Interpretable Instance-Based Text Classification for Social Science Research Projects

Helena Löfström, Tuwe Löfström and Ulf Johansson

Abstract In this study, two groups of respondents have evaluated explanations generated from an instance-based explanation method called *WITE* (Weighted Instance-based Text Explanations). One group consisted of 24 non-experts who answered a web survey about the words characterising the concepts of the classes and the other group consisted of three senior researchers and three respondents from a media house in Sweden who answered a questionnaire with open questions. The data used originates from one of the researchers' project on media consumption in Sweden. The results from the non-experts indicate that WITE identified many words that corresponded to the human understanding but also included some insignificant or contrary words as important. In the results from the expert evaluation, there were indications that there is a risk that the explanations could persuade the users of the correctness of a prediction, even if it is incorrect. Consequently, the study indicates that an explanation method

Helena Löfström
Jönköping International Business School, Jönköping University, Sweden
✉ helena.lofstrom@ju.se

Tuwe Löfström · Ulf Johansson
Department of Computer Science and Informatics, Jönköping University, Sweden
✉ tuwe.lofstrom@ju.se
✉ ulf.johansson@ju.se

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 5, No. 1, 2018

DOI: 10.5445/KSP/1000087327/15
ISSN 2363-9881

could be seen as a new actor which is able to persuade and interact with the humans and cause a change in the results of the classification of a text.

1 Introduction

Nowadays it is possible to automatically classify a large amount of documents with a high level of accuracy (Guidotti et al, 2018). However, many of the best performing classifiers behave like black boxes; the data is fed into the box which produces a prediction that is almost impossible to back-trace and understand (Ribeiro et al, 2016). If important decisions are to be made based on predictions, it is important for the user to understand and to evaluate the value of predictions. Understanding the reasons behind the predictions of a classifier is necessary in order to detect anomalies, to grasp the inner workings of the intelligent system and for long-term learning (Martens and Foster, 2014). If users understand that a classifier may give erroneous predictions, they could become unwilling to use automatic text classification, in spite of overall good results (Dzindolet et al, 2003). At the same time, users are more willing to use text classification if they are given a reason for the erroneous answers (Dzindolet et al, 2003). When classifying text documents, a general problem is the very high feature dimensionality, often with tens of thousands of features. The high dimensionality makes it very difficult for humans to understand the decisions made by the document classifiers (Martens and Foster, 2014).

Classifying text manually, or coding text as it is called in Social Science, is complicated and requires several steps. Text data is often messy and can contain insignificant information, e.g. page numbers or advertisements, making it a challenge to evaluate the quality of instance-based explanations. Before being able to code the entire data set, a coding scheme is required to be developed and the consistency between coders must be evaluated. If the consistency is low, the coding rules, as well as the definition of classes, must be revised (Wildemuth, 2009). After the coding of the entire data set, the consistency must be checked again since the human coders are likely to make mistakes during the coding, or new coders may have been added since the start. The coder's understanding of the classes may also change subtly during the coding process (Wildemuth, 2009). Human coding is an expensive, time-consuming and challenging task in a research project. At the same time, it is a crucial stage in the content analysis process. The transparency of the process is

seen as one of the advantages of this research method, which highlights the necessity of understandable explanations if an automatic text classification should be used (Bryman, 2012).

In the last couple of years, there has been an increased interest in interpretable machine learning at instance level. Automatic, quantitative evaluations are often used, but the explanations are to be presented to humans and are required to be understood by humans (Nguyen, 2018). Some form of human testing should be included when evaluating the explanations (Kirsch, 2017). The question should not only be if the system is interpretable, but also to whom (Tomsett et al, 2018). Different kind of actors are necessary to be approached, e.g. researchers (Tomsett et al, 2018).

This study aims at contributing to filling the gaps mentioned in the discussion above by including humans in the evaluation of an instance based explanation method applied on relevant real world data. In a first evaluation, a group of experts is asked to motivate their decision to either accept or decline the predictions and explanations using open questions in a questionnaire. This evaluation addresses the use case of the paper, which is a researcher having to spend a lot of his research budget on human coders. The intended use of this technique is to use automatic text classification to classify documents that the underlying model is confident about and to provide the human coders with the documents that the underlying model is uncertain about, along with explanations highlighting the most important words pointing to any of the categories. In a second evaluation, another group, with non-experts, has evaluated if the explanation method manages to associate the most influential words to the correct classes.

In the following section, a background is given covering text classification, how to construct comprehensive predictions, how to evaluate them, as well as a description of the suggested solution together with two metrics used in the study. Section 3 presents related work and in Section 4, the methodology used in the study is described, presenting details about the text classifier used as well as the interpretable predictive solution utilized. The method also introduces the evaluation of the results. In Section 5 results and analyses are presented. The conclusions and future work are presented in Section 6.

2 Background

Text classification is a process where a set of documents is divided into subsets that have something in common. These subsets are called classes or categories. The labels of the classes are symbolic, and intended to characterize or explain the documents in the class. The classes can describe different characteristics of the document such as genre, language, topic etc. 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 (see e.g. Baeza-Yates and Ribeiro-Neto, 2011 or Eklund, 2016). Automatic text classification most often relies heavily on pre-processing, a step which has great importance for predictive performance but also may create impediments making it harder to create interpretable predictions. When manually classifying documents, the human coders are required to read through each of the documents in the data set and from the rules in the code instructions decide which class each document belongs to. This is a time-consuming work and the costs are high, but the transparency of the work is seen as one of its major advantages (Wildemuth, 2009).

2.1 Constructing Understandable Explanations

There are different types of *explanation methods* in predictive modeling: *Model* (or global) explanation and *instance* (or local) explanation (see e.g. Martens and Foster, 2014 or Nguyen, 2018). Another type sometimes mentioned is a *rule* explanation (e.g. representing the path from root to leaf in a decision tree). The model explanation provides greater understanding of the entire classification model and its performance, while an instance explanation provides greater understanding of the model's predictions of a specific instance (see e.g. Martens and Foster, 2014 or Robnik-Šikonja and Kononenko, 2008). An explanation method makes the decisions of the classifier transparent, and is in that way independent of the accuracy of the prediction. However, it is reasonable to assume that the quality, or at least the usefulness, of the explanations increases with higher accuracy (Robnik-Šikonja and Kononenko, 2008).

Algorithm 1: WITE.

Input: The text classifier M, the vector representation x of the document.

1: Calculate the probability estimate for each class: $p = M(x)$
2: **for all** $x_f > 0$, i.e. for all words active in x **do**
3: Inactivate feature f in x: $x'_f = 0$
4: Calculate probability estimate for each class: $p'_f = M(x')$
5: Calculate the explanation coefficient: $ec_f = p'_f - p$
6: Reset x to its original state
7: **end for**

Output: ec, i.e., how much each feature / word affects each class p', i.e., probability estimates after inactivating each feature f.

When constructing an instance-based explanation method, different approaches can be used. The simplest solution is to use an interpretable model, e.g. a decision tree. For an interpretable model, the explanation for each instance is derived directly from the model. Taking a decision tree as an example, the path to the leaf predicting the instance serves as an explanation of the prediction of the model. A draw-back with interpretable models is that they are generally less accurate than more complex, opaque, models like ensembles, SVMs or different kinds of neural networks. When trying to construct an explanation using an opaque model, there are generally two distinct steps: The training of an underlying model used for prediction and a second step in which some form of explanation, based on the input features, is extracted for each instance. The solution used in this study, called *WITE* (Weighted Instance-based Text Explanations), utilizes the two-step strategy to construct explanations. WITE (see Algorithm 1) explains the predictions using the most important set of words for the prediction. Two metrics are used to measure the quality of the prediction based on the output from the underlying model, and the relevancy of each existing word:

- *Probability estimate (PE)*: A statistical probability estimate value, used to express an estimate of the probability for a document to belong to a certain class. It is intimately connected to the results of the classifier.
- *Estimate coefficient (EC)*: A value used to express in what direction each word influences the prediction.

WITE takes as input the text classifier M and the vector representation x of the document to get explained. 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. The algorithm steps through each of the active features f in x in the document, and calculates how the probability estimate is changed if the feature is excluded from the document. The estimate coefficient gets a positive value if excluding the feature strengthens the prediction, and a negative value if it weakens the prediction. The algorithm returns two lists; one list including how much each feature affects each class and one with the probability estimates after inactivating each feature f.

2.2 Evaluating Interpretability

Automatic evaluations of interpretability are frequently used since they are fast and easy to reproduce. Such evaluations focus on evaluating quantifiable metrics capturing some aspect of interpretability, most often using some estimate of explanation size (see Ribeiro et al, 2016 and Nguyen, 2018).

Many document classification tools need human understanding when making data-driven classification decisions (Martens and Foster, 2014). If the explanations are faithful and intelligible, the explanation methods are important in getting humans to use machine learning more effectively (see Ribeiro et al, 2016).

While an evaluation by humans is necessary, it is not an easy task (see Doshi-Velez and Kim, 2017). When including humans in evaluating explanations, the most common approach is to ask closed questions resulting in limited, quantitative answers (Ribeiro et al, 2016 or Nguyen, 2018). The questions often evolve around guessing the outcome of the prediction or trusting the prediction, given a set of words that most strongly affect the prediction (see e.g. Ribeiro et al, 2016, Nguyen, 2018 or Guidotti et al, 2018).

3 Related Work

In two reviews of explanations of expert systems (Lacave and Diez, 2004) and Bayesian networks (Lacave and Diez, 2002) the authors point at the important role of explanations for the acceptance of a system and write that

without it, the users would not be able to reject a system's recommendation when it is wrong and they would be reluctant to accept its advice even when it is right. This is in line with what Dzindolet et al (2003) writes, with the experience from the area of psychology. Also in Skitka et al (1999) the author points out that an automated aid does not necessarily result in a reduction of human errors, but could cause the creation of new classes of errors. Within psychology it is known that explanations influence peoples judgements, e.g. longer explanations, or explanations of scientific phenomena that contain statements or situations that are non-explanatory, affect the results (see e.g. Weisberg et al, 2008;Weisberg et al, 2015; Hopkins et al, 2016; Hopkins et al, 2019). The study by Hopkins et al shows that this effect drops, but still exists, with higher education or with experts. In Keil (2006) the authors write about the problems of gaps in explanatory understanding. They discuss how humans can use others as sources of information and that it is important to know if these "experts" are to be trusted.

In the survey of explainable AI (Adadi and Berrada, 2018) the authors write that explanation is a form of social interaction with psychological as well as cognitive and philosophical projections and that human ideas and behavior should be more visible in the field.

Today, with effective but opaque black box classifiers, interpretable machine learning is a hot topic. Several different explanation methods have been developed since Martens and Foster (2014) wrote that there were not many researchers focusing on instance-based explanations and encouraged more research in the area. The suggested solutions from others have been presented in different domains, with various ways of explaining the predictions (e.g. words in text or patches in an image). The focus of explanation methods is often one of the following: To create the most correct or best interpretative explanations or the effects on users; if they trust or distrust the system (see e.g. Lacave and Diez, 2004; Lacave and Diez, 2002; Ribeiro et al, 2016; Dzindolet et al, 2003; Chiou and Wong, 2010; Adadi and Berrada, 2018). In 2009, Carlsson et al presented an algorithm providing interpretable instance-based predictions for the chemoinformatics field where the interpretations were given as the active sub-parts of the molecules being predicted. The explanation method utilized in

this study was introduced in Löfström (2018)[1] and is inspired by the algorithm presented by Carlsson et al (2009). Martens and Foster suggested using a set of words as an explanation to the prediction, where, if one of the words were excluded, the prediction would change. In 2016 Ribeiro et al introduced an instance-based explanation method for text classification, Local Interpretable Model-agnostic Explanations (LIME), where the words are ranked according to their importance. Following the introduction of LIME, several studies have been presented which have focused on the evaluation of solutions for instance-based interpretable text classification (see e.g. Nguyen, 2018; Tomsett et al, 2018; Pedreschi et al, 2018). LIME uses a bag of words with a specified number of words to ensure an interpretable representation of the explanations. WITE explains the predictions of instances by presenting the words that are most important for the prediction in a sorted list. In Ribeiro et al (2016) the acceptance of a prediction is used as a synonym to trust. In this study the acceptance of a prediction is not automatically seen as an expression of trust, and the reasoning behind the opinion is investigated with open questions.

In the article by Nguyen (2018), it is suggested as a future work to use more specific application oriented tasks or evaluations, tailored towards specific user groups. Nguyen also suggests that if explanations are expanded with a visualization of the class distribution of the most influential words, it could make the explanations more informative. In this study, in line with the suggestions of Nguyen, a second group of non-expert respondents have evaluated the most influential words of each class (see Section 4.4).

Kirsch (2017) points out that users are often not involved or even mentioned when researchers propose methods and argued that evaluations of explanations must include some form of user testing. As Tomsett et al (2018) write, it is not enough to ask if the explanations are interpretable, the question must also include to whom it should be interpretable. The explanations may be required to be presented differently depending on the targeted users. Guidotti et al (2018) also emphasized the importance to ask what kind of decision is affected and which type of data record is more comprehensible. Tomsett et al (2018) points out that it is important to define interpretability in relation to a specific task and user group. In Ribeiro et al (2016) humans that have a basic knowledge and

[1] The implementation of the explanation method WITE began in early 2016, before the paper by Ribeiro et al was published online.

interest in the content of the data set are recruited, but as well as in the study of Nguyen (2018) the humans are not experts. In this study, expert respondents are used and the data used is connected to the respondents' expertise and to situations in which explanations could be required.

4 Methodology

This section presents the data set, the experimental setup and the evaluations. In this study the evaluation is conducted with human users, as suggested in Kirsch (2017). Two different types of evaluations are performed:

1. Instance-based explanations presented to experts;
2. In evaluation where non-experts indicated to what degree the words estimated to be most closely related to each of the two classes corresponded to the actual classes.

4.1 Data Set

The data set used in the study is a small subset concerning politics or social issues of a large corpus of manually classified – or coded – news articles from an existing research project in Sweden called "Gammelmedia" (Traditional Media). The human coders were trained and tested before classifying the documents, to make the result as consistent as possible.

The original data corpus consisted of approximately 5000 documents, but only about 1000 articles were digitally accessible as pdf-files. It contained articles covering 15 topics or classes with varying frequency. The two most frequent classes – Politics and Social Issues – were selected, and after sampling, the data set consisted of 178 documents from the two classes (with 89 articles per class). Each document was assigned to either the class Politics or the class Social Issues. The class given by the human coders was considered as the ground truth. Some of the documents used were very complex to classify since they were wandering in subject, covering several different possible topics. The documents also varied a lot in size, ranging from 27 words to 1446 words. All articles were written in Swedish, resulting in all explanatory words being in Swedish as well.

Preprocessing was done in KNIME (Berthold et al, 2008). During preprocessing, the documents were digitally transformed into text files. Numbers, punctutation characters, stop words from a list of Swedish stop words and one-letter-words were all filtered. Then the documents were transformed into indicator features using bag-of-words. Finally, the indicator variables representing words that only occurred in less than three documents were filtered. Stemming was not used since the entire words should be presented in the questionnaire, which might have affected the predictive performance slightly.

4.2 Experimental Setup

WITE was implemented in Python and all experimentation was performed using SciKit Learn (Pedregosa et al, 2011). The SciKit Learn classifier RandomForestClassifier was used with default settings. The PE produced by the underlying model summed to 1 for the two classes. Leave-one-out evaluation was used, meaning that one underlying model was evaluated for each document, using the remaining documents as training set. All documents with a PE below 0.5 was incorrectly predicted as the opposite class. Since WITE use of all words in the documents in the process, a threshold value of (0,01 EC) was introduced when selecting at most ten of the most important words (from each class) to explain each document. Since some words had a very low EC, especially in the predictions with high PE, this resulted in some cases with very few words pointing at one of the classes.

When presenting documents to the respondents, no prediction was revealed for uncertain documents with the argument that uncertain documents would most likely have had to be handed over to human coders to decide upon. By revealing an uncertain prediction, the human coders (and the respondents) could be biased by the revealed prediction, taking no or little heed of the high degree of uncertainty.

4.3 Expert Evaluation

The results from WITE were evaluated with a group of expert respondents. Three of the respondents were senior researchers working in the domain of media

science. The senior researchers were responsible for using manual classification in their research projects and familiar with the documents used, which made them possess a unique competence for evaluating the results. Using senior researchers as respondents when evaluating an explanation method applied to their own research data has not, to the best of our knowledge, been done before. Consequently, they were familiar with both the process of manual text classification and with the evaluated data and could benefit from using automatic text classification in future research projects, in order to transfer funding from manual coding to research time. They were chosen since their expert knowledge of the data and the code instructions could make them more inclined to question erroneous predictions and be more in line with the manual classification. The other three experts in the group were non-researchers working in a media house in Sweden. They have different roles in the company, but all of them are well familiar with similar texts as in the study. Furthermore, this group have, through their work, a unique knowledge of both the challenge of article categories and the experience of using automatised content clustering of news articles. They were chosen since they handle newspaper articles daily, and have a deep knowledge in the problems of fuzzy newspaper categories. In this group, the respondents are educated in journalism, sociology and data engineering and they work with data analysis in different ways.

For practical reasons, since the respondents are very busy people, they could only make an in-depth analysis of a handful of documents. Six documents were selected to represent clearly correctly predicted (PE > 0.8), clearly incorrectly predicted (PE > 0.8) and uncertainly predicted documents (PE around 0.5) from each class, to cover different kinds of situations. The uncertain predictions had a probability estimate close to 0.5 for both classes. The respondents were presented with the article text with important words highlighted, the prediction, the probability estimates and the list of top explanatory words sorted according to their weights (see Figure 1 and translation of the explanation words in Table 1). The respondents were asked to reflect on the prediction in relation to the text, if the most important words had been selected and if they agreed with the prediction. The manual classification, used as ground truth and targets, was not revealed. The expert respondents were given open questions in a self-completion questionnaire in order to evaluate if the explanations were able to provide them with insights about the causes of the predictions.

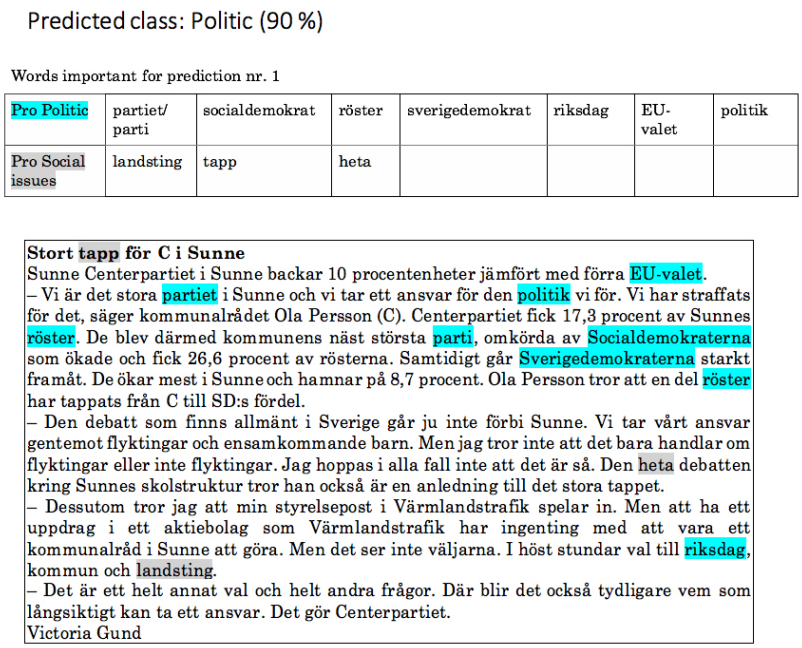

Figure 1: An example of how each document was presented to the respondents.

Table 1: Translation of the explanation words in Figure 1.

Swedish	**English**
partiet/parti	(the) party
socialdemokrat	social democrat
röster	votes
sverigedemokrat	sweden democrat
riksdag	parliament
EU-valet	EU-election
politik	politics
landsting	county/region
tapp	loss
heta	heated (as in heated discussion)

4.4 Non-expert Evaluation

To evaluate the overall performance of the words indicated as most important by WITE, a group of non-experts helped evaluate if WITE had managed to map the main characteristics of the classes. The group of non-experts were recruited from a Facebook group of adults, self-identifying as particularly gifted. The group was chosen because its members represents all parts of society, opinions and interests, and could be relied upon to provide serious answers. The total amount of members were 525, among which 24 answered the survey. The survey was implemented as a web-survey and introduced the respondents by a short text describing the purpose of the survey and how to respond to the questions. To find the most important words, EC was averaged over all the documents to achieve the overall performance for the entire data set. From each class, the fourteen words with the highest average EC were selected. The respondents were asked to choose which class they thought was best suited for each word. To avoid any bias, the words were presented in random order, unique for each participant.

5 Results

5.1 Classification and Algorithm

Before presenting the results from the human respondents, some experimental results summarizing predictive performance and overall modelling results are presented in Table 2. The accuracy of the underlying model was 74 %, indicating that the task was fairly difficult to predict.

Table 2: Summary of the predictive performance of the underlying model and the explanation method.

Class	Recall	Precision	F-measure
Politics	0.60	0.83	0.70
Social Issues	0.88	0.68	0.77

It is evident, from recall, that the underlying model performed better at predicting Social Issues than Politics. However, part of the explanation is that 64 % of the instances were predicted as Social Issues, resulting in lower precision for

Social Issues. The results indicate that the class Politics was more difficult, which is confirmed by the histogram in Figure 2. The figure plots the probability estimates produced by the underlying model for all documents coded by the human coders to belong to either class.

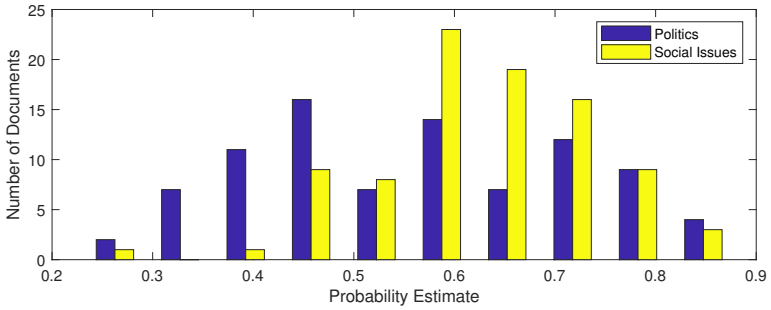

Figure 2: Histogram over the probability estimates for Documents coded as belonging to each class. The probability estimates are predicted by the underlying random forest models used.

A rather large proportion of documents, 36 %, had probability estimates in the range $[0.4, 0.6]$. If these are considered uncertain, the accuracy among the remaining documents was 85 %, with only one incorrectly predicted document coded as Social Issues. The accuracy among uncertain documents was 53 %.

5.2 Expert Evaluations

Tables 3, 4 and 5 list summaries of the answers given by the expert respondents. The columns in Tables 3 and 4 are: The class given by the human coders, and considered as the ground truth (*Class*), the respondent ID (*Resp.*), whether the respondent think that the prediction is correct or incorrect (*Prediction*), how the respondent would code the document (*Coded*) and a summary of their reasoning (*Reasoning*). The documents are represented by the first letter of their class – Politics (*P*) or Social Issues (*SI*).

Table 3 summarizes how the respondents used the explanations in their answers for the two correctly predicted documents (one from each class), and what they were missing from the explanations. The respondents generally agreed

with the predictions (see also Table 6), and would, based on the explanations, most often have coded the documents in the same way.

Table 4 summarizes how the respondents used the explanations in their answers for the two incorrectly predicted documents (one from each class), and what they were missing from the explanations. The two documents selected as incorrectly predicted documents had the highest probability estimates, i.e., they were the most clearly incorrectly predicted documents. Many of the incorrectly predicted documents were rather uncertain, especially among documents covering Social Issues (as can be seen in Figure 2). The respondents seemed confused over the amount of insignificant explanatory words, but still used both the explanatory words and the text to answer the questions (see also Table 6).

Table 3: The respondents' answers to the correctly predicted documents.

Class	Resp.	Prediction	Coded	Reasoning
P	I	P correct	P	The text is explicitly political. The prediction is correct based on the amount of clearly political words.
P	II	P correct	P	Clear focus, concerning formal politics and governance.
P	III	P correct	P	A text clearly focused on politic. Points out several words that indicate the other class.
P	IV	P correct	P	The text contains many political terms. Suggests a higher percentage of probability. Points out words that are too general.
P	V	P correct	P	The words fits well with the predicted class. The text has a clear political touch.
P	VI	P correct	P	The words indicate a high interest in politics. Points out some insignificant words and some important words that are missing.
SI	I	SI correct	SI	The text has a focus on workers rights. Some words are insignificant to the text.
SI	II	SI correct	SI	The text touches upon broader social issues, but does not involve formal political bodies.
SI	III	SI correct	SI	The text is not strictly political, but deals with a societal issue. Some words are a bit arbitrarily chosen.
SI	IV	SI correct	SI	The text is clearly about the labor market. The classification is not entirely clear. Some words are insignificant to the text.
SI	V	SI incorrect	P	The text is about unions, that are more of social issues. Understands the prediction based on the words. Thinks that there are several words in the text that have not been selected and that would have been essential for the prediction.

Table 4: The respondents' answers to the incorrectly predicted documents.

Class	Resp.	Prediction	Coded	Reasoning
SI	VI	SI incorrect	SI	The respondent writes that there is no red line among the words. Understands why it is classified into SI based on the text, but not the words that are considered insignificant.
P	I	SI correct	P	A focus on popular votes and laws put into effect. Some words are insignificant to both classes.
P	II	SI correct	Probably P	The text is hard to "pin down" since it is long and wanders in focus.
P	III	SI incorrect	SI	Many words point to social issues. The text has not a clear focus, but more a societal than a political.
P	IV	SI correct	SI	The text is correctly classified, based on the words. The respondent would although classify the text as Social Issues.
P	V	SI correct	P	Agrees with the predictions since there are several words that are associated with the predicted class. Would classify the text as the prediction since it is a debate article with a clearly political angle. Some words point to the other class.
P	VI	SI incorrect	SI	The respondent writes that the correct words have not been chosen. Would rather classify it as Social Issues.
SI	I	P correct	Probably SI	Most of the words seems too general to signify either class. Hard to classify, since a politician makes comments. Probably agree since the text does not concern itself with the political acts, but instead with a result that benefits the society.
SI	II	P partly correct	P	Understands why it is predicted as social issues since the word "kommunalråd" (local government commissioner) is involved. The respondent would relate it to formal politics but writes; "on the other hand it is tucked on in the end so I can understand why it was downplayed."
SI	III	P correct	Partly SI	Several insignificant words. A simple text, addressing a socially relevant fact.
SI	IV	P incorrect	SI	Do not think the prediction is correct, if looking at the words but understands the prediction. Would although classify the text as the prediction.
SI	V	P incorrect	P	The respondent does not think the text fits in either class, but is slightly more political. Does not consider any of the words as significant to either class. Points out the word "kommunalrådet" as political.
SI	VI	P correct	SI	The respondent writes that there is a strong indication that it concerns social issues, but the line is fine and it could be politics as well. Does not point at any of the words as either significant or insignificant.

Since the documents were presented as uncertain in Table 5, not revealing any preferred prediction, prediction agreement was not relevant. Instead of including the Prediction column, two other columns are included: whether the respondent found the explanatory words significant, i.e. sufficient to decide how to code the document (SW); if the explanatory words characterizes the text.

The respondents wrote (see Table 5) that these documents either were hard to classify or had a high amount of insignificant words. In the majority of answers, the respondents wrote that they could not classify the documents as belonging to either class (see also Table 6) and based their decisions on insignificant explanatory words. In several cases, they based their decision partly on words in the documents that were not listed as explanatory words, e.g. "kommunstyrelsen" (municipal government) or "kommunens skolor" (municipal schools).

Table 5: The respondents answers to the documents with uncertain predictions (1/2).

Class	Resp.	SW	Char.	Coded	Reasoning
P	I	No	Partially	None	The text is not easy to classify. None of the words signify either class. Some words characterizes the text, but do not help with the classification.
P	II	Yes	Yes	P	Clear case of politics, due to the focus on the explanatory word "upphandling" (procurement), the reference to "kommunens skolor" (municipal schools) in the document and the potential of breaking the law.
P	III	Some	No	P	Most of the words point to politics. The explanatory words pointing to Social Issues are insignificant and important words in the document pointing to both classes are not identified as explanatory.
P	IV	Some	Partially	None	The respondent writes that the words are quite vague and may well explain why the algorithm is uncertain about the classification.
P	V	No	No	None	The respondent writes that all the words are too vague.
P	VI	No	No	None	The respondent writes that the words have no connection to either class and that the problem is that the classifier does not find the correct words.
SI	I	Some	Mostly	P	The word "kommunstyrelsen" (municipal government) in the document in combination with the explanatory words "omröstning" (vote) and "förslag" (proposal) signify the class Politics more than any of the explanatory words pointing to Social Issues.
SI	II	Yes	Yes	P	The text has a political focus, due to the emphasis on the word "kommunstyrelsen" (municipal government) in the text and that the parents would fight the law.

Table 5: The respondents answers to the documents with uncertain predictions (2/2).

Class	Resp.	SW	Char.	Coded	Reasoning
SI	III	Some	Partially	None	The words characterize the document to some extent, but several important words in the text that point to the class Social Issues are omitted.
SI	IV	Yes	Yes	P	The respondent would classify the text as the classifier, but points out that the text deals with both social issues and politics. Removal of any words would not help with the classification of the text.
SI	V	Yes	No	SI	The respondent would classify the text as the classifier, since the political terms are in majority. But there are words that the respondent did not understand why they where chosen. Points out several words as insignificant.
SI	VI	No	No	P	The respondent writes that there are too few words with a strong connection to the classes to be able to characterise the text.

Table 6: The number of documents coded by the respondents in accordance with the ground truth (i.e., the previously coded labels) as captured by the Coded columns of Tables 3, 4 and 5.

	Correct	Incorrect	Neither Class
Correctly Predicted Documents	11	1	0
Incorrectly Predicted Documents	7	5	0
Uncertain Documents	3	4	5

5.3 Non-expert Evaluations

The results of the non-expert survey were analysed on two levels. First the general explainability: To which extent the words are recognized to belong to each class by the majority of respondents. Secondly the local explainability: How many of the respondents that recognized which class each different word belonged to. Table 7 summarizes the general explainability, using the number of words that was associated by the majority of respondents to the same class as WITE had associated the word with.

Table 7: The number of words that the majority of non-expert respondents associated with the the same class as the WITE model had associated it with.

WITE / Human	Politics	Social Issues
Politics	10	4
Social Issues	2	12

The result show that most of the words that WITE most strongly associated with the two different classes were also associated with the same classes by a majority of respondents. However, as many as 6 out of 28 words were in fact associated with the opposite class by the majority. This indicates that WITE uses several words when predicting a class that might not be associated with the same class by humans. These results align with the response from the expert respondents, who point at insignificant words being used by WITE.

When looking at the local explainability presented in Table 8, the distribution of votes can be seen per word. For words associated by WITE with Politics, as can be seen in Table 8a, 10 out of 14 words were associated with the same class by at least 80 % of the respondents. Two of these words were not associated with Politics (which was the class that WITE had associated it with). When considering words that WITE associated with Social Issues, as seen in Table 8b, only one word had more than 80 % of the respondents associating the word with Social Issues. 7 of the 14 words were associated with the same class by only 50–60 % of respondents (two of which were associated with Politics by the majority).

Although a higher percentage of the explanation words in the class Social Issues were associated by the respondents to belong to that class, the average agreement in association by the respondents and WITE for the different classes were similar, 68 % for the class Politics, and 70 % for the class Social Issues. In summary, while more words had a distinct mismatch between how humans and WITE associate words with Politics, the agreement on the remaining words were rather strong. For Social Issues, on the other hand, the picture is almost reversed, with a much larger degree of disagreement on which class the words were associated with.

Table 8: The explanation words describing each class.

(a) The distribution of answers from the web survey for the class Politics.

% Correct	Swedish	English
92 %	Regeringen	The Government
64 %	USA	The United States of America, USA
92 %	Partiet	The (political) party
84 %	President	A President
96 %	Valet	The Election
80 %	Kommunstyrelsen	The Municipality Board
100 %	Parti	A political Party
32 %	Japan	The country Japan
16 %	Ljus	Light
80 %	Politiken	The Politics
72 %	SD	Abbreviation for the political party of the Swedish Democrats
48 %	Pågå	Be Ongoing
20 %	Användas	To be Used
80 %	Sverigedemokraterna	The political party of the Swedish Democrats
68 %	Average vote in favor of Politics	

(b) The distribution of answers from the web survey for the class Social Issues.

% Correct	Swedish	English
76 %	Kritik	Critique
64 %	Varje	Each
68 %	Problemen	The Problems
60 %	Skolan	The school
56 %	Allra	A Controversial Insurance Company
60 %	Öppnar	Opening
80 %	Omkring	Around or Surrounding
80 %	Verksamheten	The Business or The Activity
60 %	Skapa	Create
84 %	Bo	Live
56 %	Hända	Happening
44 %	Förändra	To Change
40 %	Nationella	National
76 %	Samhälle	Society
70 %	Average vote in favor of Social Issues	

5.4 Discussion

Summarizing, the respondents agreed in most cases to the prediction, often pointing to the explanatory words as a reason to their decision. Although the respondents sometimes questioned the prediction, especially for the incorrectly predicted documents, they often ended up agreeing nonetheless. Although they did not know if it was a correct or incorrect prediction, they found evidence in the explanatory words to support the prediction. As an illustration, respondent II understood the prediction in one of the incorrectly predicted documents, based on the importance of one of the explanatory words (see Table 4).

Several of the documents (see Tables 4 and 5) were not easy to classify. The (senior researcher) respondents sometimes presented arguments that contradicted each other, although they have worked with the data and have been involved in writing the code instructions for the human coders classifying the documents (see Table 5).

Previous research has shown that explanations of predictions could alter the users perception of the predictions, often in a positive direction (see e.g. Ribeiro et al (2016) or Dzindolet et al (2003)). When using open questions, the reasoning behind this change of perception could be observed. The results from the study could indicate that if the words are considered important to the text, the respondent could get persuaded that the prediction is correct. At the same time it is important to ask what the goal of the explanation is: To make users accept the prediction or to make them able to inspect and correct an incorrect prediction.

Considering the disproportional amount of incorrect and unsure predictions in the questionnaire, it follows that the results from the respondents indicate a more negative picture of WITE and that the distribution of the different categories of predictions is important to consider in the analysis of the results. The intention was to evaluate WITE in an unfavourable situation, but it is important to realize that it does not mirror the results of the entire data set. The accuracy of the underlying model was 74 %, which indicates that the documents were fairly difficult to classify. From recall and precision (see Table 2) it is also possible to see that the class Politics was more difficult to predict than the class Social Issues. In the questionnaire, the accuracy did not reach more than 33 % and the amount of documents with an unclear prediction were almost twice as high as in the underlying model, 66 % in the questionnaire to 36 % in the model. In other words, the results from the questionnaire could be suspected to be considerably

less accurate than the results from of the entire data set. Nevertheless, the distribution of answers from the questionnaire (see Table 6) still mirrors the results of the classification fairly well.

In the non-expert evaluation, the average agreement between the respondents and WITE in association of the words were similar between the classes, about 70 %. The agreement on the words among the respondents were generally much higher for words associated by WITE with Politics, even when the respondents agreed on the opposite class. Even if the average agreement was similar between the classes, the words that WITE associated with Social Issues were much more ambiguous to the respondents, with less agreement internally among the respondents. Nguyen (2018) suggests that an explanation could be more informative with a visualization of the class distribution for the most influential words. The results from the non-expert evaluation shows that the most influential words associated by WITE with each of the classes were most often associated in a similar way by humans. However, several words considered insignificant or irrelevant by respondents were included by WITE, which could be seen in both the expert and non-expert evaluation. One possible reason is the rather small data sample, with only 178 documents. Even though leave-one-out was used in the experimentation, to maximize the size of the training set, it is still a very small sample to learn from, especially considering that the number of features is many times larger than the number of instances. An expectation is that a larger corpus of documents would, at least to some extent, solve these issues, since insignificant words would be less important to the underlying classifier, thus favoring words more strongly related to the concepts targeted (i.e., Politics and Social Issues).

6 Conclusions and Future Work

When explaining the predictions from a black box text classifier, a common approach is to present those words in the document that are most important for the prediction. As discussed in sections 1 and 3 it is unusual to evaluate the results with humans and no scientific research has been done, to the best of our knowledge, evaluating if such explanations are helpful when experts (e.g. researchers) work with real world data sets.

In this study the results from an instance-based explanation method was evaluated by three experts from a media house and three senior researchers,

with a data set from one of the researchers' own studies. The evaluation was conducted with open questions in a self completion questionnaire. In the questionnaire, a disproportional amount of incorrect and uncertain predicted documents were used, which caused a high amount of insignificant explanatory words. The evaluation showed that the respondents used the explanatory words, but that the noise from insignificant words could confuse and influence the interpretability of the explanations.

Although the expert respondents had either an intimate knowledge of the data or the problems with categorising and analysing news articles, they hesitated to contradict the prediction indicated by the prediction estimate. They used the explanatory words to either questioning or agreeing with the prediction, but in most cases they ended up agreeing with what the prediction estimate indicated. Although they most often wrote that the predictions were correct, they could argue for the other class or even write that they would classify the document differently.

Text classification is a difficult task and there is no definite true answer to which class the documents should belong, it is a question of interpretation of the text. The human classification is seen as most accurate, the golden standard or ground truth. Although the expert respondents either had an intimate knowledge about the data or of this type of problem, and one of them had written the code instructions, it did not make them immune from interpreting the text from a new angle when the explanatory words and the prediction was presented to them.

As discussed earlier in Section 3 it is known within psychology that explanations may influence humans. Studies have also shown that this risk decreases with higher education or expertise knowledge. Earlier research about explanation methods show that explanations of predictions may increase the acceptance of predictions, but since open questions have not been used, the reasons why has not been known. The assumption is that the increased acceptance is based on a higher degree of trust.

In this study, the results from the expert respondents indicate, just as in the larger psychological studies (see Section 3), that even human experts could get convinced by explanations. Although the sample of experts is small, it shows that this tendency exists in explanation methods. In this sense an explanation method could be seen as a possible new actor which is able to persuade and interact with the humans and cause a change in the human perception of the document. In a critical situation, as in medical diagnosis or in a situation of

military decisions, this could lead to fatal consequences if the prediction is incorrect. Since the number of respondents is small, the results must be seen as indicative and suggestive for future research directions rather than actual proof in any specific direction.

In the results from the non-expert evaluation the majority of the words where correctly identified to each class. As was also discussed in Section 3, it may be possible that inclusion of a list of words more generally describing the classes could make an explanation more informative by helping the respondent get a better grasp of the class concept in general. E.g., in a situation similar to that of the questionnaire, it may be that adding general words describing the classes could help the user to question an incorrect prediction.

It would therefore be interesting, as a future work, to study if the increased acceptance of the predictions when using explanation methods could be a result of the explanation's possibility to convince the respondents, and if words providing a general description of the classes could lessen this effect.

Acknowledgements This work was supported by the Swedish Knowledge Foundation (INSiDR FO2018/9 and DATAKIND 20190194), by the Swedish Governmental Agency for Innovation Systems (Airflow 2018-03581) and by the region Jönköping (DATAMINE HJ 2016/874-51).

References

Adadi A, Berrada M (2018) Peeking Inside the Black-box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6:52138–52160, Institute of Electrical and Electronics Engineers (IEEE). DOI: 10.1109/ACCESS.2018.2870052.

Baeza-Yates R, Ribeiro-Neto B (2011) Modern Information Retrieval – The Concepts and Technology behind Search, 2nd edn. Addison Wesley, Boston (USA). ISBN: 978-0-321416-91-9.

Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2008) KNIME: The Konstanz Information Miner. In: Data Analysis, Machine Learning and Applications, Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (eds), Springer, Berlin, Heidelberg (Germany), Studies in Classification, Data Analysis, and Knowledge Organization, pp. 319–326. DOI: 10.1007/978-3-540-78246-9_38.

Bryman A (2012) Social Research Methods, 4th edn. Oxford University Press, New York (USA). ISBN: 978-0-199689-45-3.

Carlsson L, Helgee EA, Boyer S (2009) Interpretation of Nonlinear QSAR Models Applied to Ames Mutagenicity Data. Journal of Chemical Information and Modeling 49(11):2551–2558, ACS Publications. DOI: 10.1021/ci9002206.

Chiou A, Wong KW (2010) Auto-explanation System: Player Satisfaction in Strategy-based Board Games. In: Entertainment Computing Symposium (ECS2010), Nakatsu R, Tosa N, Naghdy F, Wong K, Codognet P (eds), Springer, Berlin, Heidelberg (Germany), IFIP Advances in Information and Communication Technology, Vol. 333, pp. 46–54. DOI: 10.1007/978-3-642-15214-6_5.

Doshi-Velez F, Kim B (2017) Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint. URL: https://arxiv.org/abs/1702.08608.

Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP (2003) The Role of Trust in Automation Reliance. International Journal of Human-Computer Studies 58(6):697–718, Elsevier B.V. DOI: 10.1016/S1071-5819(03)00038-7.

Eklund J (2016) With or Without Context: Automatic Text Categorization Using Semantic Kernels. PhD thesis, Borås (Sweden), University of Borås. ISBN: 978-9-198165-48-7, URL: http://hb.diva-portal.org/smash/record.jsf?pid=diva2%3A906045&dswid=4314.

Guidotti R, Monreale A, Ruggieri S, Turini F, Pedreschi D, Giannotti F (2018) A Survey Of Methods For Explaining Black Box Models. ArXiv e-prints. URL: https://arxiv.org/abs/1802.01933. Provided by the SAO/NASA Astrophysics Data System.

Hopkins EJ, Weisberg DS, Taylor JC (2016) The Seductive Allure is a Reductive Allure: People Prefer Scientific Explanations that Contain Logically Irrelevant Reductive Information. Cognition 155:67–76, Elsevier B.V. DOI: 10.1016/j.cognition.2016.06.011.

Hopkins EJ, Weisberg DS, Taylor JC (2019) Does Expertise Moderate the Seductive Allure of Reductive Explanations? Acta Psychologica 198:102890, Elsevier B.V. DOI: 10.1016/j.actpsy.2019.102890.

Keil FC (2006) Explanation and Understanding. Annual Review of Psychology 57:227–254, Annual Reviews. DOI: 10.1146/annurev.psych.57.102904.190100.

Kirsch A (2017) Explain to Whom? Putting the User in the Center of Explainable AI. In: Proceedings of the 1st International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), Besold T, Kutz O (eds). URL: https://hal.archives-ouvertes.fr/hal-01845135.

Lacave C, Diez FJ (2002) A Review of Explanation Methods for Bayesian Networks. The Knowledge Engineering Review 17(2):107–127, Cambridge University Press. DOI: 10.1017/S026988890200019X.

Lacave C, Diez FJ (2004) A Review of Explanation Methods for Heuristic Expert Systems. The Knowledge Engineering Review 19(02):133–146, Cambridge University Press. DOI: 10.1017/S0269888904000190.

Löfström H (2018) Time to Open the Black Box: Explaining the Predictions of Text Classification. Master's thesis, University of Borås, Borås (Sweden).

Martens D, Foster P (2014) Explaining Data-driven Document Classifications. MIS Quaterly 38(1):73–100, Society for Information Management and The Management Information Systems Research Center, Carlson School of Management, University of Minnesota. DOI: 10.25300/MISQ/2014/38.1.04.

Nguyen D (2018) Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'18): Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans (USA), pp. 1069–1078. DOI: 10.18653/v1/N18-1097.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research (JMLR) 12:2825–2830. URL: `http://www.jmlr.org/papers/v12/pedregosa11a.html`.

Pedreschi D, Giannotti F, Guidotti R, Monreale A, Pappalardo L, Ruggieri S, Turini F (2018) Open the Black Box Data-driven Explanation of Black Box Decision Systems. arXiv preprint. URL: `https://arxiv.org/abs/1806.09936`.

Ribeiro MT, Singh S, Guestrin C (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery (ACM), New York (USA), pp. 1135–1144. ISBN: 978-1-450342-32-2, DOI: 10.1145/2939672.2939778.

Robnik-Šikonja M, Kononenko I (2008) Explaining Classifications for Individual Instances. IEEE Transactions on Knowledge and Data Engineering 20(5):589–600, Institute of Electrical and Electronics Engineers (IEEE). DOI: 10.1109/TKDE.2007.190734.

Skitka LJ, Mosier KL, Burdick M (1999) Does Automation Bias Decision-making? International Journal of Human-Computer Studies 51(5):991–1006, Elsevier B.V. DOI: 10.1006/ijhc.1999.0252.

Tomsett R, Braines D, Harborne D, Preece A, Chakraborty S (2018) Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. ArXiv e-prints. URL: `https://arxiv.org/abs/1806.07552`. Provided by the SAO/NASA Astrophysics Data System.

Weisberg DS, Keil FC, Goodstein J, Rawson E, Gray JR (2008) The Seductive Allure of Neuroscience Explanations. Journal of Cognitive Neuroscience 20(3):470–477, MIT Press. DOI: 10.1162/jocn.2008.20040.

Weisberg DS, Taylor JCV, Hopkins EJ (2015) Deconstructing the Seductive Allure of Neuroscience Explanations. Judgment and Decision Making 10(5):429–441. ISSN: 1930-2975, URL: http://journal.sjdm.org/15/15731a/jdm15731a.pdf.

Wildemuth BM (2009) Application of Social Methods to Questions in Information and Library Science. Libraries Unlimited, Westport (USA). DOI: 10.5860/0710082.