire data set.

The evaluations (sub-objective 3) of the AITCE (sub-objective 2), applied to the data (from sub-objective 1), will make it possible to fulfill the main objective if the AITCE has the potential to provide researchers with insights about the causes of the predictions and decision support about documents requiring human classification.

Obviously other methods could also have been used to address the main objective. Instead of a self-completion questionnaire, a structured interview could be used, which is one of the most commonly used methods in social research (Bryman, 2012, p. 209). It was not an obvious choice to choose a self-completion questionnaire before the structured interview, but as Bryman (2012, p. 232) points out, the structured interview is in many ways a questionnaire that is administered by an interviewer. The difference in presenting the questions lies mostly in the situation of either using the modern technique (internet or e-mail) or performing an interview via telephone. The use of Internet or e-mail offers some advantages, as this way of conducting research is cheaper, quicker to administer and not the least, as Bryman (2012, p. 233) writes, the self-completion questionnaire has an almost complete absence of interviewer effects.

The advantage that finally made the choice between the two different ways of conducting the research, was the convenience for the respondents. The three respondents in the research are very busy senior researchers in different universities across Sweden. The possibility to complete the questionnaire when they had time, made it possible to avoid a hurried judgment from the respondents.

All three respondents were connected to the project ‘Gammelmedia’ from where the data was collected/taken. The respondents knowledge about the data could have influenced the results of the questionnaire, but it was not seen as negative. Since some of the respondents had been trained together with the coders, it was expected that they would be more inclined to question incorrect predictions and be more critical towards insignificant explanation words. Thus evaluating the AITCE in a non favourable, more realistic, situation.

The second evaluation was a web based survey, where the prospective respondents were invited from a Facebook group of adults. The group was chosen since it has members from all parts of the society, opinions and interests. It is only their self-identified
identity as gifted in any area, that binds the members together. the total amount of mem-
bers were 525 members and 24 answered the survey.

From the discussion above it is obvious that the self-completion questionnaire was a
strong alternative to the web survey. But, as it is suggested by Sheehan and Hoy (in
Bryman (2012, p 670)), there is a tendency to use e-mail surveys in smaller and more
homogeneous groups, whereas web surveys are more often used in the study of larger
groups of on-line users, which characterizes the group of adult non-experts. But as Bry-
man (2012) also points out, surveys have characteristics from both structured interviews
and self-completion questionnaires, which made the choice of web survey the best alter-
native.
Chapter 4

Construction of the Data Set

4.1 Adaptation of Method

The data that is used within the thesis consists of articles collected from different newspapers, in a restricted time span\(^1\), in a Swedish research project called ‘Gammelmedia’. The data has been manually classified by human coders within the ‘Gammelmedia’ project, following a document of instructions. The human coders were trained for several weeks and tested to maximize accuracy and agreement between the coders.

The data set consists originally of approximately 5000 articles. Only approximately 1,300 were stored digitally and easy to access. Consequently, a convenience sample was done and only the digitally saved files were taken into account in the research.

The data was stored as PDF-files, where each file was a scanned page from a newspaper. The articles have been collected from ten different Swedish newspapers: Aftonbladet, Dagbladet, Dagens Nyheter, Expressen, Metro, Mitt i Huddinge, Södermalmssnytt, Sundsvalls Tidning, Svenska Dagbladet and Värmlands Folkblad.

In the code instructions the coders were instructed to work with every third news item (article). The coders chose articles by counting them from top to bottom in every PDF-file. If the coders came upon an article that was already coded, they should move on to the next one in the page. The coders were also encouraged to report their primary interpretation.

There were no obvious connection between the coded material and the articles, other than the title of each article, the media group the newspaper belonged to and the date of publication. Via the combination of the title, date and media group it was possible to trace the text of the articles, but since redundancy (meta data about a single article were found to have been stored on several different places in the coded document, with slightly different or identical names) was detected, only one meta data post was saved. The title of the articles was not always the exact title, but could be a set of keywords.

\(^1\) Between 10 January 2014 to 16 September 2014
The articles had been scanned a whole newspaper page at a time, which is difficult to handle automatically since this would result in a disturbing noise when reading the data automatically. Therefore it was necessary to clear each article from the noise. The noise being everything else but the actual article: it could be other articles, page numbers, a photo of a celebrity or the weather of the day. The noise was reduced by manually connecting only the text of the article to the meta data in the manually coded document. Words written with hyphens were managed manually and the hyphens were removed.

4.1.1 Class Selection

A pre-processing was done to find those classes that were meaningful to use in the experimental study. The code instructions were read thoroughly and each of the classes was studied to see if they were possible to use within the research. When looking for classes to use, some ambiguity was desired; It is in the situations where the assignment of class for a document is ambiguous, that explanations are most helpful.

In the data set, there were several classes that had an ambiguous character, and could cause documents to become incorrectly predicted in an automated classification, and of interest for the research study. In the coded document the classes were referred to by a number:

1 = Politics, 2 = Business/economics, 3 = Social issues (work, health, education, migration), 5 = Accidents/disasters, 6 = Culture, 7 = Entertainment, 9 = Science, technology, 10 = Media, 12 = Crime, 11 = Environment, 14 = War/conflict.

In the code instructions, the distinction between the classes above was described as follows: “This variable answers what MAIN theme of the news item is. If, for instance, a politician is involved in a car crash it is an “accident” news and not about “politics”. “Culture” refers to news about cultural objects per se such as literature, opera and theater (fine arts) or a discussion about society at large that relies on explicit intellectual and/or moral arguments. “Entertainment” refers to amusement such as TV-shows, computer games and a, largely, self-referring meta discourse about cultural objects or subjects such as celebrities, actors, and authors that does not focus on the content of the cultural object but on the performer or the discourse about the performer/cultural object.”

Because some of the classes had too few articles to be classified with a satisfying accuracy, a convenience sample was done out of two of the classes. The number of classes was chosen to ensure that the classifier had as high performance as possible, although the number of articles was low. There is sometimes a close connection between the number of classes and the performance of the classification, where an increasing number could decrease the performance (Gunnarsson, 2011, p. 227). The documents

---

2 If the text from the code instructions document is in Swedish, an English translation is written after the words within parentheses

3 It is necessary to ask for permission from the authors of the code instructions, to achieve them
within the chosen classes got the value of 1 or 2 depending on which class they belonged to (see table 4.1).

<table>
<thead>
<tr>
<th>Class</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics</td>
<td>1</td>
</tr>
<tr>
<td>Social issues</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 4.1: Class and number

### 4.2 Results

As seen in table 4.2, the documents within the data set differs in length (number of words/features). There are some documents that are only a few lines long and some that are several pages. Some are written as debate articles and some are merely informative. Although the maximal length of the texts in the two classes differs (705 to 1446) they are rather similar in minimal and mean length. After the tokenization (see chapter 2.3.1),

<table>
<thead>
<tr>
<th>Class</th>
<th>Maximal</th>
<th>Minimal</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics</td>
<td>705</td>
<td>27</td>
<td>184.5</td>
</tr>
<tr>
<td>Social issues</td>
<td>1,446</td>
<td>29</td>
<td>196.6</td>
</tr>
</tbody>
</table>

Table 4.2: Document length (number of words/features) within the classes in the data set

there were 29,867 features (see table 4.3) to work with. The stop-word list was applied, words that occurred less than 3 times in the document were excluded and the amount of features was reduced to 8,963. Although the number of documents is rather low, there was a considerable amount of features left after the stop word elimination, which is characteristic for this type of high dimensional data, making the classification complicated to explain and understand (Martens and Foster, 2014). In the original data set of 5,000 documents, only approximately 1,300 were digitally stored to be able to build a data set from. In the pre-processing of the original 1,300 documents, only 180 documents were matched to the two selected classes. 90 documents within each of the chosen classes. The text of the documents was stored and organized together with the extended meta data that was provided by the manually classified document.
Chapter 5

Implementation of the AITCE

5.1 Adaptation of Method

In the chapter of methodology the purpose of the experimental study was written as: to implement the explanation method AITCE, as mentioned in the literature chapter, to use with the results of an automated text classification. The study will contribute to fulfilling the main objective by presenting the features for each document, that most strongly affect the prediction and the most important features for each class. The results from the experimental study will be used as the in-data to the evaluations.

5.1.1 Experimental Design

The explanation coefficient (EC) of the features is calculated using the AITCE. When analyzing the text of a specific document in the data set, the EC for the features can be used to explain the prediction, by presenting to the user those features that most strongly affect the prediction.

5.1.2 Defining the Setup

The output from the implementation and the input to the evaluations, was divided into two sections:

1. The creation of the document representation of each document within the data set resulted in a matrix, that after stop word removal was used as an input to two classifiers in a pre-classification. The word occurrence was represented as in Forman (2003). Instead of a Boolean value (where each variable is assigned either true or false), a binary list (where each variable is assigned either 1 or 0) was used, to make calculations easier. Stemming was not used, since it did not make the accuracy rise notably, instead the explanation words became hard to understand. To reconnect the stem to the intended words, would be time consuming and uncertain
since different words could have the same stem in a text, but only one word be important.

The data set was pre-classified in the program KNIME to check which classifier should be used, Random forest (RF) or SVM, that provided the highest accuracy. Different classifier suits different data and a classifier that does not have an acceptable accuracy could result in a majority of insignificant words, when trying to extract the most important features. The two classifiers had a comparable accuracy, and they are both used for document classification. Since SVM could be complex, with many parameters to know how to tune to get the most out of the classifier (as seen in Chapter 2.3.4), the RF classifier was chosen.

2. The second part in the experimental study consisted of training the model, classifying the data set and applying the AITCE to the model. The AITCE was implemented as a function in Python. It takes as input the trained model (an object) and the document $x$. The RF classifier (with 100 trees) and the AITCE were implemented in Python. The results were achieved by evaluating using the Leave-one-out method repeated five times. The values presented in the thesis is, consequently, the average from the five repetitions. The output from the implementation resulted in three files, one from the classification and two from the AITCE:

- One file with a probability estimate (PE) for each of the classified articles (from the classification),
- One file with the explanation words for each article, with the EC for each explanation word (from the AITCE),
- One file with the explanation words for the classes (from the AITCE)

5.2 Results of the Experimental Study

5.2.1 Results from the Classification

<table>
<thead>
<tr>
<th>Status</th>
<th>Correct</th>
<th>Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total amount</td>
<td>131</td>
<td>47</td>
</tr>
<tr>
<td>PE above 0.60</td>
<td>93</td>
<td>17</td>
</tr>
<tr>
<td>PE between 0.40-0.60</td>
<td>39</td>
<td>26</td>
</tr>
<tr>
<td>Percent Unsure (%)</td>
<td>30</td>
<td>55</td>
</tr>
</tbody>
</table>

Table 5.1: Distribution of PE among the predictions
The classification reached an accuracy of 74%, which could be considered as relatively high\(^1\), considering the complexity and size of the data set. These so-called real world data sets may be inconsistent, noisy, and incomplete, which according to Kamruzzaman and Sarkar (2011, p. 4626) could disguise useful patterns.

Table 5.1 shows the distribution of correct and incorrect predicted documents in the data set. The majority of incorrect predictions was found to have a PE of 0.4-0.6. When the PE of a prediction in an automatic text classification lies around 0.5, it is said to be unsure and the prediction tends to be almost arbitrary. As seen in figure 5.1 26 of the total 47 incorrect predictions lied within the unsure interval, but only 39 of total 131 correct predictions.

![Figure 5.1: Distribution of documents within the interval 0.4-0.6 among correct and incorrect predictions.](image)

5.2.2 Article Selection Criteria

Six documents were chosen from the results of the classification, based on their PE from the table of classification results (see table 5.2). A higher amount of documents in the

---

\(^1\)In Ribeiro et al. (2016, p 3) 54% is seen as an acceptable level of accuracy
questionnaire would have been beneficial, but it would make it too time-consuming to answer and could lower the answering frequency, which could be seen in the frequency of answers from the respondents. Respondent 3 sent the response of the questionnaire 6 month later than requested and said it was partly due to the scope of the questionnaire. Below follows a description of the six documents.

The number of documents was chosen to satisfy the following requests:

1. Two documents (one for each class) where the prediction had a high PE, and was correct. When a prediction has a high PE, approaching 1, it indicates that the classifier has found strong evidence in the data for that prediction. This is the ideal situation, when the classifier produces a high PE for a document and the prediction is true. The goal with these examples was to investigate if the AITCE was able to catch the words that described the prediction. To reach this goal, it was necessary to find if the respondents recognized the significant words with the help of the AITCE and if they agreed with it.

2. Two instances (one for each class) where the prediction had a high PE, and was incorrect according to the human classification. This is not a desirable situation; the classifier has found strong evidence in the data for the prediction, but it is incorrect. The goal with these examples was to investigate if the respondents could, with the help of the AITCE, detect that the prediction was incorrect and the words that had caused the false prediction.

3. Two instances (one for each class), that had a PE about 0.5. When the PE approaches 0.5, the prediction becomes arbitrary, and the probability for a correct prediction drops to about 50%. The goal with these instances was to investigate if the AITCE could help the respondents to find why the prediction became uncertain, by detecting the insignificant words, and if they could find significant words.

<table>
<thead>
<tr>
<th>Instance ID</th>
<th>PE Politics</th>
<th>PE Social issues</th>
<th>Predicted class</th>
<th>True class</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>143</td>
<td>0.90</td>
<td>0.10</td>
<td>Politics</td>
<td>Politics</td>
<td>True sure</td>
</tr>
<tr>
<td>94</td>
<td>0.09</td>
<td>0.91</td>
<td>Social issues</td>
<td>Social issues</td>
<td>True sure</td>
</tr>
<tr>
<td>111</td>
<td>0.76</td>
<td>0.23</td>
<td>Politics</td>
<td>Social issues</td>
<td>False sure</td>
</tr>
<tr>
<td>18</td>
<td>0.19</td>
<td>0.81</td>
<td>Social issues</td>
<td>Politics</td>
<td>False sure</td>
</tr>
<tr>
<td>90</td>
<td>0.48</td>
<td>0.52</td>
<td>Social issues</td>
<td>Social issues</td>
<td>Unsure</td>
</tr>
<tr>
<td>50</td>
<td>0.48</td>
<td>0.52</td>
<td>Social issues</td>
<td>Politics</td>
<td>Unsure</td>
</tr>
</tbody>
</table>

Table 5.2: Selection of articles for evaluation.
<table>
<thead>
<tr>
<th>Feature</th>
<th>EC for class Politics</th>
<th>EC for class Social issues</th>
<th>PE for class Politics if deleted</th>
<th>PE for class Social issues if deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>partiet</td>
<td>0.042</td>
<td>-0.042</td>
<td>0.860</td>
<td>0.140</td>
</tr>
<tr>
<td>parti</td>
<td>0.036</td>
<td>-0.036</td>
<td>0.866</td>
<td>0.124</td>
</tr>
<tr>
<td>socialdemokrat</td>
<td>0.024</td>
<td>-0.024</td>
<td>0.878</td>
<td>0.124</td>
</tr>
<tr>
<td>röst</td>
<td>0.024</td>
<td>-0.024</td>
<td>0.878</td>
<td>0.124</td>
</tr>
<tr>
<td>sverigedemokrat</td>
<td>0.024</td>
<td>-0.024</td>
<td>0.878</td>
<td>0.124</td>
</tr>
<tr>
<td>riksdag</td>
<td>0.018</td>
<td>-0.018</td>
<td>0.884</td>
<td>0.116</td>
</tr>
<tr>
<td>EU-valet</td>
<td>0.16</td>
<td>-0.16</td>
<td>0.886</td>
<td>0.114</td>
</tr>
<tr>
<td>politik</td>
<td>0.14</td>
<td>-0.14</td>
<td>0.888</td>
<td>0.112</td>
</tr>
<tr>
<td>landsting</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.912</td>
<td>0.088</td>
</tr>
<tr>
<td>tapp</td>
<td>-0.008</td>
<td>0.008</td>
<td>0.91</td>
<td>0.09</td>
</tr>
<tr>
<td>heta</td>
<td>-0.008</td>
<td>0.008</td>
<td>0.91</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table 5.3: Feature EC, Article 1

**Article 1**

The first article, with instance ID 143 (see table 5.3), was correctly predicted with a 0.9 PE to belong to the class Politics. It is a rather short text with a clearly political message; the Swedish party ‘C’ (short version of the Swedish word ‘Centerpartiet’, the Center party) has lost votes in a town called Sunne in the EU election.

The AITCE found several words that pointed to the class Politics: parti/partiet (the party), socialdemokrat (Social democrat, a Swedish political party), EU-valet (EU election), sverigedemokrat (Swedish democrat, a Swedish political party), riksdag (parliament), röster (votes) and politik (politics).

Three words were found to indicate the other class Social issues: landsting (county council), tapp (loss), heta (heated).

**Article 2**

The second article, with instance ID 94 (see table 5.4), is also a rather short text, correctly predicted with a 0.91 PE to belong to the class Social issues. The article refers to how the (labour union) LO is informing youths about their rights when working during summer holidays. It has a clear message about working rights, and youths.

The AITCE found the following words that pointed to the class Social issues: året (the year), jobb/jobbet (job/the job), träffa (meet), vecka (week), illa (bad), villkor.
(terms), arbetsmarknad (employment market).
Two words were found to indicate the class Politics: pressmeddelande (press release) and hoppas (hope).

<table>
<thead>
<tr>
<th>Feature</th>
<th>EC for class Politics</th>
<th>EC for class Social issues</th>
<th>PE for class Politics if deleted</th>
<th>PE for class Social issues if deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>året</td>
<td>-0.04</td>
<td>0.04</td>
<td>0.13</td>
<td>0.87</td>
</tr>
<tr>
<td>träff</td>
<td>-0.022</td>
<td>0.022</td>
<td>0.112</td>
<td>0.888</td>
</tr>
<tr>
<td>jobb</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.11</td>
<td>0.89</td>
</tr>
<tr>
<td>vecka</td>
<td>-0.018</td>
<td>0.018</td>
<td>0.108</td>
<td>0.892</td>
</tr>
<tr>
<td>illa</td>
<td>-0.018</td>
<td>0.018</td>
<td>0.108</td>
<td>0.892</td>
</tr>
<tr>
<td>villkor</td>
<td>-0.012</td>
<td>0.012</td>
<td>0.102</td>
<td>0.892</td>
</tr>
<tr>
<td>jobbet</td>
<td>-0.008</td>
<td>0.008</td>
<td>0.098</td>
<td>0.902</td>
</tr>
<tr>
<td>arbetsmarknad</td>
<td>-0.008</td>
<td>0.008</td>
<td>0.098</td>
<td>0.902</td>
</tr>
<tr>
<td>pressmeddelande</td>
<td>0.006</td>
<td>-0.006</td>
<td>0.084</td>
<td>0.916</td>
</tr>
<tr>
<td>hoppas</td>
<td>0.006</td>
<td>-0.006</td>
<td>0.084</td>
<td>0.916</td>
</tr>
</tbody>
</table>

Table 5.4: Feature EC, Article 2

Article 3

The third article, with instance ID 111 (see table 5.5), was incorrectly predicted with a 0.76 PE to belong to the class Politics. It is an article covering a broad range of topics, from the demolition of a beggar shantytown to a referendum in Switzerland.

The AITCE found several words that pointed to the class Politics: regeringen (government), politik (politics), rösta (to vote), Moderaterna (a Swedish political party), välfärden (welfare), Schweiz (Switzerland), franska (french).
Three words were found to indicate the class Social issues: exempelvis (for example), sällsynt (rare) and bär (carries).

Article 4

The fourth article, with instance ID 18 (see table 5.6), is very short, not more than a few lines. It was incorrectly predicted with a 0.81 PE to belong to the class Social issues. The article covers an opening ceremony of a bus stop (together with a walkway and a cycle way) in a town called Vålberg. It is written by a political elected person, which is indicated by a letter within parentheses after the name of the author of the article.
Table 5.5: Feature EC, Article 3

The AITCE found several words that pointed to the class Social issues: satsning (investment), konstatera (note), ort (district), alldeles (entirely), gång- (walk), kommunalrådet (local government commissioner), land (part of the last name of the author, also meaning country). No words were found to indicate the other class, Politics.

Table 5.6: Feature EC, Article 4
Article 5

The fifth article, with instance ID 90 (see table 5.7), was predicted with a 0.52 PE to belong to the class Social issues. The article concerns how a municipality bought 200 used iPads without going through the required procedures.

The AITCE found several words that pointed to the class Social issues: egentligen (actually), händelse (event), väga (weigh), färre (fewer), uppgift (task).

Several words were found to indicate the class Politics:
Timrå (a Swedish town), kvalitet (quality), avslöja (reveal), lagen (the law), Eriksson (the last name of the boss in charge), syfte (purpose), upphandling (procurement), lämna (leave), kommun (municipality).

<table>
<thead>
<tr>
<th>Feature</th>
<th>EC for class Politics</th>
<th>EC for class Social issues</th>
<th>PE for class Politics if deleted</th>
<th>PE for class Social issues if deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>egentligen</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.498</td>
<td>0.502</td>
</tr>
<tr>
<td>händelse</td>
<td>-0.018</td>
<td>0.018</td>
<td>0.496</td>
<td>0.504</td>
</tr>
<tr>
<td>färre</td>
<td>-0.012</td>
<td>0.012</td>
<td>0.49</td>
<td>0.51</td>
</tr>
<tr>
<td>väga</td>
<td>-0.012</td>
<td>0.012</td>
<td>0.49</td>
<td>0.51</td>
</tr>
<tr>
<td>uppgift</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.488</td>
<td>0.512</td>
</tr>
<tr>
<td>Timrå</td>
<td>0.038</td>
<td>-0.038</td>
<td>0.44</td>
<td>0.56</td>
</tr>
<tr>
<td>kvalitet</td>
<td>0.022</td>
<td>-0.022</td>
<td>0.456</td>
<td>0.544</td>
</tr>
<tr>
<td>avslöja</td>
<td>0.014</td>
<td>-0.014</td>
<td>0.464</td>
<td>0.536</td>
</tr>
<tr>
<td>lagen</td>
<td>0.014</td>
<td>-0.014</td>
<td>0.464</td>
<td>0.536</td>
</tr>
<tr>
<td>Eriksson</td>
<td>0.014</td>
<td>-0.014</td>
<td>0.464</td>
<td>0.536</td>
</tr>
<tr>
<td>syfte</td>
<td>0.012</td>
<td>-0.012</td>
<td>0.466</td>
<td>0.534</td>
</tr>
<tr>
<td>upphandling</td>
<td>0.012</td>
<td>-0.012</td>
<td>0.466</td>
<td>0.534</td>
</tr>
<tr>
<td>lämna</td>
<td>0.012</td>
<td>-0.012</td>
<td>0.466</td>
<td>0.534</td>
</tr>
<tr>
<td>kommun</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.468</td>
<td>0.532</td>
</tr>
</tbody>
</table>

Table 5.7: Feature EC, Article 5

Article 6

The sixth article, with instance ID 50 (see table 5.8), was predicted with a 0.52 PE to belong to the class Social issues. The article covers how a group of parents was trying to hinder a sports association from preventing the construction of a local sports hall.
The AITCE found several words that pointed to the class Social issues: samt (as well as), informera (inform), barn (children), lokaltidning (local newspaper), ungdöm (youth).
Several words were found to indicate the class Politics: kommun (municipality), omröstning (recall), förslag (suggestion), planer (plans), ytterligare (further), klubb (club), veckan (the week), grupp (group).

<table>
<thead>
<tr>
<th>Feature</th>
<th>EC for class Politics</th>
<th>EC for class Social issues</th>
<th>PE for class Politics if deleted</th>
<th>PE for class Social issues if deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>kommun</td>
<td>0.098</td>
<td>-0.098</td>
<td>0.382</td>
<td>0.618</td>
</tr>
<tr>
<td>omröstning</td>
<td>0.036</td>
<td>-0.036</td>
<td>0.444</td>
<td>0.556</td>
</tr>
<tr>
<td>förslag</td>
<td>0.032</td>
<td>-0.032</td>
<td>0.448</td>
<td>0.552</td>
</tr>
<tr>
<td>planer</td>
<td>0.03</td>
<td>-0.03</td>
<td>0.45</td>
<td>0.55</td>
</tr>
<tr>
<td>ytterligare</td>
<td>0.012</td>
<td>-0.012</td>
<td>0.468</td>
<td>0.532</td>
</tr>
<tr>
<td>klubb</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.47</td>
<td>0.53</td>
</tr>
<tr>
<td>veckan</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.47</td>
<td>0.53</td>
</tr>
<tr>
<td>grupp</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.47</td>
<td>0.53</td>
</tr>
<tr>
<td>samt</td>
<td>-0.012</td>
<td>0.012</td>
<td>0.492</td>
<td>0.508</td>
</tr>
<tr>
<td>informera</td>
<td>-0.008</td>
<td>0.008</td>
<td>0.488</td>
<td>0.512</td>
</tr>
<tr>
<td>barn</td>
<td>-0.008</td>
<td>0.008</td>
<td>0.488</td>
<td>0.512</td>
</tr>
<tr>
<td>lokaltidning</td>
<td>-0.006</td>
<td>0.006</td>
<td>0.486</td>
<td>0.514</td>
</tr>
<tr>
<td>ungdöm</td>
<td>-0.006</td>
<td>0.006</td>
<td>0.486</td>
<td>0.514</td>
</tr>
</tbody>
</table>

Table 5.8: Feature EC, Article 6
Chapter 6

Evaluation of the AITCE

6.1 Adaptation of method

The analysis of AITCE was built on a self-completion questionnaire, given to three senior researchers that have a deep knowledge in the data set used in the research, and a web based survey given to non-experts in a Facebook group. The researchers were presented with the results of the classification together with the results of the AITCE, and asked to give feedback by answering the questions sent to them via e-mail, while the web survey was shared as a link in the Facebook group.

6.1.1 Formulation of questionnaire

The self-completion questionnaire differs from the situation of interview in a fundamental way: there is no interviewer to ask the questions. Instead the respondents must read the questions and answer them themselves. Bryman (2012, p. 233) emphasizes that this requires the questionnaire to be especially easy to understand and follow, just as the questions have to be very easy to answer.

The challenge lies in the task to ask few, simple questions, but still catching the respondents view of the AITCE. Bryman (2012, p. 233) also writes that it is important to only ask a very small amount of open questions, since the respondents tend to not want to write a lot. In the questionnaire, all the questions are open, but there are not many questions asked and they are formulated not to require much writing.

The questionnaire consisted of two sections: first a short presentation of the results from the AITCE with questions and secondly two different suggestions of how to present the results from the AITCE. The questions of the first section of the questionnaire focused on if the AITCE could produce insights to the predictions, if the respondents were able to use the explanation words and understand why the automatic classification had predicted a certain class.
Martens and Foster (2014, p. 17) write that some kind of relevant feedback or interactive solutions with the users could enhance the quality PE close to 0.5, were chosen to see if the respondents could detect both the significant words in the table with explanation words, as well as the insignificant words, and give suggestions of significant words from the text that could increase the quality of the classification.

All the respondents were familiar with the data set used in the research, since they had been working with the project from which the data was taken. For this reason, it was not necessary to introduce the context of the data.

To make the respondents more familiar with the concept and the look of the result of an automatic classification, figures of the results both from the human classification and the automated classification (see figure 6.1) were placed beside each other. The results of the automatic classification is to the left. This served as an introduction, together with a short text about the research itself. The introduction was followed by a presentation of the six chosen articles. The prediction as well as the PE of each article was presented to the respondents, but it was not revealed if the prediction was correct. The users of automatic classification could have a tendency to misuse and knowing the correct prediction could cause the respondents to agree out of misuse. It was assumed that because the respondents knew the data well, they would be more inclined to question the incorrect predictions.

### 6.1.2 Design of Presentation

The second part of the questionnaire consisted of a suggestion of user interface. It was designed to enhance the perception of which words that have been most important for the prediction. The presentation focused on putting the explanation words into their context. The respondents were presented with the full text of the article as well as the explanation words. The words were highlighted in different colours in the text as seen in figure 6.2, and ordered in the list according to their importance for the prediction.

In the second version the explanation words were presented to the respondents in the form of a word cloud. Word clouds have been used for many years and have been very popular, although their actual importance for perception has been discussed. In the second version of the presentation the explanation words were presented in a word cloud.

<table>
<thead>
<tr>
<th>Politik</th>
<th>Samhälle</th>
<th>The prediction</th>
<th>The True value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.72</td>
<td>0.28</td>
<td>Politik</td>
<td>Politik</td>
</tr>
<tr>
<td>0.33</td>
<td>0.67</td>
<td>Samhälle</td>
<td>Samhälle</td>
</tr>
</tbody>
</table>

Figure 6.1: The two types of classification.

<table>
<thead>
<tr>
<th>Politik</th>
<th>Samhälle</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Each word was given a size corresponding to how strongly it affected the prediction (see figure 6.3).

### 6.1.3 The Questions

The questions were formulated with the intention of gradually introducing the respondent to the results of the AITCE and to be able to fulfill the goals formulated in chapter 5.2.2. For convenience, and to avoid translations, both questions and answers were formulated in English.

**Questions for Articles 1 to 4**

- When looking at the words in combination with the estimated certainty of the prediction (in parentheses), do you think the prediction is correct? Why/why not?

- Read the text below and look at the words again. Do you understand why the text has been assigned into its respective class? Why/why not?

- Would you yourself map the text into its respective class? Why/why not?
Figure 6.3: The second type of presentation

- According to you, are there any words in the table above that point to the other class, than the predicted one?

Questions for Articles 5 and 6

- When looking at the words in combination with the estimated certainty (in parentheses), would it be possible to categorize the document, based on the words above? Why/why not?

- Read the text below and look at the words again. Do you agree that the words characterize the text? Why/why not?

- Would you yourself assign the text into its respective class? Why/why not?

- Would the removal of one or more words from the list help to categorize the document?

Questions for the Presentation

- Which one of the versions do you prefer? Why?

- Would a mix of both of them make the explanation better? Why/why not?
6.1.4 Construction of the Web Survey

The web survey was introduced to the respondents by a short text describing the purpose of the survey and how to address the questions. The fourteen words from each class with the highest EC were selected from the results of the AITCE. The respondents were asked to choose which class they thought was best suited for each word. The words were presented at random from the two classes, in a way that the respondents couldn’t know to which class the words belonged to.

The respondents were invited from a Facebook group of adults with 524 members, and 24 persons answered the web survey. The respondents began a discussion on Facebook about the web survey where the purpose and the thought behind the questions were discussed with the author.

Class Politics

The class Politics was characterized by the following words with the highest EC: regeringen (the government), USA (the country USA), partiet (the party), president (president), valet (the election), kommunstyrelsen (municipal government), parti (party), Japan (the country Japan), ljus (light), politiken (the politic), SD (short name for a political party, Sverigedemokraterna), pågå (to last), användas (be in use), sverigedemokraterna (a political party).

Class Social issues

The class Social issues was characterized by the following words with the highest EC: kritik (critique), varje (every), problemen (the problems), skolan (the school), allra (very), öppnar (open up), omkring (around), verksamheten (the business), skapa (create), bo (to live), hända (happen), förändra (change), nationella (national), samhälle (society).

6.2 Results of the self completion questionnaire

Two articles that were chosen for the questionnaire were correctly predicted, article 1 and 2, and two articles that were chosen for the questionnaire were incorrectly predicted; article 3 and 4. Articles 5 and 6 had an estimate coefficient of around 0.5 which is at a level where the prediction has more of an arbitrary character, making it more or less meaningless to categorize them as correct or incorrect. Instead the were categorized as unsure.
6.2.1 Correctly predicted articles

Article 1

The respondents had no problems agreeing with the prediction of the article, based on what they said were clearly political words and since it, according to the respondents, concerns formal politics and governance. Although respondent 3 wrote that considering the words not taken account for that point to the other class and the number of insignificant words, the probability estimate for the article should have been lower, around 60%. Asked if they understood why the text had been predicted to belong to its respective class, all but respondent 3 did. Both Respondent 1 and 3 pointed out the clearly political message of the text, focusing on politics and polls, which indicated the class Politics.

When asked if there were any words that pointed to another class than the predicted one, respondent 1 was unsure about why the word landsting (county council) pointed to Social issues rather than Politics. Respondent 2 answered that the word pointed to the formal dimension of politics, but it was understood that the word is "somewhat problematic to separate from other social issues (such as being in contact with landstinget as a patient)". Respondent 3 wrote that the words "tapp" (loss) and "heta" (heated) were neither pro Politic or pro Social issues and perceived as rather arbitrary chosen. Respondent 3 also mentioned five words that indicated the class Social issues, but had not been pinpointed as important: Värmlandstrafik (a company), Sunne (the name of the municipality), ensamkommande barn (unaccompanied minors), flyktingar (refugees) and skolstruktur (the structural school organization).

Article 2

Respondent 1 was not totally convinced of the prediction, but was able to see certain parallels between some of the words and the classes they had been predicted to belong to. Respondent 2 wrote that the text touched on broader social issues as work rights, but couldn’t see that it "involves formal political bodies”. Respondent 1 wrote that the words jobb (job), villkor (terms) and arbetsmarknad (employment market) pointed to Social issues. Respondent 3 wrote that several of the explanation words were a bit arbitrarily chosen, and mentioned more important words from the text: arbetsvillkor (terms of employment) rättigheter (rights) and skyldigheter (liabilities). When asked if they perceived the prediction as correct, all of the respondents agreed.

When asked if there were any words that pointed to another class than the one predicted, respondent 1 had no example, but wrote that some of the words “confuse me by simply being mapped to either category”, since they were perceived as not belonging to either class. Respondent 2 wrote that the word pressmeddelande (press release) could not be solely connected to politics since "many other actors utilise them as well". Respondent 3 did not mentioned any words from the table, but pointed at two words in the text: LO (a federation of labor unions) and kampanj (campaign).
6.2.2 Incorrectly predicted articles

Article 3

The respondent 1 as well as respondent 2 agreed with the prediction, but not respondent 3. Respondent 1 wrote that it was due to the focus on "the popular vote and putting new laws into effect", but pointed out that some words (EU and EES) that indicated the class Politics did not exist in the list of explanation words. Respondent 2 agreed to the prediction since there were aspects of formal politics involved, mentioning law-making and political agreement, as the EES-deal. Respondent 3 answered that the text had a more societal focus, where social problems were discussed and "not politics per se".

Respondents 1 and 2 did not see any words that pointed to the other class, nevertheless respondent 1 remarked that there were words that the respondent found confusing that they had been predicted to either class. Respondent 2 would probably had classified the article in line with the automatic classifier, but was not entirely convinced, since this kind of text is “hard to pin down since it is long and wanders in focus”. Respondent 3 again referred to the text and found many words that pointed to Social issues, that were not to be found in the table with the explanation words: omvärlden (the surrounding world), käkstad (shanty town), etniskt (ethnical), abort (abortion), sexuella övergrepp (sexual harrasments) and tvångsvräkningar (forced evictions).

Article 4

About this article, respondent 1 wrote that none of the words signified that the article belonged to any of the classes and the prediction was not understood. Two words that could be an exception were the words satsning (investment) and kommunalråd (local government commissioner). Where the latter pointed to the class Politics. Respondent 1 and respondent 3 agreed with the prediction, writing that the text was referring to a social relevant fact, but respondent 3 found the explanation words were arbitrarily chosen.

The respondent also remarked that the article would be difficult to classify, since a politician made comments about the development. Respondent 2 was not convinced about the prediction, since the word kommunalråd (local government commissioner) was involved, and pointed to the class Politics. But since the indication on the class was so discreetly written, in the end of the article, respondent 2 understood why it was downplayed.

6.2.3 Uncertain articles

Article 5

According to respondent 1, the explanation words did not really point to either class, and respondent 3 wrote that more of the explanation words point to the class Politics than Social issues. Respondent 1 listed three words that had some meaning: lagen (the law),
upphandling (procurement) and kommun (municipality) but wrote that none of the two last words made the text more Pro Politics than Pro Social issues. Respondent 2 on the other hand wrote that it was a rather clear case of Politics due to the focus on the word upphandling (procurement), the potential breaking of law and references to the words kommunens skolor (the schools of the municipality) as well as förvaltningschef (administrative director), which according to respondent 2 indicated formal political structures.

When asked if the words characterized the text, respondent 1 answered that some words were central to the text, but it did not help with the classification, and other words that would help were missing in the table; such as förvaltningschef (administrative director), utbildningsförvaltningen (education administration) and omorganiserings (reorganisation). Respondent 3 answered that other words should have been chosen to characterize the text and suggested the words: tjänstemän (civil servant) and utbildningsinsats (education effort) to indicate Social issues. Respondent 3 also wrote that important words i.e. skattebetalarna (taxpayers) were omitted in the table, indicating the class Politics.

When asked if the removal of one or more words could help with the classification, both respondent 1 and 3 wrote that it would be hard to classify, but mentioned several insignificant words: händelse (incident), färre (few), väga (weigh), kvalitet (quality), Eriksson (a name, Eriksson) or lämna (leave). Respondent 2 also questioned the value of the word Eriksson.

**Article 6**

When asked if the words in combination with the estimated certainty would make it possible to classify the document, respondent 1 wrote that the words kommunstyrelsen (municipal executive board) in combination with omröstning (vote) and förslag (suggestion) signified the Pro Politics class, and that the words within the Pro Social issues frame did not. Respondent 2 also agreed with the prediction indicated by the probability estimate, writing that it was because the emphasis on the word kommunstyrelsen (municipal executive board) and that the parents would fight its ruling. Respondent 3 wrote that the explanation words to some extent characterized the text, but important words indicating the class Social issues were omitted and others seemed rather arbitrarily chosen, making it hard to classify the text based on the explanation words.

Respondent 1 agreed that the most important words had been selected and that they characterized the text, but wrote that the words samt (as well as), ytterligare (additional), klubb (club) or veckan (the week) were insignificant to either class. Respondents 2 and 3 mentioned almost the same words as respondent 1, but respondent 3 also added the word informera (inform) and respondent 2 had some thoughts about the word omröstning (vote).
6.2.4 Presentation

When asked about the presentation, respondent 1 and respondent 3 preferred the first version and not a mix of them. Respondent 1 wrote that the first version made a clearer case, when seeing the whole text as well as the explanation words, and that the second version made the explanation words pointing at the class Pro Social issues looked more significant than they were in the context. Word clouds seemed to be of a not entirely positive character. Respondent 1 expressed mental fatigue and wrote "I am bored to death by word clouds".

Respondent 2 also preferred the first version of the presentation since it provided a context that made it easier to understand, but did not share the dislike of the word clouds. Instead, respondent 2 liked version two for a quick overview.

Both respondents 1 and 2 found the study highly interesting and when they were asked for suggestions for changes that could make the tool more useful, none of them could think of anything special. At the same time, all of the respondents raised questions through the questionnaire about the insignificant words and the selection of the explanation words. Respondent 1 expressed this in the answer to the last question; "I would be highly interested in seeing the documentation and logic behind the choice of certain words how the tool relates them to the different frames".

6.3 Results of the web survey

The Facebook group where the link was shared had at the specific date 524 members, and 24 answered the survey. According to Bryman (2012) there is growing evidence that this kind of online surveys generate lower response rates than postal surveys. As the members of Facebook groups fluctuate over time and not all of the members may have been online when the link was shared, it is also very difficult to calculate a precise response rate.

The majority, 71.43%, of the words from the class Politics were answered to belong to the class. 28.57% of the words in the class were answered to belong to the class Social issues, which corresponds to four words: Japan (the country Japan), ljus (light), pågå (to last) and användas (is being used). The words with the highest estimate coefficient were those that most respondents answered as belonging to the class Politics (see figure 6.4). They were clearly political words as regeringen (the government), partiet (the party) and valet (the election). Although the word USA was seen as belonging to the class Politics, the word Japan was seen as belonging to the class Social issues. The other three words that were considered to not belong to the class are all words that have no obvious connection to the concept Politics: ljus (light), pågå (to last) and användas (is being used).

The majority, 85.71%, of the explanation words in the class Social issues were answered to belong to the class, although not as distinctly as in the class Politics. Only
14.29% of the words were answered to belong to the other class, which corresponds to three words: kritik (critique) förändra (change) and nationella (national). Although, the word förändra (change) had a close to even distribution of answers, where 55% answered it to belong to Politics and 44% to Social issues. The distribution, close to 50-50, is also seen in several other words as seen in figure 6.5.

Although a higher percentage of the explanation words in the class Social issues were answered to belong to the class than in the class Politics (71.43% to 85.71%), the mean values of the explanation words from the different classes were similar. The mean values of how many respondents that answered that the explanation words belonged to the classes were 68% for the class Politics, and 70% for the class Social issues.
Figure 6.5: Results of class Social issues
Chapter 7

Concluding Discussion

7.1 Analysis and reflections

In the chapter of problem formulation a main objective was formulated:
Develop and evaluate an explanation method which can provide researchers with insights about causes of the predictions and provide decision support about documents requiring human classification.

Three sub-objectives were derived from the main objective and the work with the thesis was divided according to them; 1. where a suitable data set was constructed, 2. where an explanation method (AITCE) was implemented and 3. where the results from AITCE were evaluated. The evaluations (sub-objective 3), that investigated if the AITCE could provide researchers with insights about the predictions, were based on the results of the AITCE (sub-objective 2) that used the output from a RF classifier, applied on the constructed data set, using data from an existing research project (sub-objective 1). To investigate how AITCE could provide decision support with insights about documents requiring human classification, the predictive results from the experimental study (sub-objective 2) were analyzed. By fulfilling the three sub-objectives, the main objective is considered to have been accomplished.

7.1.1 The data set

The data set used in the thesis originates from an existing research project, where a large amount of money and time has been spent to collect and classify the data, with specific instructions on how to work. In this thesis the data set has been further developed, to make it possible to be applied together with an automatic text classifier. This type of data set is not easy to classify, which could be seen in the results from the AITCE, where insignificant words appeared even among the explanation words with the highest EC.

The length of the documents ranges between a few lines to several pages. Some of them are strictly informative and others wander in subject, touching on different areas,
which according to the responding researchers make them hard to classify. The classes are not easy to distinguish from each other, which sometimes makes it difficult to classify the documents. This situation occurred several times in the self-completion questionnaire, when the respondents were of different opinions of how a document should be classified, and found arguments in the text for their respective opinion.

The small amount of documents in the data set probably also partly caused the classifier not to find all the characteristics of the documents, causing the AITCE to present insignificant explanation words. Automatic classification is likely to work better with larger data sets, since it increases the possibility for the model to detect the relevant patterns within the documents during the learning phase.

As Kamruzzaman and Sarkar (2011, p.4626) point out, the data mining algorithms are assumed to be "nicely distributed, containing no missing or incorrect values where all features are important. The real world data may be incomplete, noisy, and inconsistent, which can disguise useful patterns". It is this kind of ambiguous real world data that researchers work with, which made it important to use within the study. To be able to fulfill the main objective, to evaluate if the AITCE could provide researchers with insights about causes of the predictions of an automatic text classification, it was important to get as close to the reality as possible.

7.1.2 The implementation

The results from the AITCE mirrored the results of the classification. Where the PE of the documents was lower, the amount of insignificant explanation words increased. The amount of insignificant words also seen in documents with higher PE could partly depend on the rather small data set it was trained on. If a larger data set had been used, the amount of insignificant words might have decreased.

The explanation words of the unsure predictions indicate that the classifier was not able to detect the characteristic of the documents. The PE of the documents did not rise or sink noticeably if a single explanation word was deleted, but fluctuated around 0.5. Only one explanation word made the PE rise above 0.6 for one of the documents, the word kommun (municipality). The deletion of the word would cause the prediction of the class Social issues to get stronger (0.618) and made the prediction of the class Politics sink to a PE of around 0.4 (0.382), which indicates that the classifier identified the word as belonging to the class Politics. Although, the PE of the document would still just slightly rise above the unsure interval (between 0.-0.6). It would require the deletion of several words, or the addition of one or more class-significant words, to make the document predictions less arbitrary.
7.1.3 The Evaluation

The Self Completion Questionnaire

The self completion questionnaire was based on only six documents, since a higher number of documents was expected to make the questionnaire too time consuming for the respondents. Only two of the six documents were correctly predicted, while two documents were unsure and two were clearly incorrect. This made the questionnaire complex to evaluate, since the respondents in the majority of cases were exposed to an incorrect prediction and many insignificant explanation words. The two clearly incorrect documents that were chosen were in fact the two documents in the data set that had a highest PE among those incorrectly predicted. The respondents seemed confused over the amount of insignificant explanation words, but still used both the explanation words and the text to answer the questions.

The data set is not easy to classify according to the respondents, which could be seen in the arguments the researchers presented to support their choice of class. They sometimes presented arguments that contradicted each other, although they have worked with the data and been involved in writing the code instructions for the human coders classifying the documents.

When the PE approaches 0.5, the prediction becomes arbitrary. The respondents wrote that these documents were hard to classify and sometimes wrote arguments for a different class than what the classifier predicted, but ended up in most cases to agree with the prediction that the PE indicated as slightly more likely.

All of the respondents liked the first version of presentation best. They wanted the quick overview from the explanation words, but they also wanted to be able to see the whole text with the words highlighted, to understand the context, which could indicate that they wanted to be able to check the classifier’s prediction and the explanation words against the full text.

Considering that only one third of the documents in the questionnaire were correctly predicted, that one third of the documents came from unsure predictions and that the accuracy of the entire model was around 74%, it is clear that the responses from the respondents indicate a more negative picture of the AITCE and that the distribution of the different categories of predictions is important to consider in the analyze of the results. The intention was to evaluate the AITCE in an unfavourable situation, but it is important to realize that is does not mirror the results of the entire data set.

The Web survey

A clear majority (71.43% Politics and 85.71% Social issues) of the respondents in the web survey recognized which class the words belonged to, although the order in which the words were presented were randomly chosen and no information of the classes was presented.
- **The Class Politics**
The respondents easily identified the first 7 words with the highest EC in the class. They are all, but one, distinct political words which could explain the result. The last 7 words had a more even result. Two words that are interesting are USA and Japan, where USA was answered to belong to Politics and Japan Social issues, although both of them are names of countries. The difference could be due to that since the latest presidential election in the USA, there have been much political discussions concerning the country in the newspapers. And in that way the name of the country, USA, may be perceived as more political than the other country, Japan. The other three words that most respondents identified as belonging to the class Social issues instead of Politics, are all rather neutral words, and it is possible that that is the reason they were perceived as more related to Social issues.

- **The Class Social issues**
The results from the class Social issues are more uniform than from the class Politics. The words with the highest EC are not considered by more respondents to belong to the class than those with the lowest value. In fact, the word with the lowest EC is considered by more respondents as belonging to the class, than the word with the second highest value. Only three words are marked as not belonging to the class. One of the words that by the majority of the respondents was considered as not belonging to the class, was the word kritik (critique) and the word with the highest EC in the class description. It is not impossible that this word indeed represents the class Social issues, if many critical articles in the subject Social issues were published at the time when the newspapers were written. Just as the case with the names of the two countries USA and Japan, the movement of subjects of discussions in society since the newspapers were published may play a not entirely negligible role.

7.2 Conclusion

The focus has been on analyzing if AITCE could provide researchers with insights about the causes of the predictions of documents and if it could serve as decision support about documents requiring human classification. To be able to evaluate the results from AITCE, a data set from an existing research project was constructed. Two evaluations were done; one at instance-level with senior researchers and one with non experts at model-level. The evaluations indicate that AITCE is able to provide the researchers with insights about the predictions, but a more extensive research with a more comprehensive data set is required to verify the findings.

Since it was not possible to have a higher amount of documents in the questionnaire, it was important to evaluate the results of AITCE in a non-favourable situation; To see if the explanations for the predictions were able provide the researchers with insights about
the theme, even in the most difficult situations. The data set was complicated, with documents that strongly differed in length, often touching on several different topics. The two classes that were chosen were also not easy to distinguish from each other. In the questionnaire, the respondents were presented with a majority of incorrect predictions. The AITCE was in other words evaluated in a very difficult situation, where the respondents of the self-completion questionnaire were exposed to a quantity of insignificant explanation words. However, the respondents used the explanation words in their analysis of the documents, to either questioning or agreeing with the predictions the PE indicated. E.g., in the documents with clearly incorrect predictions (number 3 and 4), the explanation words highlighted that the prediction was influenced by insignificant words, which the respondents realized and they questioned the predictions. When taking into account the proportions of correct, incorrect and unsure predictions, it is obvious that both the unsure and the incorrect predicted documents got a too prominent position in the questionnaire. To evaluate how the AITCE performed with the entire data set, an evaluation at model level, evaluating how well the AITCE could generalize the explanation words of the entire data set to explain the characteristics of the classes, was performed using a web survey. Although the respondents couldn’t trace from which class the explanation words originated, a clear majority of the respondents recognized which class the words belonged to. When regarding the distribution of the different categories of predictions, where the correct predictions are nearly 3 times more common than the incorrect, and the results of the web survey, the indications that the AITCE could provide researchers with insights about the causes of the predictions was strengthened.

When the predictions in the questionnaire were unsure, the explanation words had a very low EC. The respondents named the most insignificant words, and pointed out the significant. Even if all of the insignificant word where deleted and the significant ones from the table were marked as important, the prediction did not change noticeable, indicating that this type of document requires human expertise. The indication was further strengthened by the distribution of incorrect predicted documents, where the majority where found to belong in the unsure interval (PE between 0,4-0,6).

The conclusion is that the sub-objectives have been fulfilled and by this also the main objective. There are some uncertainty regarding how the evaluations have been affected by the small amount of documents in the data set and the low number of respondents. These questions are discussed further below.

7.3 Repeatability

The data set used in the thesis originates from an existing research project and has been further developed in the thesis to suit the research. With a comparable data set, although it is not easy to find, it is probable that similar results may be obtained.

Since an explanation method makes the automatic classification transparent, the result of the AITCE depend on the results of the classification. The AITCE is general and
works with different classifiers and the source code can be provided if asked for.

7.4 Sources of errors

7.4.1 The Data set

A larger data set might have strengthened the training phase of the classification, making it easier to find words that characterize the documents, creating a higher degree of certainty in the results of the AITCE. However, the choice of classes probably also influenced the accuracy. More distinct classes would likely have increased the accuracy, since it would have made the classification of the individual documents easier. In text classification there is a connection between larger data sets and a better, more fine-tuned, result. In the classification process the patterns within the data set generalizes to be applied to new, unseen, data. When the classifier is trained on a larger data set, the patterns are easier to find, which results in the finding of less insignificant words, which would likely strengthen the explanations produced by the AITCE.

7.4.2 The number of respondents

It is common that the evaluations in this type of research are constructed by the researchers through simulations, rather than reaching out to human respondents. When human respondents are used, they are often non-professionals; students or recruited from web sites. This way of gathering a respondent group is the most common way found in the literature, and one of the evaluations is constructed based on these principles.

The other evaluation in the thesis, the self-completion questionnaire, focused on the expert opinion from senior researchers, that are potential users of the tool and are interested in saving time and money. It was not easy to find a group of respondents in Sweden and the number of respondents in this evaluation could have been higher, making it easier to draw conclusions from the results. Looking at other similar research studies in the literature, the human respondent groups tend to vary between zero (0) human respondents to 100 non-experts (in Ribeiro et al.). The simulated user group requires zero (0) humans; the user group where some sort of sample is done could be about 10-12 humans; and the respondent group without any sort of sampling, recruited from a web page, could be up to 100 humans. It is obvious that the evaluation in the thesis would have benefited from a larger respondent group, but it was seen as of paramount value to get the opinions of human experts.

7.4.3 Choice of research method

Considering the results of the self-completion questionnaire, the choice of evaluation method could have been different. It took too long to get the answers from the respon-
ents, which made it impossible to do a follow-up with questions. If an interview was used, instead of a self completion questionnaire, the follow-up questions would have been made possible.

7.5 Ethical aspects

7.5.1 The respondents

In this research the number of respondents in the self-completion questionnaire was low. When being part of a small respondent group, it is hard to be anonymous. Contact was taken directly by asking the respondents if they wanted to be part of the research. Consequently, the information consent was made by the author asking the respondents if they wanted to be part of the research and what that would entail. Although the answers were sent via e-mail showing the sender, the answers were made anonymous in the research.

There is also a risk of pressure to answer the questionnaire, in spite of lack of time, when the respondents know that they are a part of a small participant group. Unfortunately, this could not be avoided, but by having personal contact with the respondents, they were assured of respect if the time was lacking.

7.5.2 Sustainability

The researchers, i.e. the respondents, are responsible to use the money provided for their research as effectively as possible. By using the kind of explanation method presented in this thesis, researchers could get the possibility to use a larger part of the funding to research.
Chapter 8

Future Work

The work in the thesis is the first step to develop a complete explanation tool, for researchers when working with automatic text classification. The idea is to use the best of two worlds; the time gain from the automatic text classification, with the human associative ability to detect errors and context, where the computer fails. It could include warnings when the PE for documents drops below 0.6 and the classification process tends to have more of an arbitrary nature.

The explanation aid would be used by researchers or other persons, positioned to evaluate, code or in other way control an automatic text classification. Since the fully implemented explanation aid would be a too extensive work to fit in a master thesis, the work has been limited to the functionality of AITCE. A first step as future work, a more extensive evaluation of the AITCE is suggested. To see if the indications from the evaluations could get verified and developed, with a structured interview together with an analyze of a larger, more extensive data set.
Ahlberg, E., Spjuth, O., Hasselgren, C., and Carlsson, L. (2015). Interpretation of con-
formal prediction classification models. In SLDS, The third International Symposium
on statistical learning and Data Sciences, pages 323–334.

Wesley, Harlow, 2. ed. edition.


applied to ames mutagenicity data. Journal of chemical information and modeling,

The role of trust in automation reliance. International Journal of Human-Computer
Studies, 58(6):697–718.

Eklund, J. (2016). With or without context: Automatic text categorization using semantic

Flach, P. (2012). Machine learning: the art and science of algorithms that make sense of
data. Cambridge University Press.


Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., and Lane, J. (2016). Big data and social
science: A practical guide to methods and tools. CRC Press.


Löfström, T. (2015). On Effectively Creating Ensembles of Classifiers: Studies on Creation Strategies, Diversity and Predicting with Confidence. PhD thesis, University of Borås, Faculty of Librarianship, Information, Education and IT. At the time of the doctoral defense, the following paper was unpublished and had a status as follows: Paper 8: In press.


Appendix A

The Self-completion Questionnaire
Introduction

The focus of this research has been to find out if the use of an explanation method could help humans with automatic text classification, and in that way increase the quality of automatic text classification.

The questionnaire consists of two parts; in the first part six articles are presented together with the words that explain why the classification method has assigned the article as the current class. The second part consists of questions related to increased usability and ease of use.

Automatic text classifiers are in general black boxes, where your input is the documents to be classified and the output is a table, filled with numbers. It is generally not possible to ask the classifier why it has mapped certain documents to certain classes. Human classifiers may add comments, but there is generally no explanation available for the automatically classified document.

<table>
<thead>
<tr>
<th></th>
<th>Politik</th>
<th>Samhälle</th>
</tr>
</thead>
<tbody>
<tr>
<td>The prediction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Politik</td>
<td>0.72</td>
<td>0.28</td>
</tr>
<tr>
<td>Samhälle</td>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td>The True value</td>
<td>Politik</td>
<td>Samhälle</td>
</tr>
<tr>
<td></td>
<td>Politik</td>
<td>Samhälle</td>
</tr>
</tbody>
</table>

Fig. 1 The output from different kinds of classification

Figure 1 show examples of how the result from an automated text classification (left) could look like, in comparison with the result from a human classification (right). When the value of a class is set to 1 (in the human classification) the text belongs to that class. The automated classifier ‘predicts’ a document to belong to a certain class. Based on which one of the different classes that gets the highest prediction value, i.e., probability estimate, in the the class columns ‘Politik’ and ‘Samhälle’ in fig. 1. The prediction is based on the True value (i.e., the class assigned by humans) from a set of similar, already classified, documents. The number of predicted documents where the assigned class equates the true values decides how accurate the automated text classification has. The sum of the probability estimates for all categories equals 1, or 100%. The first row in the table left in figure 1 shall be understood as the classification method being 72% sure that the document belongs to the category Politics (‘Politik’), and 28% sure that it belongs to the class Social issues (‘Samhälle’).

The number and characteristic of classes in an automated text classification could of be higher, but out of convenience (to make this evaluation easier) only two classes was chosen.

The explanation method used within this research, is based on an algorithm that find the words that has the biggest impact on the prediction of each class. The words are presented as an explanation on why the classifier predicts the document to belong to a specific class. In the questionnaire the words are presented in order of magnitude, in decreasing order from left to right.
Predicted class: Politic (90 %)

Words important for prediction nr. 1

<table>
<thead>
<tr>
<th>Pro Politic</th>
<th>Pro Social issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>partiet/ parti</td>
<td>landsting</td>
</tr>
<tr>
<td>socialdemokrat</td>
<td>tapp</td>
</tr>
<tr>
<td>röster</td>
<td>heta</td>
</tr>
<tr>
<td>sverigedemokrat</td>
<td></td>
</tr>
<tr>
<td>riksdag</td>
<td></td>
</tr>
<tr>
<td>eu-valet</td>
<td></td>
</tr>
<tr>
<td>politik</td>
<td></td>
</tr>
</tbody>
</table>

Questions:
1. When looking at the words in combination with the estimated certainty of the prediction (within parentheses), do you think the prediction is correct? Why/why not?
2. Read the text below and look at the words again. Do you understand why the text has been mapped into the actual class? Why/why not?
3. Would you yourself map the text into the actual class? Why/why not?
4. According to you, are there any words in the table above that point to the other class, than the predicted?

**Stort tapp för C i Sunne**

Sunne Centerpartiet i Sunne backar 10 procentenheter jämfört med förra EU-valet.

Victoria Gund
Predicted class: Social science (91%)

Words important for the prediction nr. 2

<table>
<thead>
<tr>
<th>Pro Social issues</th>
<th>året</th>
<th>jobb/jobbet</th>
<th>träffa</th>
<th>vecka</th>
<th>illa</th>
<th>villkor</th>
<th>arbetsmarknad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro Politic</td>
<td>pressmeddelande</td>
<td>hoppas</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Questions:

5. When looking at the words in combination with the estimated certainty of the prediction (within parentheses), do you think the prediction is correct? Why/why not?
6. Read the text below and look at the words again. Do you understand why the text has been mapped into the actual class? Why/why not?
7. Would you yourself map the text into the actual class? Why/why not?
8. According to you, are there any words in the table above that point to the other class, than the predicted?

LO kontrollerar att ungdomar jobbar under schysta villkor
Värmland.
Lisa Harkman
Predicted class: Politic (76 %)

Words important for the prediction nr. 3

<table>
<thead>
<tr>
<th>Pro Politic</th>
<th>regeringen</th>
<th>politik</th>
<th>rösta</th>
<th>moderaterna</th>
<th>välfärden</th>
<th>schweiz</th>
<th>franska</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro Social issues</td>
<td>exempelvis</td>
<td>sällsynt</td>
<td>bär</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Questions:
9. Read the text below and look at the words again. Do you understand why the text has been mapped into the actual class? Why/why not?
10. Would you yourself map the text into the actual class? Why/why not?
11. According to you, are there any words in the table above that point to the other class, than the predicted?

Sverige i en ding ding värld


Anna Dahlberg

HÖGDALEN, SVERIGE. Romska tiggare vräks och deras tält körs till soptippen. FINLAND. Finska skolan får högsta betyg för sina goda undervisningsresultat. SPANIEN. Kvinnor demonstrerar mot nya abortlagar och Norge gör som Spanien FRANKRIKE. Polisen bär bort kvinnor och barn för att sedan riva 165 kärstäder.
Predicted class: **Social science (81%)**

Words important for the prediction nr. 4

<table>
<thead>
<tr>
<th>Pro Social issues</th>
<th>satsning</th>
<th>konstatera</th>
<th>ort</th>
<th>alldeles</th>
<th>gång-</th>
<th>kommunalrådet</th>
<th>land</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro Politic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Questions:
12. When looking at the words in combination with the estimated certainty of the prediction (within parentheses), do you think the prediction is correct? Why/why not?
13. Read the text below and look at the words again. Do you understand why the text has been mapped into the actual class? Why/why not?
14. Would you yourself map the text into the actual class? Why/why not?
15. According to you, are there any words in the table above that point to the other class, than the predicted?

---

**Busspunkt invigd**

i Vålberg.

På fredagen invigdes den nya gång- och cykelvägen och busspunkten i Vålberg. Det utgör finalen på den stora ombyggnaden av Långgatan som pågått under drygt ett år.

– Jag tror att denna satsning är jätteviktig för Vålberg. Tidigare var trafikmiljön inte optimal här och många hade en alldeles för hög hastighet genom orten, konstaterar kommunalrådet Henrik Lander (C).
Köpte 200 Ipads – utan upphandling
Timrå kommun har gjort hundratals inköp utan upphandling. För skattebetalarna kan okända summor ha slö-sats bort.
– Vi får beklaga att missen är gjord, säger Stefan Eriksson, upphandlings- och it-chef.
   Ann-Christin Isaksson, förvaltningschef på barn och utbildningsförvaltningen, är en av de ansvariga för upphandlingen.
   – Det beror lite på att vi köpte in begagnade Ipads.
   Har ni gjort en miss?
   – Jag tycker naturligtvis det är väldigt viktigt att vi håller oss till de upphandlingsregler som finns.
   Tycker du att ni har gjort det?
   – Alltså, vi skulle kunna bli mycket duktigare på det här.
   Stefan Eriksson, upphandlings- och it-chef, är på semester när Dagbladet konfronterar honom med uppgifterna. Han är inte ovetande om att tjänstemän på kommunen har kringgått upphandlingsreglerna.
   – Inköpen föregicks inte av någon upphandling. Det är inte rätt. Om jag inte minns fel tror jag att det handlar om 200 begagnade Ipads.
   Kan inte konsequensen av att bara ett företag får lämna anbud vara högre priser och sämre kvalitet?
   – Jag vet egentligen inte hur det här har gått till väga. Men jag är övertygad att priserna hade varit lägre om det hade gjorts en upphandling.
   Är det här en isolerad händelse?
   – Ja, jag tycker nog att det är det.
   Samtidigt uppger Stefan Eriksson att en stor utbildningsinsats ”som redan var planerad” kommer att hållas i höst. Efter Dagbladets avslöjande kan även en stor omorganisering vara att vänta.
   – Det är kanske så att man ska fundera på att ha färre inköpare i kommunen för att kunna styra upp att vi följer… så att vi gör bra affärer för Timrå kommun och skattebetalarna, säger Eriksson.

Pontus Hellsén
Predicted class: Uncertain (Politic 52 %, Social issues 48%)

Words important for the prediction nr. 6

<table>
<thead>
<tr>
<th>Pro Social issues</th>
<th>samt</th>
<th>informera</th>
<th>barn</th>
<th>lokal-tidning</th>
<th>ungdom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro Politic</td>
<td>kommun-styrelsen</td>
<td>omröstning</td>
<td>förslag</td>
<td>planer</td>
<td>ytterligare</td>
</tr>
</tbody>
</table>

Questions:

19. When looking at the words in combination with the estimated certainty (within parentheses), would it be possible to categorise the document, based on the words above? Why/why not?

20. Read the text below and look at the words again. Do you agree that the words characterises the text? Why/why not?

21. Would the removal of one or more words from the list help to categorise the document?

---

Föräldrar tar strid för ny idrottshall

Kommunstyrelsen sa förra veckan ja till en ny idrottshall på Stuvsta iP. Hallen kan dock stoppas av Stuvsta iF:s styrelse, något som upprört en stor grupp föräldrar i klubben. Risken finns att det inte blir någon hall alls om klubben överklagar, säger Sanna Schönwell som tillsammans med många andra föräldrar med barn i Stuvsta idrottsförening, ororar sig över styrelsens övervägande att överklaga.


– Framförallt vill vi informera så många som möjligt om att styrelsen överväger ett överklagande. Därefter får var och en bilda sig en egen uppfattning, säger Sanna Schönwell.

Placeringen är kanske inte perfekt, enligt föräldrarna, men de positiva aspekternas överväger vida de negativa. Områdets skolor får en fungerande idrottshall, ungdomarna i Stuvsta får en ungdomsgård och en danslokals, och Stuvsta IF får nya fräscha omklädningsrum, samt en 5-manna konstgräsplan.

– Dessutom, med ytterligare en idrottshall i kommunen har vi större möjligheter att få ytterligare träningstider under vintern, säger Sanna Schönwell. På tisdag har klubben årsmöte. Då sker en omröstning i ärendet. Sedan är det upp till styrelsen att fatta beslut.

– Barnen är de stora förlorarna om hallen stoppas, säger Sanna Schönwell.

Monika Ruborg
Part two: Presentation of results

The results from the explanation method could be presented in different ways to enhance the perception of which words that has been important for the prediction of class. Within this research two ways of presentation have been chosen; one that focus on putting the words within its context and one that focus only on the words and the context they indicates.

Questions:
1) Which one of the versions do you prefer? Why?
2) Would a mix of both of them make the explanation better? Why/why not?

First version.

In this version the words that are important for the prediction of the document are highlighted within the text. Blue for the predicted class and grey for the other class.

<table>
<thead>
<tr>
<th>Words important for the prediction of the document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro Politic</td>
</tr>
<tr>
<td>Pro Social issues</td>
</tr>
</tbody>
</table>

Predicted class: Politic (90 %)

Stort tap för C i Sunne

Sunne Centerpartiet i Sunne backar 10 procentenheter jämfört med förra EU-valet.

Victoria Gund
Second version

In this version, the words that are important for the prediction of the document is viewed in so called ‘word clouds’. In these clouds the size of the words are equal to their individual importance for the prediction.

Predicted class: Politic

Pro Politic (91%)         Pro Social issues (9%)