Co-funded by the
Erasmus+ Programme
of the European Union

Performance of machine learning algorithms: case study of BørneTelefonen

MASTER THESIS
to obtain the Erasmus Mundus Joint Master Degree
in Digital Communication Leadership (DCLead)

of

Faculty of Cultural and Social Sciences
Paris Lodron University of Salzburg

and

Technical Faculty of IT and Design
Aalborg University in Copenhagen

Submitted by
Sharona Boonman MSc
s1061377
sharonaboonman@gmail.com

Prof. Anders Henten (Aalborg University)
Prof. Josef Trappel (Salzburg University)
Prof. Leah Lievrouw (University of California, Los Angeles)

Department of Communication Studies
Salzburg, July 31st 2020

Table of Contents

List of Figures

List of Tables

1. Executive summary

It can take the Danish child helpline, called BørneTelefonen, over a month to answer the letters from Danish children seeking help. Every child deserves an answer, but some problems are more severe than others. The current response rate of BørneTelefonen is especially concerning for neglected children. A way to deal with this problem is by automatically classifying the incoming messages into categories with answer priorities. However, the implications and limitations of automated classification need to be considered. Therefore, the main research question of this study is 'To what extent can machine learning algorithms classify incoming messages to organizational helplines in comparison to human coders?'. Two sub questions are formulated to help answer the main research question. The first question is related to the technical possibilities of automatically classifying text messages *('How accurate are machine learning algorithms when classifying incoming messages in comparison to human coders?').* The second question focusses on the possible limitations *('What are possible technical, social, cultural, and ethical limitations when using machine learning algorithms to classify incoming messages?').* To answer the first sub research question, seven machine learning algorithms were applied to a dataset of 5664 messages of BørneTelefonen. Human coders labeled these messages in terms of neglect. When comparing the results of the machine learning algorithms against the performance of the human coders, support vector machine (SVM) performed most optimal. SVM had an F1 score of 94 percent for messages which were not labeled as neglect, and 23 percent for messages labeled as neglect. These scores are expected to improve when human-guided machine learning is applied to future incoming messages. Thus, machine learning classifiers offer great potentials to classify incoming messages. Concerning the results of the second sub question, technical limitations are discussed for the data used in this research, including the language of the messages, errors, and chat language. Also, potential societal implications may arise when using a machine learning classifier. Concerning the main research question, while it is algorithmically possible to automatically classify incoming messages, it might be necessary to inform people that their message is classified by an algorithm, and give them the chance to opt out of it. Further research is needed to understand the possible ethical issues involved to ensure fair and responsible use of machine learning for classifying incoming messages.

Keywords: classification, machine learning, natural language processing, children, child welfare, Danish language, ethics.

2. Introduction

Children with questions and problems can usually approach people, such as their parents, teachers, and their general practitioner. Still, some children find it difficult to go to these people to seek help. They can have questions about their sexuality that they find uncomfortable discussing with someone they know. Also, they can have concerns related to the alcohol consumption of their parents which they want more advice on without any (negative) consequences. In these cases, a child often wants to stay anonymous and get one-time advice. Luckily, child helplines offer help. Traditionally, it was possible to call these helplines to discuss the issue. Nowadays, children can choose to chat with adult volunteers, send a text message, and write a letter to adult counselors or other children. Also, children can read letters from other children, watch videos about questions and issues, and even do an online quiz which is followed by advice (e.g. the quiz called '*Do I have a crush?*') (BørneTelefonen, 2019; Fukkink & Hermanns, 2009; Sindahl, 2011; van Dolen & Weinberg, 2019).

Children welcome the additional options that child helplines are providing. Most children mainly experience written counseling and conversations as helpful. With this form of digital communication, the child is in control, experiences a high degree of anonymity, and can re-read the conversation when needed. Furthermore, this form of counseling is silent, which might help the child to feel less concerned with whether or not someone overhears them (Andrade, 2003; Caplan, 2003; Sindahl, 2011; Sindahl et al., 2018). Still, there are some limitations regarding this type of counseling. The most present limitation is the fact that written counseling is more time-consuming than counseling over the phone (BørneTelefonen, 2019; Fukkink & Hermanns, 2009; Sindahl, 2011).

The voluntary adult counselors of the Danish children helpline, called BørneTelefonen, answer such letters. This is done on a first-come-first-served basis. Currently, it takes the team on average seven days to reply to a letter. But, this can increase to a duration of more than a month. This is a long time for most children and can especially be critical for those with severe problems (Cho et al., 2013; Sindahl, 2011). Ideally, the Danish child helpline has the human capabilities to answer the letter of every child in time. However, if these human resources are not present it could at least be valuable if the incoming messages could be automatically classified, e.g. in terms of neglect, which would allow for prioritization of the critical cases. The classification could support experts to provide timely replies to children with problems related to neglect. One of the aims of this research is to

gain a better understanding of the differences between human and machine classifications for text messages. This knowledge could help to understand how appropriate it is to use machine learning algorithms to automatically classify letters, e.g. of Danish children in terms of neglect.

2.1 Relevance of research

2.1.1 Societal relevance

This research aims at classifying letters from children that are experiencing neglect to give them timely advice. Therefore, mainly neglected children are expected to benefit from this research.

Furthermore, this research can help BørneTelefonen volunteers to allocate their resources more effectively. Also, it is expected that volunteers gain a greater sense of purpose when they also get to work with urgent letters instead of just the next one on the endless-seeming pile of incoming letters (Sindahl, 2011).

The results of this research also contribute to potential long-term societal benefits. This is because research has shown that helping children with serious problems (regarding neglect) at an early stage allows them to be happier and function better in the society later in their life (Fergusson et al., 2005; Sindahl et al., 2018; Wisse & de Meij, 2015).

The Danish child helpline, BørneTelefonen, and child helplines in other countries can benefit from this research as it can facilitate the prioritization of incoming letters. Also, having a greater understanding of the effect of using machine learning classifiers could be useful for many other organizations that are considering to optimize processes regarding incoming messages or letters, such as consumer-oriented businesses.

2.1.2 Scientific relevance

This research consists of several components that are less frequently researched, especially the combination of these components seems to be unique. The main components that are expected to contribute to the scientific community are the fact that various machine learning algorithms are used for text classification, that natural language processing is applied on the Danish language, and that natural language processing is applied on the text of children containing e.g. typos, slang, and abbreviations. Also, ethical machine learning is considered of great value to both data science and social sciences. Especially because there is currently a lack of understanding about biases in data of Danish children who are experiencing neglect, this study combines technical and social elements in digital communication technology studies.

The scientific findings of this study are relevant to many different organizations and in many different fields. For example, the fact that natural language processing is applied to the Danish language could be helpful for other studies dealing with data from a language that is only spoken by a small portion of the population. The same counts for dealing with errors in messages from children.

2.2 Research questions

Main research question

To what extent can machine learning algorithms classify incoming messages to organizational helplines in comparison to human coders?

Sub research question 1

How accurate are machine learning algorithms when classifying incoming messages in comparison to human coders?

Sub research question 2

What are possible technical, social, cultural, and ethical limitations when using machine learning algorithms to classify incoming messages?

Seven algorithms are applied to one corpus of 5664 messages written by children to BørneTelefonen, a Danish child helpline. The results of each algorithm will be compared with results from human classifiers. Also, the implications and limitations of using machine learning algorithms instead of human classifiers are discussed.

3. Literature review

As follows, the findings of relevant theories and studies are summarized to better understand the issue of prioritizing incoming letters to a child welfare organization on neglect (Saunders et al., 2009). A top-down approach is followed. First, literature on child helplines is reviewed. Second, linguistic research is presented. Third, the field of computational linguistics, called natural language processing, is reviewed. Fourth, scientific literature on machine learning algorithms for natural language processing are discussed.

3.1 Child helplines

A child helpline is a counseling service for children and young people to get confidential support (without the consent of parents) about their issues, which is free of charge about their issues. Globally, there are at least 178 child helplines which are located in more than 146 countries (Fukkink & Hermanns, 2009). One of the advantages of child helplines is that they lack the barriers frequently associated with other health service organizations. Often, child helplines are the first point of contact with any kind of child protection organizations (Fukkink & Hermanns, 2009; Sindahl et al., 2019; van Dolen & Weinberg, 2019).

Most child helplines can be contacted for any type of problem, but some child helplines focus on a specific target group (e.g. bullied people, transsexuals, or domestic violence). One of the reasons why child helplines exist is to reduce child abuse by giving the right tools and support to people. There are four main types of child abuse according to the World Health Organization: emotional (or psychological) abuse, sexual abuse, physical abuse, and neglect (Butchart et al., 2006). Child neglect means that a child does not get all the required basic needs including health care, housing, emotional support, education, clothing, and security. Children can go to the child helpline to explain their problem and to get the required support or information (Areen, 1974).

In 3.1.1 the audience of child helplines is being described. This is done by explaining the kind of children that ask for advice, and the type of topics they request advice on. Also, the different kinds of support that children request is being presented, as well as how children feel after having received help from a counselor. Subsubsection 3.1.2 describes the different communication channels that child helplines use to provide advice.

3.1.1 Audience

Children and young people seek assistance from helplines for different types of support, e.g. emotional or informational. Some children need help with issues that are characterized by high levels of emotional distress, like violence, depression, and suicidal tendencies. These situations are often critical and potentially harmful. Other people seek assistance on practical questions that require more informational advice, such as how to get friends, how to kiss, and which contraception to use (van Dolen & Weinberg, 2019).

Given the different needs of children, it is expected that the counselor should listen more (i.e. not type) when the child is seeking emotional support. Also, the counselor should manage the duration of the chat as longer chat negatively influences the immediate well-being of the child. On the other hand, when the child is looking for instrumental support (i.e. information) then the counselor should play a larger role in the conversation as this creates a positive perceptions of quality. This means that the impact of the counselor's relative word count on children's perceived quality and direct well-being would change depending on the type of support the child seeks. Therefore, the counselor that is assigned to help the child needs to be sensitive to early indicators of the reason for the chat (Cohen, 2004; Kaufmann & Beehr, 1986; van Dolen & Weinberg, 2019).

The children who contact child helplines often experience relatively severe emotional issues (Fukkink & Hermanns, 2009). These feelings diminished after contacting a child helpline, and the perceived burden of their problems was reduced. Also, the service succeeded in increasing the general well-being and the feeling of empowerment of the children that contacted them (Fukkink & Hermanns, 2009; Sindahl et al., 2019). Therefore, it can be said that child helplines are a key resource for the state of the mental health of many children (Fukkink et al., 2016).

The most common identified gender at online youth web-counseling services is female, the second-highest identified gender is male, and there is a small group of people that identify themselves in other categories. Kooth is an example of an online web-counseling service for youth. 71 percent of Kooth users are girls, compared to 52 percent in youth (face-to-face) health care services. One of the expected reasons why girls make more use of online mental support has to do with the fact that girls engage more in social media than boys (Eriksson et al., 2012; Glasheen & Campbell, 2009). Furthermore, boys are

more likely to use an online web-counseling service at a younger age than girls (Beardsmore, 2015; Glasheen & Campbell, 2009).

3.1.2 Communication channels

Child helplines provide their services through different communication channels, like text-message (SMS), email, chat, and telephone. These channels share the characteristics of being anonymous, dialogue-based, and mediated. A telephone conversation is voice-oriented, which is in contrast to text-message, chat, and email which are text-based, and therefore require typing and literacy skills. Chat and telephone are synchronous and require constant presence, whereas text-message and email communication are asynchronous, and do not require constant presence (van Dolen & Weinberg, 2019).

Children perceived the quality of the advice given through text-based communication channels to be higher than of the voice-oriented help (Fukkink & Hermanns, 2009). Text-based communication gives children the feeling that they are anonymous, at ease and in control, as no one can overhear the conversation and they can keep re-read the response of the counselor in times of need (Andrade, 2003; Caplan, 2003; Sindahl, 2011; Sindahl et al., 2018). Still, various challenges related to online counseling caused by the time delay, the anonymity of the counselor, and the lack of nonverbal communication methods, can arise. Luckily, these issues were generally not considered insurmountable, and technologies have the potential to reduce these identified concerns (Fletcher-Tomenius & Vossler, 2009).

This explains why more child helplines are shifting from counseling over the phone to various text-based approaches (Sindahl et al., 2019). However, timely advice must be given as long waiting times for children that need emotional support could harm their well-being (van Dolen & Weinberg, 2019). One of the possible ways to reduce waiting time for children that need a certain type of support (e.g. emotional support) is to make use of prioritization techniques (Fletcher-Tomenius & Vossler, 2009; Hirschberg & Manning, 2015; Lopez & Kalita, 2017). Currently, it seems that most child welfare organizations do not use prioritization techniques for incoming letters. The letters are mostly handled on a first-come-first-served basis (Davidson et al., 2017; Smith & Donovan, 2003).

3.2 Linguistics

Effective communication is salient to child welfare work, and therefore it is key that the number and the significance of communication issues should be reduced (Kriz & Skivenes, 2010). Communication is the process of transferring messages or information between two or more people while language is a tool of communication. The field of linguistics is concerned with the nature of language and (linguistic) communication, and it includes the study of semantics and syntax (Akmajian et al., 2017; Jakobson, 1961).

3.2.1 Children language

Unsurprisingly, children's early language skills are important for later school performance (Magnuson et al., 2009). Children's language skills are positively correlated with age. In general, girls are ahead of boys in combining words, productive vocabulary, and communicative gestures in various kinds of language communities (Eriksson et al., 2012). Furthermore, children with a minority background and/or low-income households hear 30 million fewer words than their affluent counterparts in the early years of life. This harms their school achievements, and day-to-day (writing) skills (Golinkoff et al., 2019; Sperry et al., 2019). Depressed children are also more likely to experience communication difficulties in their families. Besides, there are linguistic differences identified in written communication for children with and without severe problems, e.g. text length (Katalin, 2010; Magnuson et al., 2009).

3.2.2 Identify children that need help with linguistic analysis

The use of linguistic analysis for online consultations to identify the needs by people is a new, but promising field. Jones et al. (2019) conducted a linguistic analysis of e-consultations to find whether people have mental problems. They identified various promising linguistic features, warranting the potential of further research when larger samples are used. Sumner et al. (2012) also found that linguistic analysis of online communication has great potential, as it has a statistically significant relationship with personality. Rini et al. (2015) successfully applied linguistic analyses to indicate cognitive processing, negative emotions, and practical problems.

Natural language processing (NLP) can be used in a linguistic analysis where children with a problem (e.g. children that experience neglect) can be clustered separately from the rest. NLP is a field of study at the intersection of linguistics and artificial intelligence, as displayed in Figure 1. The term artificial intelligence (AI) includes all fields of research that investigate the application of human intelligence (in terms of thinking and actions) by machines (Russell & Norvig, 2016). This field is too broad for this research, and therefore this study only focuses on the subdomain machine learning and its subdomain called deep learning. The following subsection describes the fields of natural language processing and machine learning in detail.

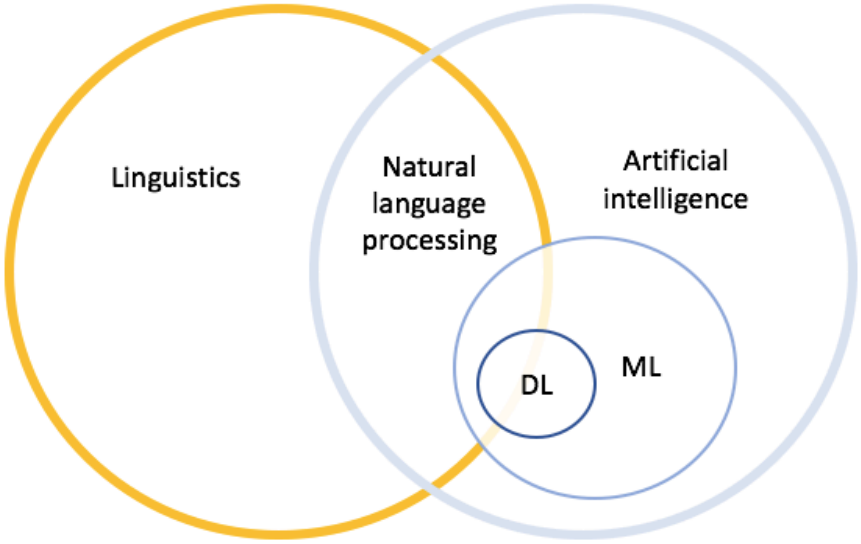

Figure 1. The interrelationships of linguistics, NLP, and AI (Ding, 2019).

3.3 Natural language processing

Natural language processing is an area of research that combines computer science, artificial intelligence, and linguistics. It also assesses the interactions between human language and computers. Natural language processing is a method for computers to analyze, understand, and derive meaning from human language in an intelligent and useful way. Natural language processing is a challenging field of study as it deals with human speech instead of programming languages. Human speech can contain many complex variables, including abbreviations, dialects, social context, slang, and typos (Hirschberg & Manning, 2015; Lopez & Kalita, 2017). Still, some NLP techniques can reduce noisiness in texts which is particularly useful for messages written by children

(Baldwin et al., 2013; Dey & Haque, 2009; Sahakian & Snyder, 2012; Sindahl, 2011). Also, NLP can offer help in the analysis and understanding of non-English texts, including Danish texts (Braasch, 2002; Derczynski, 2019; Owoputi et al., 2013).

3.3.1 Messages written by children

There are several reasons why natural language processing is often classified as a difficult approach. Ambiguities in language and the difficulty of expressing intent through semantically accurate language. These challenges are even more persistent in text written by (young) children (Carrell et al., 2017; Kreimeyer et al., 2017; Wu et al., 2016).

Despite the difficulties of applying NLP on text written by children, the technique achieved high score in classifying text written by children in cases of issues, such as possible abuse and suicidal behavior (Amrit et al., 2017; Katalin, 2010; Seedall et al., 2019).

3.3.2 Messages written in Danish

Danish, a North Germanic language, is a language that is mainly spoken in Denmark. Denmark is a country that is famous for its scientific and technological innovations, but this is not reflected in the language (Kirkedal et al., 2019; Kotov, 2017). Kirkedal et al. (2019) researched NLP tools and researched that were developed for the Danish language, and noticing that the availability of progressive modern technologies is limited.

Kirkedal et al. (2019) mentioned that it is necessary to change the Danish language technology to directly engage and support the global standard in NLP if users of the Danish language want to benefit from the advantages of natural language processing. Applied Danish natural language processing, Danish syntactic tools, and Danish semantic processing are the main pillars of this kind of strategy. Before this, Kotov (2017) described a similar idea to reduce the manual labor of analyzing a rare language as Danish. This solution involved the creation of a pipeline for processing the word-lists, this reduces labor due to the automation of lemma ascribing processes, and part-of-speech tagging. Also, Derczynski (2019) describes a set of basic machine learning-based tools for automatic processing of Danish documents. The tools use NLP models which are trained over previously annotated text.

3.3.3 Text classification

One of the fundamental tasks in natural language processing is text classification (Lai et al., 2015). Text classification aims to label the received text in several pre-defined categories, e.g. in the categories neglect/priority and no neglect/priority (Khan et al., 2010; Lai et al., 2015). This can be done manually, automatically, or semi-automatically.

Labeling text can be a difficult task, even for humans. This is because labeling can be highly subjective. If a person is unable to label properly, then it is important to remember that the computer will at least perform as bad as the person itself. This is why labeling data should be done by someone with a lot of knowledge about the problem that is to be solved from a human standpoint (Bowker & Star, 2000). Defining clear rules of what should receive which label is a good approach to classify text. It is hereby important to stick to the set of defined rules. Also, it is recommended to start defining the easiest examples first. The hardest ones can be left until the end as it is expected that the human being has then a better comprehension of the problem. Another option is to build a model based on the pre-labeling of the easiest examples. The harder examples can be provided to the model at a later stage and be evaluated by a human being to see if the labeling is being done optimally. This strategy, however, only works when there are sufficiently easy examples in the dataset. It is therefore not recommended to use this strategy with a small dataset (Bowker & Star, 2000).

Rule-based systems are not the only type of systems that can classify text. It is also possible to create a machine learning-based system or to build a hybrid system. To get a better understanding of how text classification is done, the study reviews several scholars which used text classification for a variety of purposes in the following three paragraphs.

Pestian et al. (2016) developed methods of natural language processing that can conduct a text classification on suicide notes to differentiate between genuine, and elicited suicide notes. The suicide notes were binary labeled in elicited or genuine by eleven mental health professionals and 31 psychiatric trainees. Their decisions were compared with nine machine learning algorithms. Some of the relevant features were word count, vowel spacing, and prosody. The result was that the trainees accurately classified 49 percent of the letters, the professionals 63 percent, and the best machine learning algorithm 78 percent of the letters. Based on this study it can be concluded that NLP has the potential to classify notes based on the text.

Amrit et al. (2017) created algorithms that can identify child abuse based on structured notes and unstructured text data. Data of a public health organization was used which included notes about the visits of children. Less than 30 percent of the letters were labeled with abuse or no abuse. A bag-of-words approach was used where the most important feature is the frequency that a word occurs in the texts. Other features that were included were the average amount of characters per consult, lexical diversity, and gender. It was found that machine learning algorithms have the potential to correctly improve the number of registrations that are identified as child abuse. Identifying child abuse is an important step towards reducing its effects and preventing child abuse. The research found that utilizing natural language processing for children's health-related issues supports the idea that this is beneficial and feasible.

Perron et al. (2019) studied whether computer models could be used to gain insights from written summaries from investigations about substance-related issues among families to detect neglect or abuse. The unstructured text was converted to numerical values based on bag-of-words. Afterward, a dictionary and the term-frequency approach were used to analyze the text. For the dictionary approach, domain knowledge is required to create a term list of priory specifications. This is not necessary for the term-frequency approach as the frequencies are computed based on a standard pre-determined formula. An accuracy score of over 90 percent was achieved by a set of algorithms. The authors found that expert human reviewer ratings were interchangeable by computational algorithms. Therefore, the researchers concluded that NLP is an efficient, and cost-effective solution to extract meaningful insights on the topic of neglect or abuse in child welfare.

Classification is an example of a supervised learning technique in which a computer learns from its input data and uses this learning to classify new observations (Maglogiannis, 2007). Support vector machines, K-nearest neighbors, and naïve Bayes are examples of supervised machine learning models that are claimed to perform quite well in the classification of binary labeled text due to its simplicity and generally high performance (Colas & Brazdil, 2006; Khan et al., 2010).

3.4 Machine learning

Machine learning is a field of study where algorithms learn from data. The main idea is to predict a class (output) based on a previously unseen feature vector (input). Learning can be defined as the progressive improvement in performance on a certain task.
The most common form of machine learning is supervised learning, where example input-output pairs are necessary for the learning task. In contrast, unsupervised learning does not require labeled data to conduct unsupervised learning, as the goal of unsupervised learning is to find hidden structures from unlabeled data. Semi-supervised learning is another type of machine learning that also makes use of unlabeled data for training. Typically, semi-supervised learning used a small amount of labeled data and a large amount of unlabeled data (Kubat, 2017; Zhu, 2005).

In this study, supervised learning is used due to the availability of labeled data (Lison, 2015). Supervised learning techniques that can be used for the task of labeling neglect are regression and classification. In regression, the relationships between a dependent variable and one or more independent variables would be estimated. Classification would be the identification of categories based on certain features (Kirkedal et al., 2019; Kotov, 2017). In this study, classification is preferred over regression as the available output is already classified in the following classes: yes (‘Ja’), and no (‘Nej’).

Various machine learning algorithms that can binary classify letters on neglect are discussed in the following subsubsections, to better understand the differences in intuitions.

3.4.1 Support vector machine

A support vector machine (SVM) is a separating hyperplane that can analyze data used for classification. The input data should be marked as belonging to one or the other category. The SVM training algorithm is a non-probabilistic binary linear classifier which builds a model that assigns new inputs to one of the two categories (Mavroforakis & Theodoridis, 2006). An example is demonstrated in Figure 2, where the black triangles could be letters that are about neglect, and the white squares could depict letters that are not about neglect. The uninterrupted line between the squares and triangles is the separating hyperplane. When a new letter comes in it is either predicted to be about neglect or not depending on which side of the separating hyperplane the item is being positioned.

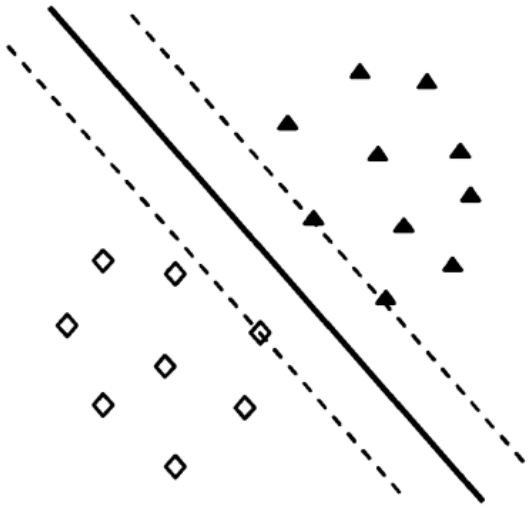

Figure 2. Separating hyperplane in SVM (Mavroforakis & Theodoridis, 2006).

3.4.2 K-nearest neighbors

K-nearest neighbors (KNN) is another algorithm for classification tasks (but it can also be used for regression tasks). KNN can classify a data point to the class it is closest to in the training dataset. The distance between the point that needs to be classified and all other points in the training dataset is computed, this can be done with many distance metrics. For continuous variables, the Euclidean distance is of common use. The point is classified as the most common class in the k-nearest points. It is important to take into account that the KNN algorithm will provide different results based on the chosen value of the number of nearest neighbors (K) (Jiang et al., 2012).

3.4.3 Naïve Bayes

Naïve Bayes (NB) is a classification method that assigns a probability to every possible value in the target range. The resulting distribution is then condensed into a single prediction. Naïve Bayes is based on Bayes rule stated in Equation 1.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

Naïve Bayes assumes that all variables/features in the feature vector are independent, and it is a high bias/low variance learner (Caruana & Niculescu-Mizil, 2006; Jiang et al., 2012; Mandal & Sen, 2014; Tan & Zhang, 2008).

3.4.4 Logistic regression

Logistic regression (LR) is a classification algorithm used to assign observations to a discrete set of classes (e.g. neglect or no neglect). It is a statistical model which uses a logistic function to model a binary dependent variable. Logistic regression transforms its output using the logistic sigmoid function to return a probability value. If the probability is above 0.5 it is rounded off to '1' and if it is below 0.5 it is rounded to '0'. The threshold value is thus 0.5, as displayed in Figure 3 (Kleinbaum et al., 2002; Wright, 1995).

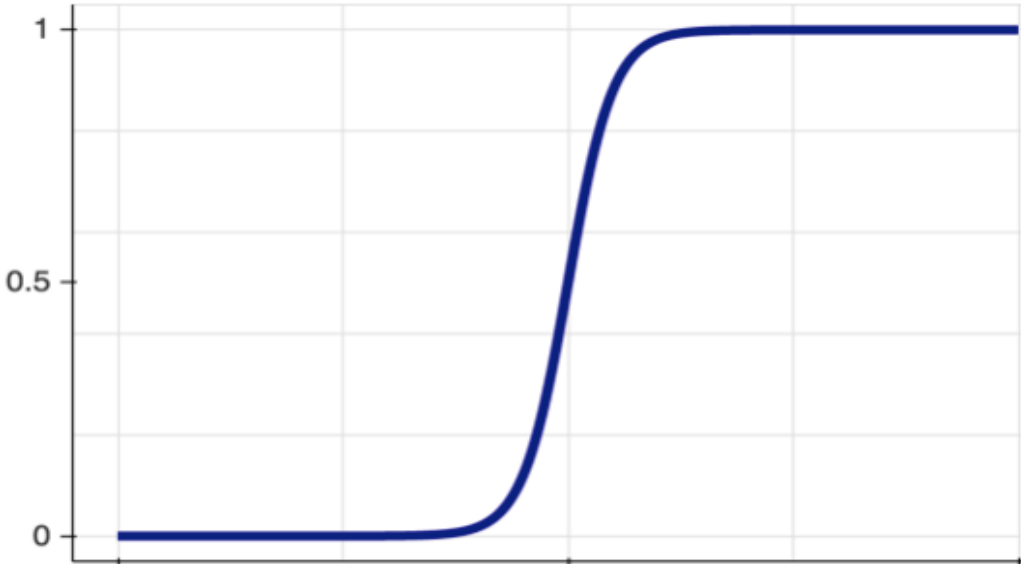

Figure 3. Logistic regression (Kleinbaum et al., 2002).

3.4.5 Decision tree

Decision tree (DT) is a popular classification technique that represents classification rules in a tree form. The goal of DTs is to create a decision tree from training data to correctly determine the labels for new examples. DTs are often preferred over other machine learning algorithms because of the simplicity, comprehensibility to uncover data structure, and their classification speed (Ludwig et al., 2018). Zacharis (2018) believes that DTs are not only good in performance prediction, but also at providing rich information about feature importance.

Decision trees are also low bias/high variance learners, meaning that small changes in the training data can have a large impact on the tree being produced (Ludwig et al., 2018). Zacharis (2018) explains that the simplicity of decision trees often comes at the cost of decreased performance. Therefore, the model performance and model understandability trade-off should be considered.

3.4.6 Random forest

One of the disadvantages of decision trees is that they are highly sensitive to the data that they are trained on, where minor changes to the training set can already result in significant changes in the tree structures. Random forest (RF), or random decision forest, can tackle this disadvantage by allowing individual trees to randomly sample from the dataset with replacement (Pal, 2005). This results in a (large) number of individual DTs that are structured differently and operate together as an ensemble. In other words, random forests can correct for the shortcoming of decision trees to overfit the training data (Liaw & Wiener, 2002). The difference between a DT and a RF is illustrated in Figure 4. The RF algorithm relies on various decision trees that are all trained slightly differently; all of them are taken into consideration for the final classification (Silipo, 2019).

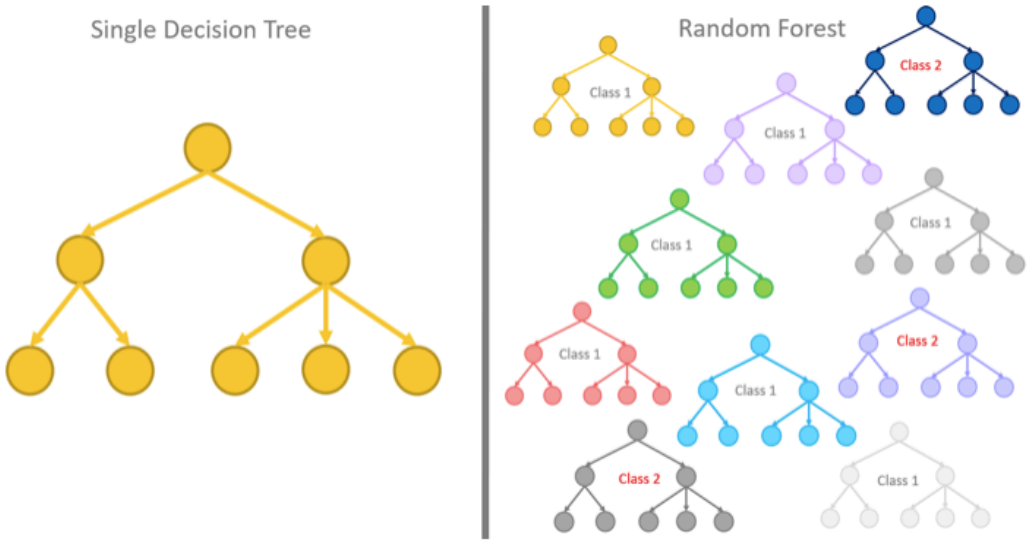

Figure 4. Single decision tree and random forest (Silipo, 2019).

3.4.7 Neural networks

Nowadays, deep learning methods are becoming more important to address various natural language processing tasks. Yin et al. (2017) even claim that deep neural network (DNN) revolutionized the field of natural language processing. DNNs are part of the larger set of machine learning methods (see Figure 1) which are based on artificial neural networks (ANN). Where ANN can be defined as a biologically-inspired programming paradigm that enables computers to learn from observed data. They gain knowledge by detecting relationships and patterns in data, and they learn by considering examples, instead of by programming (Agatonovic-Kustrin & Beresford, 2000). ANNs can reward weights that support correct predictions and punish weights that lead to incorrect predictions (corrective feedback loop). When a non-linear step is available in the process, ANNs can detect all kinds of interaction between independent variables, without having any doubt, even in complex nonlinear relationships between the predictor variables (Shahiri et al., 2015).

A DNN is an ANN with various hidden layers of units between the input and output layers. DNN employs several processing layers to learn hierarchical representations of data. DNNs proved to be successful in tackling difficult learning problems in several domains including natural language processing. Also, DNN is becoming more feasible for people to use these models due to its open-source libraries and increased access to high-performance computing resources (Gong et al., 2019; Lopez & Kalita, 2017; Tai et al., 2015; Yin et al., 2017).

The two main types of DNN architectures are Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). Both are largely researched to tackle several NLP tasks (Yin et al., 2017).

Convolutional neural network

A CNN is quite similar to a traditional neural network, since it is also made up of neurons that have weights and biases (Lopez & Kalita, 2017; Yin et al., 2017). The main difference between CNN and NN is the number of layers. In ordinary NNs every input neuron is in connection to every output neuron in the next layer. However, in CNNs there are only a few layers of convolutions with nonlinear activation functions connected to the results. This structure forms local connections, where every input region is linked to a neuron in the output. Every layer applies many different filters and combines their outcomes (Lopez & Kalita, 2017). CNN is often rated as a good method to extract position-invariant features. However, it is mainly used in computer vision (Yin et al., 2017).

Recurrent neural network

In contrast to CNNs, RNNs are good at using sequential information, or to model units in a sequence of data. RNNs can handle time-series data (Lopez & Kalita, 2017; Yin et al., 2017). The structure of a simple recurrent neural network is demonstrated in Figure 5. A is the update rule (which is applied to the previous outputs), x_t indicates the time steps, and h_t stands for the hidden state vector for each t (Goodfellow et al., 2016). In ordinary NNs inputs are independent of each other, but this often leads to bad performance. Specifically, in the field of text analysis and in human language, words are related. Thus, it is not advisable to look at a single word before looking at the word(s) that came before it.

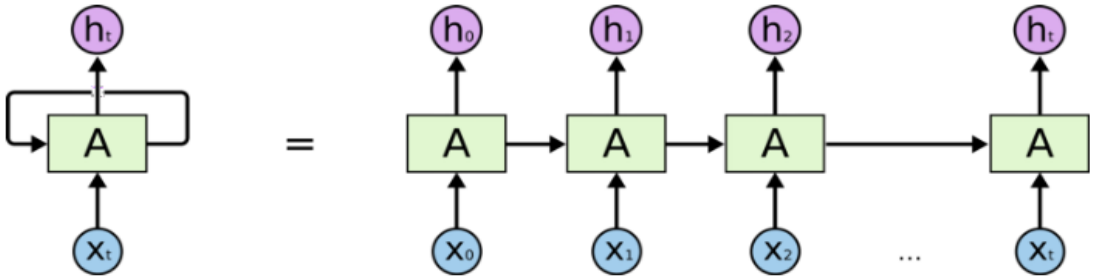

Figure 5. Standard recurrent neural network (Goodfellow et al., 2016).

RNNs are recurrent as they carry out the same task for all the elements in a sequence. RNNs are thus capable to remember what was calculated before. In theory, RNNs can remember information in long sequences. However, in practice, RNNs are only capable of looking back at a few steps (Lopez & Kalita, 2017).

Gated recurrent unit and long-short term memory

The two famous recurrent units used in RNN are gated recurrent unit (GRU) and long-short term memory (LSTM) recurrent unit. Both aim at effectively tracking long-term dependencies.

LSTM and GRU are designed to adaptively update or reset its memory content. The LSTM does this via input, forget, and output gates. The input gate controls the amount of the new state that should be kept, the forget gate regulates the amount of the current memory to delete, and the output gate controls the amount of the cell state that has to be displayed to the next layers of the network.

The GRU only makes use of a reset and upgrade gate. The reset gate is located between the previous recurrent unit and the next candidate unit to forget the previous state. The update gate determines the amount of the candidate unit/activation that should be used in the cell state update. The GRU, unlike the LSTM, displays its complete memory content every time step, and it finds a balance between the old and the new memory content by only making use of leaky integration, meaning that a small amount of the input gradually leaks over time (Chung et al., 2015; Morchid, 2018).

3.4.8 Ethical machine learning and biases

The results by machine learning algorithms can have consequences for individuals and groups, especially when a decision-making process is automated. This is because factors, such as explainability and predictability, are lost when automating human cognitive tasks (Bolander, 2019). As such, algorithmic decisions could be unfairly biased against certain subpopulations based on gender or age (Anderson & Anderson, 2011; Kusner et al., 2018; O'Connor & Conway, 2016). A substantial critique toward algorithmic decisions is the fact that these decisions are hard to understand because machine learning systems cannot explain their behavior and reasoning as human beings can (Bolander, 2019).

Despite the potential of machine learning techniques to simulate aspects of human cognition, fundamental differences between machines and human beings exist. These include different capabilities, skills, advantages, and weaknesses. Some of the tasks that have proven to be complex for human beings have turned out to be simple for algorithms, and vice versa (Bolander, 2019).

Information bias, or misclassification bias, is a type of bias that arises from measurement error. Specifically, the classification error occurs when a participant of the study is placed in the category or population subgroup due to a measurement or observational mistake. When this occurs, the relationship between exposure and outcome is distorted. Usually, misclassification of exposure can introduce significantly more bias into a study than misclassification of outcome. There are two (main) types of misclassification: differential and non-differential. A differential classification error occurs when the bias depends on other components. In contrast, a non-differential classification does not depend on the values of other variables. Differential (or non-random) misclassification bias occurs when the information errors or bias differs between groups (e.g. the group of children who experience neglect or in the group of children who do not experience neglect). Non-differential (or random) misclassification bias occurs when the information is incorrect but the same across groups (e.g. both for children that experience neglect and that do not experience neglect). It happens when exposure is unrelated to other components, or when a problem is unrelated to other variables (including exposure). The direction of bias introduced by non-differential misclassification errors is usually towards the null value of the evaluated parameter. Non-differential misclassification bias does not suppress or inflate estimates of effects but rather dilutes the exposure effect (towards the null) (Bowker & Star, 2000; Flegal et al., 1991; Patten, 2015).

Machine learning algorithms should take into account that the initial data may be biased to avoid perpetuating or creating discriminatory practices. A way to tackle this is by developing a framework for modeling fairness using tools from causal inference. Where counterfactual fairness captures the intuition that a decision is fair towards an individual if it is the same in (1) the actual world, and (2) a counterfactual world where the person belonged to a different demographic group (Kusner et al., 2018). Therefore, a counterfactual fairness framework can solve a large variety of fairness modeling problems. As such a framework can propose a machine learning algorithm that can take into account various social biases that may arise towards individuals based on ethically sensitive attributes and compensate for these biases effectively (Amrit et al., 2017; Kusner et al., 2018; O'Connor & Conway, 2016; Schnoebelen, 2017).

4. Methodology

This research is based on the Danish children's welfare organization called Børns Vilkår. It is an institution for children and young people to write about their problems and/or concerns, to read about what others have asked, and to receive answers on their questions. The methodology is structured in a way that is presented in Figure 6. First, the preliminary exploratory study is described, this includes a description of the research problem, research objectives, and the research questions. Later, the steps taken during the data analysis process are presented, this includes a description of the data set (feature(s) and target), pre-processing of the data, and machine learning algorithms. Lastly, it is presented how the proposed machine learning algorithms are being evaluated, and how the discussion and conclusion of this research are formed (Bach et al., 2016; Sindahl et al., 2018; Watson, 1994).

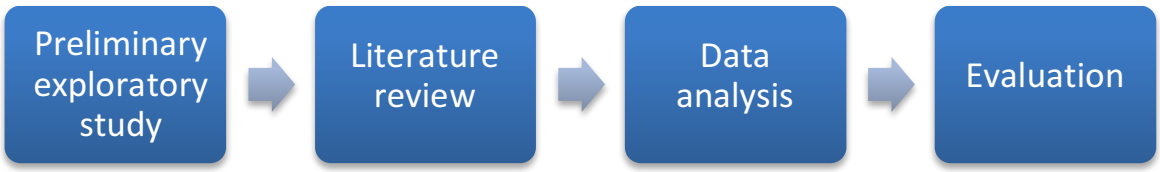

Figure 6. Structure of the methodology of this research.

4.1 Preliminary exploratory study

Because the Danish child welfare organization (Børns Vilkår) can take more than one month to answer incoming letters from children in need, this study is aimed at understanding this problem better, formulate objectives, and come up with relevant research questions.

The preliminary exploratory study, an initial exploration of the problem (Bach et al., 2016; Sindahl et al., 2018; Watson, 1994), consists of two parts: interviewing machine learning experts at IT consultancy firms that work together with Børns Vilkår and interviewing digital experts at Børns Vilkår.

First, the issues of Børns Vilkår were discussed during various interviews with Frederik Hestvang, an analytics specialist at the Danish digital media agency IIH Nordic in Copenhagen. He has experience in working together with Børns Vilkår on opportunities and challenges in digitalization. The initial problem was discussed with Mr. Hestvang, potential solutions were offered, and questions for Børns Vilkår were defined. Meanwhile, there was contact with data science consultants of ITelligence who just started working on a tool to optimize the chat feature of BørneTelefonen, a Danish child helpline which is part of the Children's Welfare organization Børns Vilkår, with the use of natural language processing. The contact was conducted over email, interviews, and through a transformation lab of ITelligence.

Second, the issue and potential solutions were further discussed with Kathrine Flindt, Digital business developer, and Charlotte Smerup, digital consultant at Børns Vilkår. This input was used in the final stage of defining the research problem, objectives, and the research questions of this study.

During the preliminary exploratory study, it became evident that the practical problem of Børns Vilkår is that children who experience neglect have to wait too long before they get a reply to the letter sent to BørneTelefonen. Child neglect is chosen as the prioritization target by Børns Vilkår and can be described as a form of child abuse, which can either be mentally or physically (Butchart et al., 2006). There are several solutions to this problem, however, in this research is assumed that the comprehensive solution to solve the problem regarding long waiting times is a system that can automatically label the incoming letters in a binary fashion. This labeling is based on two criteria, 'Nej' if a letter is not about neglect, and 'Ja' if a letter is about neglect. This labeling should be done automatically in the same way as volunteers of BørneTelefonen would have done it. Therefore, the supervised machine learning algorithm should be trained on labels that are defined by volunteers of BørneTelefonen.

Børns Vilkår values that letters in which children describe that they experience neglect are answered by a volunteer of BørneTelefonen as soon as possible. Still, Børns Vilkår believes that all children deserve an answer to their letters. Therefore, the ideal prioritization model does not select *which* letters should be answered, and which ones should not. Instead, it should prioritize *when* a certain letter is being answered. In the end, all children that write a letter should get a reply from the volunteers of BørneTelefonen.

4.2 Data analysis

4.2.1 Data set – feature(s) and target

The dataset used in this research is provided by Børns Vilkår in the form of a Microsoft Excel spreadsheet, which includes 5664 letters that are collected between July 2018 and September 2019. The dataset was compliant with the European Union General Data Protection Regulation (GDPR), meaning that the data was treated confidentially and the letters were anonymized (GDPR.eu, 2020).

Letters in the database

The data set includes various columns including one column that displays the text of the letters. In this part of the methodology is being explained how the text written by the child is converted to the text of the excel spreadsheet.

On the internet website "bornetelefonen.dk/brevkasse", a child has the option to read letters, write letters, read responses, and write responses (to some of the letters). If a child chooses to write a letter it is being presented with the interface displayed in Figure 7. The official text on this interface is written in Danish, but the text on Figure 7 is translated to English for the readability of the audience of this thesis.

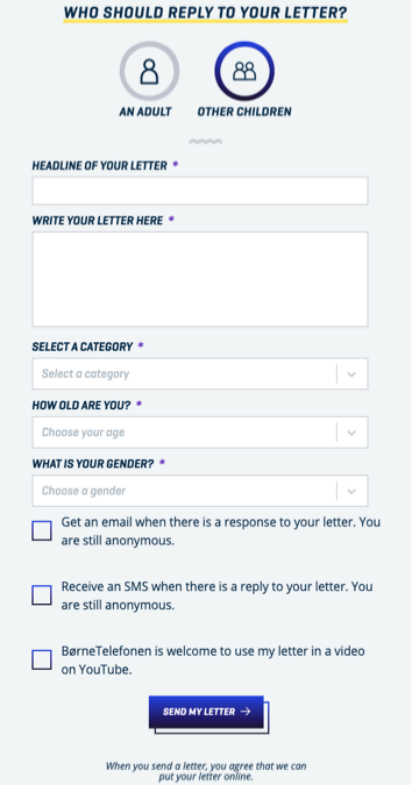

Figure 7. Interface of letter submission (BørneTelefonen, 2019).

As presented in Figure 7, the child can select whether an adult or other children respond to their letter. The children have to write a heading of the letter, the actual letter, a category, the age, and gender. Only letters where the child selected that it wants an adult to reply are being included in the database. Letters that contained personal information of the child (such as CPR number, name, and address) were excluded from the database and thus excluded from the excel file that was sent by Børns Vilkår. The letters in the excel spreadsheet are a one-to-one copy of the letters written on the website of BørneTelefonen. Therefore, the letters in the excel file are original and may include errors such as typos, and grammatical mistakes.

Data-set features

As seen in the previous paragraph, the dataset includes the text, the title, the age, the gender, and the category for each letter. It also includes other features, which are the reply, the waiting time, and the status.

Age ranges from 8 to 23 years old and has an average of almost 14 years old. For gender, there are many options, but mainly the options boy and girl were selected. The majority of the letters, 4264 letters, came from girls, 1193 letters from boys, and 207 letters from people that did not identify themselves as boy (‘Dreng’) or girl (‘Pige’). There were 33 categories that children could pick from. The waiting time varied from 0 to 32 days and is on average 7 days. There are 3 statuses for the letters, they are either private (‘private-svar’), publicly published (‘publish’), or the letter is rejected (‘afvist’) indicating that there is no personal response to the letter. The distribution of these responses is displayed in Figure 8.

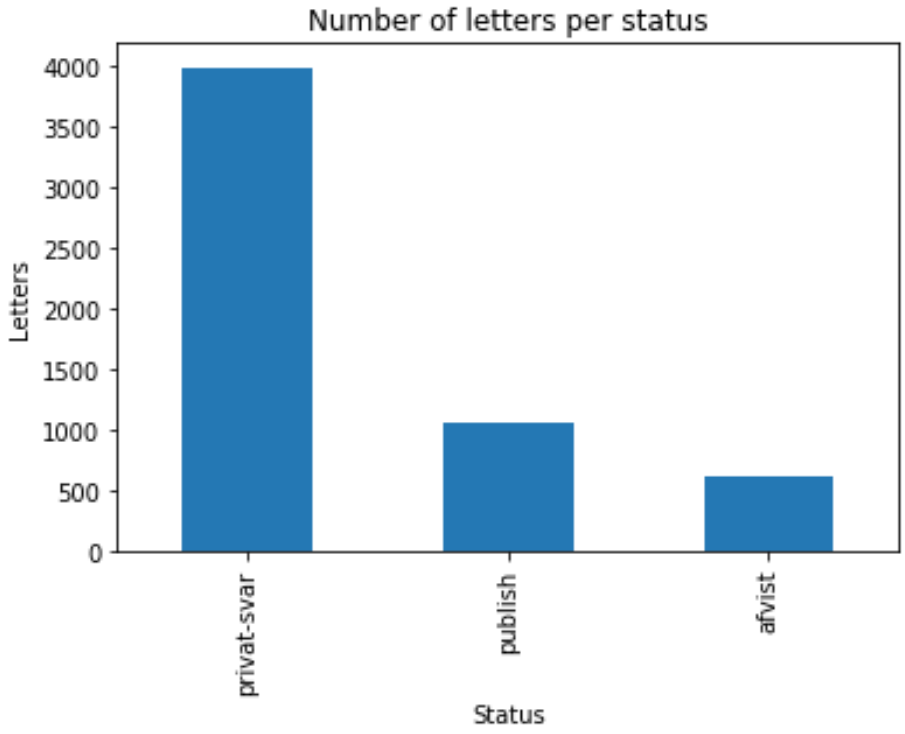

Figure 8. Statuses for the letters: private, publicly published, or rejected.

Usually, all provided features that seem relevant for the analysis are used in the creation of a machine learning algorithm. However, this analysis only uses the text of the letters as input. This decision was made based on two reasons. (1) Generalizability. Every text classification task has at least the input text available, all other features are not always available. Furthermore, the response is not accessible yet when a new letter comes in. Therefore, it is chosen to base the analysis only on the input data, as this would allow us to use the analysis for other applications too. (2) Reduce bias and noise. ITelligence also did research based on the data of Børns Vilkår and used a counterfactual fairness model on the fair prediction in clustering chats Danish children have with volunteers of BørneTelefonen to create an ethical machine learning algorithm. Specifically, they studied at age and gender bias. They found that including age was causing more noise and that gender did not affect the model from clustering gender-specific issues. Considering the similarity of the data and research of ITelligence and this study it is assumed that doing a counterfactual fairness framework would result in similar findings. Therefore, it is decided in this study to do supervised learning only on the text of the chats and leave out the features of age and gender.

Target description

To classify the letters, binary labels on neglect are provided by volunteers of BørneTelefonen. This is done to train the machine learning algorithms to handle incoming letters in the same way as would be done by the volunteers of BørneTelefonen.

When volunteers of BørneTelefonen assess a letter based on neglect they ask themselves whether they would report this to the authorities if they had the contact details of the child. For example, if a child is being physically abused by their alcoholic parents than this is a situation that would be reported to the authorities if the identity of the child was known. Therefore, this is being labeled as 'Ja' (indicating neglect). However, if a child wants to commit suicide due to factors not related to neglect this is being labeled as 'Nej' (indicating no neglect), even though the situation could be urgent. In this research is assumed that the labeling done by the volunteers is correct. It is even used as the ultimate truth in the development of machine learning algorithms.

The dataset has unbalanced classes as about 89 percent of the letters are labeled with 'Nej' (indicating no neglect) and about 11 percent of the letters with 'Ja' (indicating neglect). This difference is being considered by stratifying the labels in the train-test split of 80 and 20 percent, respectively. Stratification means that the train_test_split method of sklearn returns training and test subsets that have the same proportions of class labels as the input dataset.

The percentage of letters that are and are not labeled as neglect are displayed in Figures 9 and 10. This is done for all genders and all age groups. The reason for this is to get a better understanding of the potential biases that the experts who labeled the data might have had.

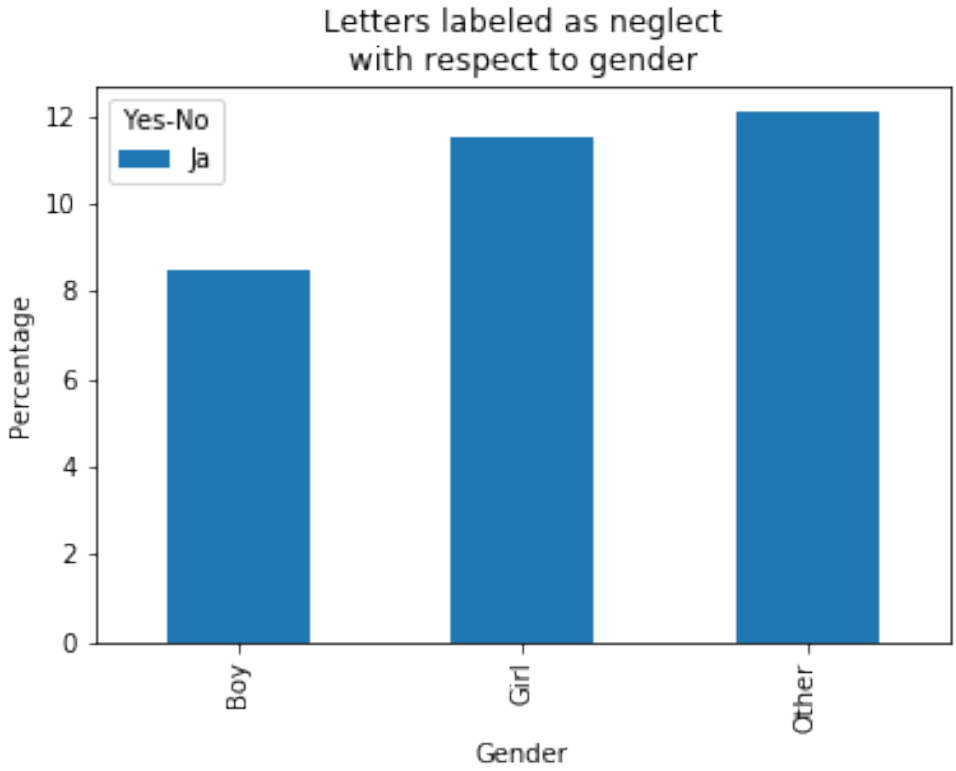

Figure 9. Yes labels per gender category.

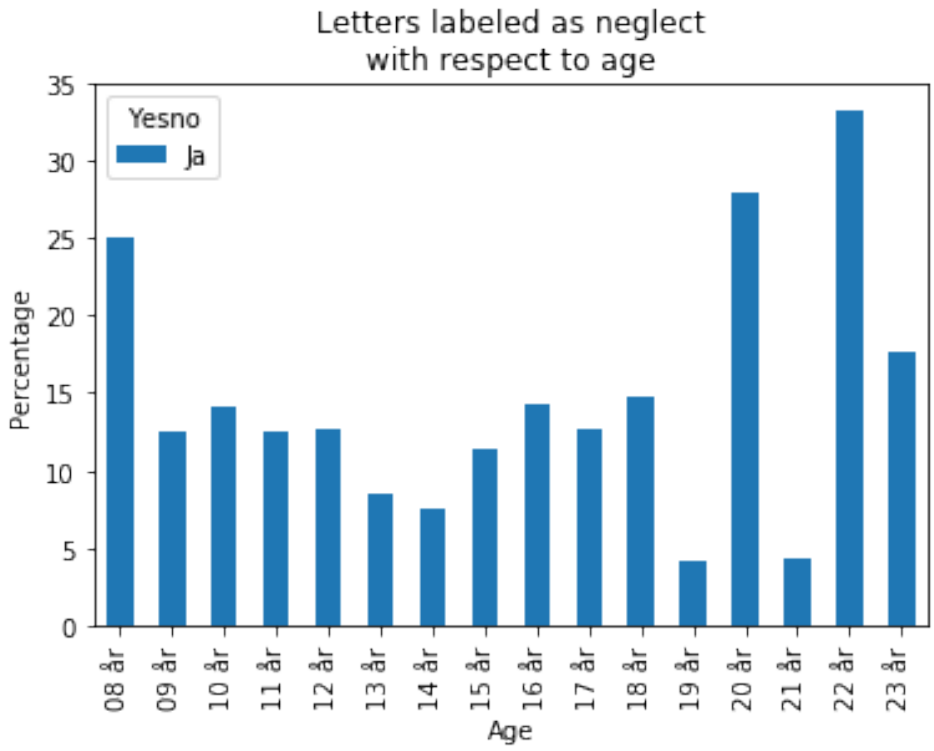

Figure 10. Yes labels per age group.

4.2.2 Preprocessing

After selecting the input feature and target, the data was processed. The natural language processing tasks of standardizing the text, simplifying the text, and extracting information regarding neglect is included in this process. As part of this procedure, Danish stop words were removed. Stop words are a set of commonly used words in any language, examples of stop words are ‘and’, ‘is’, and ‘the’ (Carrell et al., 2017; Derczynski, 2019).

Clean input text

Firstly, the letters that are going to be analyzed are cleaned. Specifically, the following actions are taken with the data: remove all special characters, remove all single characters,
substitute multiple spaces with a single space, convert all letters to lowercase, and apply lemmatization for Danish words (Khan et al., 2010).

Text representation

To represent the text, every row of the dataset is converted to a single document of the corpus. The features depend on the chosen feature creation method. The most common feature creation methods are word count vectors, TF-IDF, and word-embeddings. In this study, TF-IDF is chosen as this is the most common feature creation method for text classification. TF-IDF is a score that represents the relative importance of a term in the document and the entire corpus. TF stands for Term Frequency, and IDF stands for Inverse Document Frequency (Joachims, 1996; Yun-tao et al., 2005). In this study, TfidfVectorizer of sklearn.feature_extraction.text is being used to apply TF-IDF to the letters. Danish stop words are being removed as part of TF-IDF. This is done with the Natural Language Toolkit (NLTK) corpus which includes all kinds of natural language data sets, including a list of Danish stop words.

Split the data

Afterward, the data is being split into an 80/20 split. This means that 80 percent of the letters (of the data set) are used for training the algorithms, and the remaining 20 percent of the letters is used afterward to test the performance of the machine learning algorithms. An 80/20 split is chosen instead of a 90/10 split as a test set of 10 percent seems to be too small considering the size of this data set. A bigger test set, namely 20 percent, is thus preferred. Also, a training set of 80 percent is expected to lead to a sufficient amount of data to train the algorithm. The reason why no 70/30 split is chosen has to do with the fact that there would not be enough data left to train the model and to generate a reasonable outcome (Baesens, 2014; Schoenherr & Speier-Pero, 2015).

4.2.3 Machine learning algorithms

In this study, the following popular natural language processing machine learning algorithms are being compared: support vector machine (SVM), Naïve Bayes (NB), logistic regression (LR), decision tree (DT), random forest (RF), recurrent neural network (RNN) and k-Nearest Neighbors (KNN).

The input data, including the features, are kept the same for the training of all machine learning algorithms to allow for better comparison between the algorithms (Colas & Brazdil, 2006; Khan et al., 2010).

The way the performance of the machine learning algorithms is compared is described in the next subsection.

4.3 Evaluation

The main benefit of having labeled data is that supervised machine learning can be used, where input-output pairs are available for the learning task. This makes it easy to compare different algorithms as it can be assumed that the labels given by volunteers of BørneTelefonen are correct. Algorithms can thus be compared by stating a score which is based on the number of letters that are labeled correctly, and the number of letters that are labeled incorrectly. Using supervised learning also means that the researcher does not need to have Danish language skills or skills in processing children's language. Instead, this is covered by the work (labeling) done by the volunteers of BørneTelefonen.

In the evaluation phase, the performances of the algorithms are being assessed and compared. The performance scores used in this study are explained in the upcoming paragraphs.

In the identification of children experiencing neglect, it is relevant to minimize a type II (false negative) error without increasing the type I (false positive) error significantly. Precision calculates the true positive out of the sum of true positives, and false positives. In this research, precision is of interest because it is the fraction of letters describing neglect issues that have been correctly retrieved over the total number of letters. Recall, or sensitivity, is the fraction of relevant instances (true positives), over the total number of relevant instances (sum of true positives and false negatives). In this study, recall is valuable and could be defined as the fraction of letters about neglect that have been correctly detected over the total number of letters about neglect.

Both precision and recall are relevant for this study, therefore is decided to use both of these performance evaluation metrics when assessing the performance of the machine learning algorithms. Also, it is chosen to select the harmonic average of both evaluation metrics (precision and recall), called the F1 score. The F1 score formula is stated in Equation 2.

$$F1 = \frac{2 * precision * recall}{precision + recall} \quad (2)$$

The F1 score ranges from zero to one, where an F1 score of zero means worst recall and precision, and an F1 score of one means perfect recall and precision.

Afterward, the results of this study are being discussed. This includes the differences in performance in terms of precision, recall, and F1 score of the various machine learning algorithms. The discussion also includes how ethical the optimal algorithm is. Lastly, the conclusion of this study includes suggestions for further research (Kemmis & McTaggart, 2005; Stringer, 2013).

5. Results

In this section, the performances of different machine learning algorithms on the text classification task of neglect are presented and compared. The programming language used in this study is Python 3.7.6. There is an appendix available in which the used packages, libraries, and Python code can be found (the author can be contacted to receive this appendix).

5.1 K-nearest neighbors

Running a KNeighborsClassifier of sklearn neighbors with all default parameters (n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=None, **kwargs) results in the performance which are displayed in Table 1.

Table 1. *Evaluation metrics for the performance of k-nearest neighbors*

Confusion matrix			Precision	Recall	F1 score
1 (TN)	122 (FP)	Yes ('Ja')	0.25	0.01	0.02
3 (FN)	1007 (TP)	No ('Nej')	0.89	1.00	0.94

Afterward, hyperparameter tuning was conducted on the following hyperparameter (a hyperparameter is a parameter whose value is set before the learning process begins): n_neighbors (set to either 3, 5, 7 or 9). With the use of GridSearchCV (with CV=2) of sklearn model_selection the best hyperparameter was found, where GridSearchCV is an algorithm that can do an exhaustive search over specified parameter values for an estimator. The number of neighbors leading to the best performance was 9. This resulted in the following performance.

Table 2. *Evaluation metrics for the performance of KNN after hyperparameter tuning*

Confusion matrix			Precision	Recall	F1 score
1 (TN)	122 (FP)	Yes ('Ja')	1.00	0.01	0.02
0 (FN)	1010 (TP)	No ('Nej')	0.89	1.00	0.94

The confusion matrix is made by assuming that the letters that are correctly classified with no ('Nej') indicating no neglect (according to the labeling of the volunteers of BørneTelefonen) are positive, and the letters classified with yes ('Ja') indicating neglect are negative. True positives are letters that are correctly classified with no ('Nej'), true negatives are letters that are correctly classified with yes ('Ja'), false positives are letters that are said to be not about neglect but they are, and false negatives are letters that are falsely labeled as yes.

When looking at Table 2, it can be seen that almost all letters are being classified to be positive. Precision for yes is calculated as TN/(TN+FN) or (1/(1+0)) which gives a precision score of 1.00. Recall of no is 1.00 as it is calculated by doing TP/(TP+FN) or (1010/(1010+0).

In general, it can be said that the KNN algorithm is performing well incorrectly classifying letters that are not dealing with neglect. This is in line with the expectations of the researchers. Usually, one of the biggest disadvantages of KNN is the fact that KNN does not work well with large datasets (Jiang et al., 2012). However, the data set of this study is small, indicating that this is not a (big) disadvantage in this study. Another big disadvantage of KNN is the fact that it is performing badly when there is high dimensional data. As high dimensional data has a large number of irrelevant attributes which makes the distance function inappropriate and inaccurate (Chinmay, 2019). This is problematic as the dimensions in the features are large in this research, as is typically the case in natural language processing.

The recall score for yes is extremely low. This has to do with the fact that classes that are represented by the small sample size ('Ja') are overwhelmed by a large number of prototypes of the dominated group ('Nej'). To tackle this problem, Boiculese et al. (2013) recommend using a method of weighting the prototypes for each class of the k-nearest neighbors to cope with the uneven distribution of data. The proposed method is supposed to increase the classification rate in terms of recall measure.

5.2 Support vector machine

Running a LinearSVC of sklearn svm with all default parameters (penalty='l2', loss='squared_hinge', dual=True, tol=0.0001, C=1.0, multi_class='ovr', fit_intercept=True, intercept_scaling=1, class_weight=None, verbose=0, random_state=None, max_iter=1000) results in the performance which is presented in Table 3.

Table 3. *Evaluation metrics for the performance of support vector machine*

Confusion matrix			Precision	Recall	F1 score
19 (TN)	104 (FP)	Yes ('Ja')	0.44	0.15	0.23
24 (FN)	986 (TP)	No ('Nej')	0.90	0.98	0.94

When comparing the different loss functions 'squared_hinge' and 'hinge' it could be seen that when the loss was set to 'hinge' all values were predicted to be Nej. Resulting in a recall score for 'Nej' of 1.00, and a score of 0.00 for the precision, recall and F1 score of 'ja'.

An advantage of SVM is that it works well with unstructured and semi-structured data like text. This can be seen when looking at the performance of the SVM in Table 3. Also, SVM algorithms have generalizability in practice, indicating that the risk of over-fitting is reduced. Furthermore, SVM scales relatively well to high dimensional data which is beneficial for natural language processing as the text is represented in many dimensions (Colas & Brazdil, 2006; Mavroforakis & Theodoridis, 2006).

A disadvantage of SVM is that it is hard to visualize the impact of some of its hyperparameters, such as C, which makes it difficult to fine-tune them. Also, it is generally difficult to understand the final model making. This makes it difficult to optimize its performance based on intuitive logic related to the research topic (Colas & Brazdil, 2006; Mavroforakis & Theodoridis, 2006).

Unfortunately, SVM is also struggling to perform well with imbalanced datasets (Colas & Brazdil, 2006). This indicates the large performance difference between yes and no. It is expected that recall of yes is lower than of no as there are more positive examples ('Nej') than negative ('ja').

5.3 Naïve Bayes

Running a GaussianNB of sklearn naive_Bayes with all default parameters (priors=None, var_smoothing=1e-09) results in the following performance.

Table 4. *Evaluation metrics for the performance of naïve Bayes*

Confusion matrix			Precision	Recall	F1 score
75 (TN)	48 (FP)	Yes ('Ja')	0.16	0.61	0.25
402 (FN)	608 (TP)	No ('Nej')	0.93	0.60	0.73

Naïve Bayes is one of the few algorithms that can work with small-scale data, which is beneficial for this study due to the small number of training samples in the data set. Also, it is suitable for incremental training, indicating that it can train new samples in real-time (Granik & Mesyura, 2017). This means that its performance can improve after implementation at BørneTelefonen when more data is entering the model.

Naïve Bayes is insensitive to irrelevant features, therefore some scholars claim that it is best suited for text classification problems that have many (irrelevant) features (Frank & Bouckaert, 2006; Granik & Mesyura, 2017). Furthermore, the conditional independence assumption of naïve Bayes is often seen as one of the biggest limitations. This means that naïve Bayes assumes that all predictors are independent of each other. This is especially not the case in texts where words are related to one another.

The performance scores for yes are relatively high in comparison to other algorithms. This has to do with the fact that naïve Bayes is better at dealing with unbalanced classes than other algorithms. However, this comes at a cost as can be seen in the performance measures of no which is lower than for other algorithms (Frank & Bouckaert, 2006).

5.4 Logistic regression

Running a LogisticRegression of sklearn linear_model with all default parameters (penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None) results in the following performance.

Table 5. *Evaluation metrics for the performance of logistic regression*

Confusion matrix			Precision	Recall	F1 score
3 (TN)	120 (FP)	Yes ('Ja')	0.43	0.02	0.05
4 (FN)	1006 (TP)	No ('Nej')	0.89	1.00	0.94

Afterward, hyperparameter tuning was conducted with the use of GridSearchCV (with CV=2) on the following hyperparameters: class_weight ('None', or 'balanced'), and solver ('liblinear', or 'lbfgs'). According to sklearn, 'liblinear' is a good choice for smaller data sets. Therefore, it was expected that 'liblinear' would outperform 'lbfgs'. However, this was not the case. Also, the best performing class_weight was None. After hyperparameter tuning was thus observed that the default parameters were performing best (based on the evaluated hyperparameters).

Logistic regression is one of the most used machine learning algorithms for binary classification. This is mainly due to the number of advantages of logistic regression including its efficiency, the simplicity to implement, the interpretability, the fact that it does not require high computational power, and that scaling of the features does not have to be done. Another benefit of logistic regression is that it does not need hyperparameter tuning as could be seen above (Kleinbaum et al., 2002; Wright, 1995).

A disadvantage of logistic regression is that it cannot handle a large number of categorical variables and features. Another drawback of using logistic regression is that the performance of the algorithm is low when there are independent variables that are not correlated to the target variable and that are strongly correlated to one another. Furthermore, there is the risk of overfitting (Kleinbaum et al., 2002; Schoenherr & Speier-Pero, 2015).

5.5 Decision tree

Running a DecisionTreeClassifier of sklearn tree with all default parameters (criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, presort='deprecated', ccp_alpha=0.0) results in the following performance.

Table 6. *Evaluation metrics for the performance of a decision tree*

Confusion matrix			Precision	Recall	F1 score
24 (TN)	99 (FP)	Yes (‘Ja’)	0.20	0.20	0.20
95 (FN)	915 (TP)	No (‘Nej’)	0.90	0.91	0.90

Afterwards, hyperparameter tuning was conducted with the use of GridSearchCV (with CV=2) on the following hyperparameters: criterion ('gini', or 'entropy'), splitter ('best', 'random'), and class_weight ('None', 'balanced'). After hyperparameter tuning was thus observed that the default parameters were performing best (based on the evaluated hyperparameters).

In contrast to logistic regression, decision tree works well when the variables are correlated. This is because a decision tree works by finding the interactions between variables. Also, it required less effort for data preparation in the pre-processing phase in comparison to many other machine learning algorithms. Furthermore, it does not require normalization and scaling of the data. Another benefit is that it is very intuitive and easy to explain to stakeholders such as Børns Vilkår (Ludwig et al., 2018).

One of the biggest disadvantages of a decision tree is that a small change in the data could have a large effect on the structure of the decision tree causing instability. Furthermore, the building process is quite complex and computationally expensive (in terms of time and memory) (Ludwig et al., 2018).

5.6 Random forest

Running a RandomForestClassifier of sklearn ensemble with all default parameters (n_estimators=100, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None) results in the following performance.

Table 7. *Evaluation metrics for the performance of random forest*

Confusion matrix			Precision	Recall	F1 score
0 (TN)	123 (FP)	Yes ('Ja')	0.00	0.00	0.00
3 (FN)	1007 (TP)	No ('Nej')	0.89	1.00	0.94

Afterward, hyperparameter tuning was conducted with the use of GridSearchCV (with CV=2) on the following hyperparameters: criterion ('gini', or 'entropy'), and class_weight ('None', 'balanced'). After hyperparameter tuning was observed that 'entropy' and 'balanced' are performing best (based on the evaluated hyperparameters). Still, the confusion matrix and performance scores looked the same as Table 7 indicated that hyperparameter tuning did not result in significant changes in performance.

Random forest is an algorithm that is based on the bagging algorithm and uses the ensemble learning technique. It creates as many trees on the subset of the data and combines the output of all the trees. In this way, it reduces overfitting problems in decision trees and also reduces the variance and therefore it should theoretically have a better performance than a decision tree. When comparing the results in Table 5 and Table 6 it can be seen that recall and F1 score of the random forest are higher than of the decision tree. It might seem surprising that the other values are performing less well. However, this result is not extremely shocking taking into account unreliability. The findings of the decision tree are a lot more unreliable than of the random forest as the random forest is drawing its results from many trees increasing its reliability (Liaw & Wiener, 2002; Pal, 2005; Silipo, 2019).

A disadvantage of a random forest is that it is complex to understand as (in this case) 100 trees are drawn and afterward the outputs are combined. It of course, also requires a lot more time, computational power, and resources than building a decision tree (which is already perceived as computationally expensive in itself) (Liaw & Wiener, 2002; Pal, 2005; Silipo, 2019).

5.7 Recurrent neural network

A bidirectional model of keras layers with LSTM with parameters (return_sequences=True, input_shape=(n_timesteps, 1, loss='binary_crossentropy', optimizer='adam', metrics=[f1_m, recall_m, precision_m]) is made, where f1_m, recall_m, precision_m are defined by the respective formulas. Specifically, a bidirectional LSTM layer with output embedding dimension of size 64 is made, after an LSTM layer with output embedding dimension size 32 is created, then a dropout layer where alpha is set to 0.2, lastly a dense layer is added with 10 units. Furthermore, batch_size is set to 64, epochs to 10, and hidden_dims to 250. This recurrent neural network results in the following performance.

Table 8. *Evaluation metrics for the performance of a recurrent neural network*

Confusion matrix			Precision	Recall	F1 score
0 (TN)	123 (FP)	Yes ('Ja')	0.00	0.00	0.00
3 (FN)	1007 (TP)	No ('Nej')	0.89	1.00	0.94

It was chosen to conduct bidirectional LSTMs instead of traditional LSTMs as bidirectional LSTMs are an extension of traditional LSTMs that can improve model performance on sequence classification tasks, such as text classification problems (Greff et al., 2017; Huang et al., 2015).

The created model is very simplistic and assumes that all letters are not about neglect, as displayed Table 7. This could have been expected as (recurrent) neural networks have difficulties handling a low number of input data. Especially, when the input data varies as much as is the case in the data set that is used in this study, e.g. as the letters are of different lengths, different language is used, and linguistic errors are made (Greff et al., 2017; Huang et al., 2015).

5.8 Summary of the results section

In this subsection, the results provided in the previous paragraphs are combined and compared. An overview of the performance of the algorithms is presented Table 9. The different performance scores are compared to one another in the following subsubsections.

Table 9. *Evaluation metrics for the performance of various machine learning algorithms*

		Precision	Recall	F1 score
KNN	Yes	1.00	0.01	0.02
	No	0.89	1.00	0.94
SVM	Yes	0.44	0.15	0.23
	No	0.90	0.98	0.94
NB	Yes	0.16	0.61	0.25
	No	0.93	0.60	0.73
LR	Yes	0.43	0.02	0.05
	No	0.89	1.00	0.94
DT	Yes	0.20	0.20	0.20
	No	0.90	0.91	0.90
RF	Yes	0.00	0.00	0.00
	No	0.89	1.00	0.94
RNN	Yes	0.00	0.00	0.00
	No	0.89	1.00	0.94

5.8.1 Difference in precision for the various machine learning algorithms

Table 9 shows that the precision in terms of no (‘Nej’) is highest (with 93 percent) for naïve Bayes. This is followed by support vector machine and decision tree (with 90 percent). Lastly, all other machine learning algorithms have a slightly lower precision, namely of 89 percent. In contrast, the differences in precision in terms of yes (‘Ja’) are a lot bigger. K-nearest neighbor is the only algorithm that has a precision of 100 percent, this is followed by support vector machine that has a precision of 44 percent, linear regression of 43 percent, decision tree of 20 percent, naïve Bayes of 16 percent, and the most complicated algorithms, random forest and recurrent neural networks with 0 percent.

5.8.2 Difference in recall for the various machine learning algorithms

Recall in terms of no is optimal (100 percent) for k-nearest neighbors, linear regression, random forest, and recurrent neural network, as could be seen in Table 9. This is followed by support vector machine which has a performance of 98 percent, and decision tree which has a recall score of 91 percent. The algorithm with the lowest recall in terms of no is naïve Bayes with 60 percent. However, this algorithm has the highest recall score (61 percent). Unsurprisingly, this is followed by decision tree (with 20 percent) which was the second worse performing algorithm in terms of no. Support vector machine has a recall of 15 percent in terms of yes, and k-nearest neighbor has a performance of 1 percent. Again, the two most complex algorithms (random forest and recurrent neural network) have a performance in terms of recall of 0 percent for yes.

5.8.3 Difference in F1 score for the various machine learning algorithms

F1 score can take into account the precision and recall scores. As displayed in Table 9, almost all algorithms give an F1 score of 94 percent in terms of no. Only, decision tree and naïve Bayes have a lower F1 score, 90 percent and 73 percent respectively. In terms of yes the highest F1 score is given by naïve Bayes of 25 percent. This is followed by support vector machine with 23 percent, decision tree with 20 percent, linear regression with 5 percent, and k-nearest neighbors with an F1 score of 2 percent. The precision and recall scores of random forest and recurrent neural network were 0 percent, therefore these F1 scores are also 0 percent.

5.8.4 Differences in performance for the various machine learning algorithms

Precision, recall, and F1 score are three different ways of evaluating the performance of an algorithm. Where high precision signifies that an algorithm returns substantially more relevant results than irrelevant ones, while high recall means that an algorithm returned most of the relevant results, and the F1 score is the weighted average of the precision and recall. In this research, it might seem that precision is not that relevant, as it does not matter that a child that wrote a letter which is not about neglect is classified as being about neglect. Therefore, one might claim that this research should strive to get a high recall score, regardless of the impact it has on the precision score. Still, it is important to be aware of the negative consequences these false positives have on the true positives. If too many letters are classified as neglect this means that the children that wrote a letter about neglect have a longer waiting time than needed. Also, a high number of false positives could make specialists less attentive and potentially reluctant about these letters,

resulting in the fact that they might get annoyed and stop prioritizing them. Missing out on letters that were about neglect (false negatives) is not ideal, but not highly concerning either. These letters do not 'skip the queue', and therefore they would be treated at approximately the same time as without the use of a classification algorithm. When evaluating the differences between recall and precision it seems that recall is more important to most people. Despite, it seems important to not completely ignore the performance of precision, or F1 score (Boiculese et al., 2013; Braasch, 2002; Mertens, 2014).

The performance scores precision, recall, and F1 score are calculated for both values of the binary classification (yes and no). Logically, the performance scores for no are higher than for yes as the algorithm was trained on a dataset that included more letters that were not about neglect than letters that were about neglect. As the goal of this study is to correctly predict letters that are about neglect it is more important to correctly predict yes than no, still it is important to not entirely disregard the performance on the letters classified as no. When looking at the precision scores in Table 9 it might seem that KNN is the best performing algorithm as the performance on yes is optimal, still, the performance on the no letters (with 89 percent) is also relatively high. Despite this, it is recommended to exclude this value as the jump in performance of yes from 25 percent to 100 percent due to hyperparameter tuning seems too positive. It seems that this is a case of overfitting which could be caused by the small dataset that is used in this study. The two algorithms with realistic findings that seem to be performing best are support vector machine (with yes precision of 44 percent and no precision of 90 percent) and logistic regression (with yes precision of 43 percent and no precision of 89 percent). Both, the performance of yes and no are one percentage point higher than for logistic regression. This difference is, however, rather small. Therefore, it is decided to compare the differences in recall scores of these two algorithms equivalently. The recall performance of no for SVM is 98 percent and for LR 100 percent. This difference is also small and not extremely relevant. The difference in the performance of yes, however, is a lot bigger, namely 15 percent for SVM and only 2 percent for LR. Therefore, it seems that SVM is the best performing algorithm in this research. This can also be seen when looking at the F1 score. The F1 score of no of 94 percent belongs among the highest, and the F1 score of no of 23 percent is only beaten by naïve Bayes by 2 percentage points, as presented in Table 9. It thus seems that support vector machine algorithms (with default parameters) is working best for this study as it can best balance the benefits of correctly predicting letters not to be about neglect, and predicting letters to be about neglect.

6. Discussion

The main components that are discussed in this section are technical, social, cultural, and ethical implications and limitations related to the use of different machine learning algorithms to classify letters sent to BørneTelefonen. These components are valuable to get a better insight into the meaning of the results which were stated in the previous section. Subsection 6.1 is devoted to discussing topics that are related to the first sub research question of this study *('How accurate are machine learning algorithms when classifying incoming messages in comparison to human coders?').* Subsections 6.2, 6.3, and 6.4 are dedicated to addressing issues that are related to the second sub question of this research *('What are possible technical, social, cultural, and ethical limitations when using machine learning algorithms to classify incoming messages?').* These findings are compared in the conclusion section to answer the main research question *('To what extent can machine learning algorithms classify incoming messages to organizational helplines in comparison to human coders?').*

6.1 Technical implications of machine learning classifications

6.1.1 Input data of BørneTelefonen

In this research, a dataset is used which has many features including a relatively small number of complex letters. The letters are complex as they are relatively long and written by Danish children. Chat slang, typos, chat language, and grammatical mistakes can be found in the letters. In addition to this, the letters are written in the Danish language. Danish is less well-researched especially in the field of natural language processing. Also, the difference in writing style and content of the letters is big, which is caused by the large age range of the children that sent letters. Furthermore, the different genders are not distributed in a balanced way.

Features

The dataset which was provided by employees of Børns Vilkår includes many features that could be relevant for the binary classification of the letters. In this study, only the letters sent by the children were used as independent variables. However, one could argue that other features could also be relevant, such as the subject, category, and reply. The machine learning algorithms would likely have performed more optimally if these features were added. Some of these features (such as the reply from the volunteer of BørneTelefonen) would not be available with newly incoming letters. Therefore, it might be

understandable that this is not taken into account in this analysis. Still, other features (such as category) are available for all incoming letters. It might, therefore, be questionable whether these features are not included. For this, it is important to think about the goal of this study. If the goal of this study is to build the best machine learning algorithm for this specific organization (BørneTelefonen) then it would probably be more valuable to include these features (Martinčić-Ipšić et al., 2019). However, if the goal of this study is to build another algorithm that could be used for various organizations then it might make more sense to only include the feature(s) which are related to the text. This is recommended as it seems logical that most organizations that are interested in such an algorithm at least have some kind of written messages.

Children language

Natural language processing is being used more regularly in non-perfectly written text, such as posts/messages on social media platforms. Still, the performance of these algorithms is often lower than for correctly written posts. Also, there are still no pre-processing techniques available that can translate messages with many typos, slang, and grammatical mistakes to correctly written messages. Messages written by children can make this process even harder as their language skills can be less developed than of adults. This could cause problems in terms of ethical machine learning, as the algorithm might treat letters written with fewer mistakes and abbreviations in a better way than other letters (Sindahl et al., 2018).

Danish language

Pre-processing techniques are constantly improving on how to process text messages that are not correctly written. There is a need for improvement, especially to understand the context of social media posts for instance. However, these developments are still at an early stage: The research mainly focuses on messages written by English speaking adults. Slowly, the development of these techniques is being applied to other languages. Unfortunately, the Danish language is comparably small and not as widely spoken as English. Therefore, improvements in pre-processing techniques for this language are rather minimal. This could result in ethical dilemmas as the classification of the letters should not be biased by the correctness of someone's writing and language skills (Braasch, 2002; Derczynski, 2019).

The age range of children

In this research, letters are included from children with a variety of ages. It is not realistic to believe that children at the start of the age range (on average) have the same writing skills as children at a later age. Still, the feature age is not included in this study, and no age groups are made. This means that the letters of all age groups are processed and handled in the same way. From an ethical point of view, this is probably not the best approach. Some might argue that it is ideal to make algorithms for each age group and that the correct algorithm is (automatically) selected based on the age that the child selected when submitting the letter (Kriz & Skivenes, 2010). Unfortunately, it was not possible to do this in this study as the number of letters was limited, especially for ages at either extreme end of the normal distribution.

Gender distribution

In this study, the genders were not distributed in the same way as in the population. Usually, a way to deal with this issue is to upsample the minority group or downsample the majority group. As the number of letters in the database was rather small, both, upsampling and downsampling did not seem like an ideal solution. The risk of upsampling is that the small number of letters (mainly of boys) are included in the dataset so many times that the algorithm could result in overfitting of these letters. Downsampling is not preferred due to the risk of underfitting. There are not that many letters in general, therefore it might not be a good idea to delete a big part of the letters written by girls as this would result in a substantially smaller dataset (Hirschberg & Manning, 2015; Seedall et al., 2019). Despite this, it is important to be aware of the shape of the dataset. The algorithm is mainly trained on letters written by girls. Therefore, it might automatically classify letters of boys differently as the letters might seem distinctive from the letters written by girls. This can especially be problematic when the population of boys that write letters to child welfare organizations increases (while the number of children that consider themselves as girl or other gender stays approximately the same).

6.1.2 Labeling of the letters

All letters in the database were binary labeled in terms of neglect by human coders. It is unclear how the human coders exactly conducted the labeling process, as specific documentation of the labeling was not available. Some of the labeling concerns are about: whether an employee or volunteer did the labeling, whether all letters in the dataset were labeled by one or more people, how much relevant knowledge the human beings that labeled had, whether the same guidelines and definition of neglect were used for all the different letters, and what was done with the letters where it was not clear whether it was about neglect or not (did an extra person evaluate it, or were they classified as neglect for the sake of prevention). It is expected that the (binary) labeling of some letters is easier than of others. If this was the case, then it would be possible to start labeling the easy letters and to train an algorithm to automatically classify these letters. Afterward, the algorithm could be asked to classify the other letters, and the performance of the algorithm could be evaluated by a human being. Unfortunately, this technique could not be applied in this study. This has to do with the fact that the entire dataset was already pre-labeled by human beings. Also, no information was given on how certain the human being felt about the label that it gave to a letter. Furthermore, the dataset was rather small making this technique a bit more challenging (Bowker & Star, 2000; Khan et al., 2010; Zacharis, 2018).

In this research, supervised machine learning was used to automatically classify newly incoming letters. This means that the classification of the human coders is assumed to be correct, and therefore defined as the ultimate goal. However, it is questionable whether the human being is better at labeling the messages than the machine. First, humans can make errors such as misclicking. Humans might also be biased towards certain age groups, genders, or problems when labeling the data. Some might argue that this is alarming, as an algorithm that is trained on biased data is expected to continue behaving in this unethical manner with future incoming letters. Others believe that this bias could be seen as something positive as it helps the algorithm to understand the context. The bias can be perceived as a feature and not a bug as beliefs lead humans in certain ways (Chinmay, 2019).

Furthermore, humans might have difficulties interpreting some of the letters. An example of this is suicide detection. An important feature for classifying suicidal thoughts is the length of a text. This is a metric that could be easily interpreted by a machine. It is however questionable whether a human being would pick up such an independent variable. A human would maybe focus on sentences such as 'I want to commit suicide'. But, it might be unclear whether that is the problem, or that there is an underlying problem, such as lack of attention. On the other hand, specialists might be better at understanding subtle words in the letter, which could make them more valuable than machines. It seems important to understand the differences in qualities of human beings and algorithms as they can be quite distinct, and potentially complementary. Furthermore, understanding how the manual labeling of the letters is being done can be valuable to get a better insight into the potential risks of misclassifying the incoming letters from children (Bowker & Star, 2000; Khan et al., 2010; Zacharis, 2018).

6.1.3 Possibilities to process input data of Danish children

In this research is chosen to keep the letters in Danish, but to clean them and remove stop words. It would be possible to automatically translate the letters to English before cleaning them. This might seem beneficial as natural language processing has proven to work well in the English language. The only problem is that some letters are easier to translate than others. Letters with many mistakes will partially stay in Danish as google translate would assume these words to be names or terms. This could cause difficulties, therefore it is chosen to stick to the Danish language (Derczynski, 2019).

The letters are automatically cleaned, and stop words are automatically removed. Every letter thus undergoes a few simple steps (e.g. convert everything to lower case and remove punctuation) which are defined as pre-processing. Most likely, a Danish speaking person would have been better at doing pre-processing, as this person could understand the sentence structure and add/remove words or typos where necessary. Although the building of the algorithm(s) is easier with cleaner data, it will probably not be beneficial when considering the practical use of this algorithm. Ideally, the letters provided to the algorithm in the future are similar to the ones it was trained on. As the future letters will not undergo text cleaning by a person, this should not be done to the letters in the dataset either (Khan et al., 2010; Kirkedal et al., 2019).

6.1.4 Differences in text feature extraction techniques

After the text is cleaned it is important to extract relevant features from the text that can be used by the machine learning algorithms to (binary) predict whether the letter is about neglect or not. Term Frequency-Inverse Document Frequency (TFIDF) is the text feature extraction technique that is used in this study. It can transform the text into a meaningful representation of numbers. The technique is widely used to extract features across various NLP applications. It can extract the most descriptive terms in a document, and it allows us to compute the similarity between various letters.

However, the TFIDF technique also has some disadvantages. The main disadvantage of this research is that it is based on the (simpler) text feature extraction technique called bag of words. This means that it does not capture co-occurrences in different documents, semantics, and the position in the text. An easy example of this would be that the statement ‘I feel not good’ would be seen as neutral as it does not understand that the word ‘not’ is related to ‘good’. It thus looks at all individual words without looking at the relationships between the words (Joachims, 1996).

There are text feature extraction techniques that do not have this disadvantage, such as word embedding and topic modeling. A word embedding is a learned representation for text where words that have the same meaning have a similar representation. And, a topic model is a statistical model that extracts abstract topics from the letters. Hidden semantic structures in the text body can be discovered through topic modeling. The only problem with these techniques is that the number of categories is too large as some of the letters are relatively long and include mistakes. This in combination with the little theoretical support about these sub-linear relationships resulted in the use of TFIDF. Still, the results could have been better when word embedding or topic modeling was used, especially when the letters were shorter and cleaner (Yun-tao et al., 2005).

6.2 Societal viewpoint on automated classification

6.2.1 Economic viewpoint on automated classification

In the first instance, it might seem beneficial from an economic point of view to support welfare professionals with technical tools. Most likely, implementing a classification algorithm requires some initial startup capital to implement the algorithm and to train the staff to use an algorithm. Some annual maintenance costs could also be expected to guarantee that the algorithm performs as preferred. One of the benefits of implementing a classification algorithm is the more efficient use of human capital, a production factor that is scarce at BørneTelefonen. From an economic point of view, it seems beneficial to value efficiency, mainly in terms of labor-saving. Also, a timely response could have a positive effect on some peoples' wellbeing. However, what if sensitivity should be the pursued value. This question is discussed in the next subsubsection.

6.2.2 Effect of automated classification on children

One might believe that the machine learning algorithms of this study do not negatively affect children as the algorithm is only in charge of automatic classifying the letters and not automatically replying to letters from children. Still, the potential negative societal effects of misclassification should not be underestimated. And, even if misclassifications would not occur, children can still experience negative thoughts that should be considered. Children can be negatively surprised when they notice that some of the people they know (e.g. siblings or classmates) got a reply a lot quicker than themselves. Also, they might not understand why sometimes their messages receive an immediate reply, and sometimes it can take weeks. To reduce these questions and potential disappointments one might argue for creating transparency by informing the children about the automatic classification. However, the risk is that this makes children feel even worse. They might have the feeling that they are treated as a number. Also, this might make them even angry, disappointed, or confused when they notice that some messages get a quicker reply than theirs. They might question why their problem is not 'important enough'. It can also be that they believe that the classification is not done fairly. Or, they might change their messages based on the details of automation to enhance their chances of getting a quicker reply (Kriz & Skivenes, 2010; Sindahl et al., 2018; Smith & Donovan, 2003; van Dolen & Weinberg, 2019).

Another possibility is to give children the choice to let their messages be (partially) classified by an algorithm or by a human being. This can be achieved by adding both options, in the form of two submit buttons, at the bottom of the interface where children write and send letters, see Figure 7. This would also be in line with EU GDPR article 22(1) which states that “the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her” (GDPR.eu, 2020).

One of the ways to get a better understanding about these challenges and to develop an algorithm that does not only perform well theoretically, but that is also of practical use, could be to do participatory action research, in which children would be actively included in the design process (AI High-level expert group, 2019; Bolander, 2019; Kemmis & McTaggart, 2005).

6.2.3 Effect of automated classification on specialists

The goal of using an algorithm that automatically classifies the incoming letters is to support the employees and volunteers. Still, there is a risk that these specialists feel undervalued instead of helped. They might not understand why someone believes that an algorithm is better at doing their job than themselves as they can value sensitivity more than efficiency. Also, they can feel uncomfortable using this technology, both the thought of using it as well as the actual training of the algorithm can cause problems. This can be problematic as the algorithm requires human supervision, especially in the beginning phase where the algorithm should be informed about its mistakes to improve. The process where humans collaborate closely with a machine to perform a task is called augmentation this is in contrast to automation, which means that a machine takes over the work from a human being (Raisch & Krakowski, 2020). A benefit of this process is that specialists get a better understanding of the way the algorithm is working, this can help them in gaining trust and explainability in the algorithm (Bolander, 2019). Thus, if specialists do not agree with the way that the algorithm is classifying, it can provide support which can help them to reduce the feeling of being undervalued (Smith & Donovan, 2003). Still, all specialists must agree with the chosen classification strategy to train the algorithm in the best possible way. As expected, some specialists might agree more with the chosen strategy than others. More concerning are the specialists that do not agree at all with the type of classification that is chosen, namely to classify letters based on neglect. They might believe that another factor should have been prioritized.

Specialists can also disagree with the fact that binary classification is done, instead of classification in more category (which thus allows for more nuances). Similarly to children, it could be beneficial for specialists to include them more in the design process of creating the algorithm. This could have been done by conducting a participatory design for developing a machine learning classifier, where specialists are included as one of the stakeholders (Bolander, 2019; Kemmis & McTaggart, 2005). Still, Raisch and Krakowski (2020) state that augmentation alone should not be the goal, as augmentation cannot and should not be separated from automation. They believe that too much attention on either automation or augmentation can feed the reinforcing cycles causing negative societal and organizational results. But, if organizations use a broader perspective on both augmentation and automation, they can handle the tension better and reach complementarities that can benefit society (Raisch & Krakowski, 2020).

6.3 Cultural viewpoint on automated classification

It is also important to look at automated classification from a cultural viewpoint as there are many different cultures because culture changes over time, and as machine learning algorithms change by definition based on the input they got. There are many people with different backgrounds, beliefs, and cultures living in Denmark, and these differences are even bigger when considering the world population. All these people have different norms and values which the algorithm of this study can only partly take into account. These cultural differences can be seen in the (content of the) letters written by the children as well as in the way that a specialist reads and evaluates the letters. A cultural view that could have a significant impact on this research is the definition of neglect. People of certain cultures might have a different definition of neglect than others. Most likely, the algorithm that would be applied in Denmark would be based on the official Danish legislation and governmental guidelines defining neglect (Gilbert et al., 2011). In this way, the general culture of the country in which the algorithm is applied is chosen. This could work for many people in Denmark, but this should be considered for minority groups in Denmark and when someone plans on using this algorithm for a different purpose, e.g. to help a child welfare organization in another country. Another cultural problem is that there can be cultural changes over time, indicating that concepts such as the official Danish definition of neglect are expected to change over time. Most likely, these changes will differ based on the change in social-cultural perceptions of the population. This means that a well-performing ethical algorithm might be biased in the future (Bail, 2014).

Furthermore, machine learning algorithms are constantly learning, developing, and changing. It could anticipate cultural changes in the society and behave accordingly. But, it could also go in the wrong direction. There are, thus, two sources of divergence: cultural and technological/algorithmic.

6.4 Ethical viewpoint on how to handle incoming issues

"In AI ethics, technical artefacts are primarily seen as isolated entities that can be optimized by experts so as to find technical solutions for technical problems. What is often lacking is a consideration of the wider contexts and the comprehensive relationship networks in which technical systems are embedded"

(Hagendorff, 2020, p. 4)

To discuss the ethical aspect of the machine learning classifiers created in this study is chosen to use the ethical guidelines for trustworthy artificial intelligence that are created by an independent high-level expert group on artificial intelligence (AI HLEG) which is set up by the European Commission. According to the guidelines, trustworthy artificial intelligence should focus on the following 3 components (1) be lawful, meaning that artificial intelligence should respect all applicable regulations and laws, (2) be ethical, implying that ethical values and principles should be respected, and (3) be robust, meaning that it should be solid from a technical point of view while respecting the social environment (AI High-level expert group, 2019; Bolander, 2019).

It is expected that the machine learning classifier of this study is lawful, as the act of classification, and what is done with those classifications does not seem to be in contradiction to regulations and laws. Whether the machine learning classifier will act ethically is more concerning. This is related to the robustness of the machine learning algorithm.

Ethical values include fairness, non-discrimination, and diversity. It is valuable to reflect on the established strategy or set of procedures to avoid creating or reinforcing unfair bias in the machine learning system, both regarding the algorithmic design as well as the use of the input data (Bolander, 2019). Deliberately, it was decided to exclude features such as age and gender in the design phase of the algorithm. This was decided based on the findings of ITelligence which used a counterfactual fairness model on the fair prediction in categorizing chats of Danish children with adult counselors of BørneTelefonen to create an ethical machine learning classifier. Still, it is questionable whether the findings of the chat could be applied to the letters. And, it could be that age and gender are features that are relevant to get to a good classification. Also, discrimination can happen based on factors that are different than just age and gender. Writing style is an example of a feature that can negatively affect the classification of the algorithm, but which might not be taken into account sufficiently. Another challenge could be that the different populations of children in this data set are not correctly represented or diverse enough. E.g. the number of letters that were classified as neglect is rather small. This implies that the classifier is not working as optimal for them as for children that wrote letters that are not about neglect. These problematic use cases or specific populations could have been tested separately. Furthermore, this machine learning classifier is not trained to test and monitor for potential biases during the development, deployment, and use phase of the algorithm. It could be considered to put in place such a mechanism that can flag potential problems related to discrimination, poor performance, and/or bias of the machine learning algorithm.

Adult counselors and children (end-users) are mainly considered in the design process of this algorithm. However, others that could potentially be indirectly affected by the machine learning algorithm could have been considered too. To get a better understanding of how to handle incoming issues in the future, it is important to research various aspects, such as whether transparency of the algorithm can take away certain concerns for the stakeholders of this algorithm (e.g. specialists, children, and their parents).

Also, no adequate definitions of ‘fairness’, ‘non-discrimination’, and ‘diversity’ were used in the designing process of the machine learning classifier. Therefore, a metric or quantitative analysis to test and measure these definitions was also lacking. More mechanisms could potentially have been established to ensure fairness, non-discrimination, and diversity in the designed machine learning classifier (AI High-level expert group, 2019; Hagendorff, 2020; Khan et al., 2010; Lai et al., 2015; Seedall et al., 2019; Zacharis, 2018).

7. Conclusion

The goal of this research is to get a better understanding of how appropriate it is to automatically classify incoming messages with the use of a machine learning algorithm instead of with the help of expert human beings. For this research was chosen to focus on the Danish Child helpline, BørneTelefonen, as a case study. This study aims to see whether enhancing the speed of external communication to children with the use of machine learning classifiers would be an appropriate solution. The type of incoming messages studied in this research were digital letters from Danish children, and the classification was done in terms of neglect.

The main research question *'To what extent can machine learning algorithms classify incoming messages to organizational helplines in comparison to human coders?'* is being answered with the help of the following two sub research questions: (1) *'How accurate are machine learning algorithms when classifying incoming messages in comparison to human coders?',* and (2) *'What are possible technical, social, cultural, and ethical limitations when using machine learning algorithms to classify incoming messages?'.* These questions are addressed in the following subsections.

7.1 Performance of machine learning classifiers

In this study, the performances of various machine learning algorithms were compared to understand how well these algorithms make binary classifications of texts compared with expert human classifiers. It was found that support vector machine performed most optimally, with an F1 score of 94 percent for letters that were not about neglect and of 23 percent for letters that were about neglect. This F1 score is the weighted average of precision and recall. This value is best at 100 percent. The two machine learning classifiers that are expected to perform well with unclean data (e.g. letters with typos, and grammatical mistakes) are random forest, and recurrent neural network. But, they belong to the worst-performing classifiers in this study. This is because the dataset was too small, and the messages of the children were too complex for these algorithms to learn how to classify incoming letters in the future. If speed is not an issue, and a sufficiently large sample size is available then it would be recommended to BørneTelefonen to classify with the use of a (recurrent) neural network. If a large dataset is available, but not sufficient time, then a random forest could be a useful classification technique. Naïve Bayes has the potential to classify letters that are about neglect (the minority group) in the best possible way. This, however, comes at the cost of a low F1 score for the letters which are not

about neglect. It could be valuable to look into the possibilities naïve Bayes has to offer when automatic classification aims to correctly classify letters that are about neglect. This performance could be compared to support vector machine and decision tree which were the two machine learning classifiers with the best relative performance for this study.

7.2 Limitations of machine learning classifiers

There could be many objections to using machine learning algorithms to automatically classify incoming letters. First of all, several algorithmic complications could impact the result. For example, when the aim is to classify letters from Danish children in need, the algorithm might struggle to understand the messages. Incoming messages could be of different lengths, have different writing styles, different levels of writing, be written in a small language (Danish), contain errors, slang, abbreviations, and much more. These issues are expected to be less when dealing with messages which are written in a common language (such as English) and do not contain mistakes, and chat languages.

When a machine learning algorithm is used to classify incoming letters, then the human coders can get a different role within the organization. They might become responsible for supervising the algorithm, or they might spend (more) time on answering messages from children. This change could be seen as an improvement as it could be more efficient. Still, it is valuable to realize that some human classifiers might feel uncomfortable, undervalued, and even demotivated when machine learning classifiers are used.

Children might also experience similar feelings of discomfort and might think that they are not being heard. They might feel treated like a number instead of a person. And they might not understand why their message is not being prioritized, even though they experience it as urgent and highly problematic.

7.3 Appropriateness of automatically classifying incoming messages

Algorithmically, it is possible to automatically classify incoming letters as demonstrated in the F1 performance scores of the seven algorithms that were researched in this study. It might, however, still be necessary to fine-tune the performance of these algorithms by, for example, training them on a larger dataset. And, there might be some other technical challenges, such as using a machine learning algorithm on a small language as Danish. Despite this, it seems that machine learning classifiers have the potential to automatically classify incoming letters. Still, it is questionable whether it is wise to already apply automatic classifications to all types of classification issues in various types of industries. For example, when looking at classifying incoming letters to BørneTelefonen based on neglect, there are still many uncertain factors and potential limitations. These factors are mainly related to the impact that machine learning classifiers could have on the wellbeing of children and experts. It is not yet known how people would respond to automatic classifications. People might be happy by the efficiency that it can cause, and the increased speed in which their messages are being answered. But, it can also harm those people. Senders of the messages might experience feelings of frustration, anxiety, and disappointment. As there is the right not to be subject to decisions that are made solely based on automated processing (GDPR.eu, 2020), it is appropriate to give the sender of the message the choice of whether their message is automatically classified or not. This could be done by having two different submit buttons, and the buttons could be accompanied by an informative video which explains the implications of both options. Depending on the industry and the type of classification it could be said that it is appropriate to automatically classify (some of the) incoming letters, as the technical capabilities are available and the limitations can be reduced based on the strictness of the algorithm. After implementing the algorithm, all incoming letters can still be checked by human coders, then the value of a minimal performance score can be decided for letters that do not get any supervision from a human coder. If a misclassification does not significantly affect the subject or have any legal effects on them. Then it seems appropriate to automatically classify letters of which the algorithm is (almost) certain that it performs in the same way as a human being. Still, there are many unknown factors related to automatic classification, therefore it is important to conduct further research in this field to get a better understanding of the appropriateness to use a machine learning algorithm to automatically classify incoming letters. In the next paragraph, various points are mentioned which could be considered when doing future research on this topic.

7.4 Further research

As mentioned before, there are many uncertain factors and potential negative implications of using machine learning algorithms to automatically classify incoming letters. In future research, it is therefore wise to not only investigate the options of improving the performances of the machine learning algorithms, but also to get a better understanding on the impact it has on its stakeholders, such as children, parents, employees of Børns Vilkår, and volunteers of BørneTelefonen. Ideally, these stakeholders are closely involved in the design process of the actual machine learning algorithm, this could be done with the use of participatory action research. More messages from children should be included in this future study. It is thereby important that all subgroups in a population are appropriately represented in this dataset. This is useful to increase the performance of the minority group in the study (e.g. letters written about neglect). Also, it is desirable that incoming messages from these children are being labeled by various experts in the field. The messages that were labeled in the same way by all experts could be used to train the algorithms. The letters which were labeled in conflicting ways (e.g. some experts labeled a letter as neglect, and some others did not label it as neglect) could be given to the model to automatically classify. The experts could then discuss this performance with each other and change the labels where needed. This process could improve the algorithms' future performance. This approach fits nicely with the aforementioned participatory action research design. Various algorithmic and social questions could be included in the participatory action research cycle to enhance the knowledge on these individual components, e.g. using more variables than just the text to automatically classify the incoming messages, or to ask volunteers how they would experience using a certain kind of machine learning classifier. In future research, it is also important to focus on topics as ethical machine learning as it is key that the algorithm is not biased towards a certain group of people. A more extensive study of using a counterfactual fairness model could be helpful for this.

8. References

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717–727. https://doi.org/10.1016/S0731-7085(99)00272-1

AI High-level expert group. (2019). *Building trust in human-centric AI*. European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

Akmajian, A., Farmer, A. K., Bickmore, L., Demers, R. A., & Harnish, R. M. (2017). *Linguistics: An Introduction to Language and Communication*. MIT Press.

Amrit, C., Paauw, T., Aly, R., & Lavric, M. (2017). Identifying child abuse through text mining and machine learning. *Expert Systems with Applications*, *88*, 402–418. https://doi.org/10.1016/j.eswa.2017.06.035

Anderson, M., & Anderson, S. L. (2011). *Machine Ethics*. Cambridge University Press.

Andrade, J. A. (2003). The effect of Internet use on children's perceived social support. *The Sciences and Engineering*, *64*, 406.

Areen, J. (1974). Intervention between the Parent and Child: A Reappraisal of the State's Role in Child Neglect and Abuse Cases. *Georgetown Law Journal*, *63*, 887.

Bach, M. P., Čeljo, A., & Zoroja, J. (2016). Technology Acceptance Model for Business Intelligence Systems: Preliminary Research. *Procedia Computer Science*, *100*, 995–1001. https://doi.org/10.1016/j.procs.2016.09.270

Baesens, B. (2014). *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. John Wiley & Sons.

Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and Society*, *43*(3), 465–482. https://doi.org/10.1007/s11186-014-9216-5

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013). How Noisy Social Media Text, How Diffrnt Social Media Sources? *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 356–364.

Beardsmore, R. (2015). *Measuring National Well-being*. *1*(1), 1–9.

Boiculese, V., Dimitriu, G., & Moscalu, M. (2013). *Improving recall of k-nearest neighbor algorithm for classes of uneven size*. 1–4. https://doi.org/10.1109/EHB.2013.6707403

Bolander, T. (2019). Human vs machine intelligence: *Proceedings of Pragmatic Constructivism*, *9*(1), 17–24.

BørneTelefonen. (2019). *Brevkasse*. BørneTelefonen. https://bornetelefonen.dk/brevkasse/

Bowker, G. C., & Star, S. L. (2000). *Sorting Things Out: Classification and Its Consequences*. MIT Press.

Braasch, A. (2002). A 2002 Current developments of STO – the Danish Lexicon Project for NLP and HLT applications. *In Proceedings from the Third International Conference on Language Resources and Evaluation, Las Palmas*, 986–992.

Butchart, A., World Health Organization, & International Society for the Prevention of Child Abuse and Neglect (Eds.). (2006). *Preventing child maltreatment: A guide to taking action and generating evidence*. World Health Organization.

Caplan, S. E. (2003). Preference for Online Social Interaction: A Theory of Problematic Internet Use and Psychosocial Well-Being. *Communication Research*, *30*(6), 625–648. https://doi.org/10.1177/0093650203257842

Carrell, D. S., Schoen, R. E., Leffler, D. A., Morris, M., Rose, S., Baer, A., Crockett, S. D., Gourevitch, R. A., Dean, K. M., & Mehrotra, A. (2017). Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *Journal of the American Medical Informatics Association*, *24*(5), 986–991. https://doi.org/10.1093/jamia/ocx039

Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, 161–168. https://doi.org/10.1145/1143844.1143865

Chinmay, C. (2019). *Smart Medical Data Sensing and IoT Systems Design in Healthcare*. IGI Global.

Cho, J., Lee, W. J., Moon, K. T., Suh, M., Sohn, J., Ha, K. H., Kim, C., Shin, D. C., & Jung, S. H. (2013). Medical Care Utilization During 1 Year Prior to Death in Suicides Motivated by Physical Illnesses. *Journal of Preventive Medicine and Public Health*, *46*(3), 147–154. https://doi.org/10.3961/jpmph.2013.46.3.147

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2015). Gated Feedback Recurrent Neural Networks. *International Conference on Machine Learning*, 2069–2075.

Cohen, S. (2004). Social relationships and health. *American Psychologist*, *59*(8), 676.

Colas, F., & Brazdil, P. (2006). Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. In M. Bramer (Ed.), *Artificial Intelligence in Theory and Practice* (pp. 169–178). Springer US.

Davidson, G., Bunting, L., Bywaters, P., Featherstone, B., & McCartan, C. (2017). Child Welfare as Justice: Why Are We Not Effectively Addressing Inequalities? *The British Journal of Social Work*, *47*(6), 1641–1651. https://doi.org/10.1093/bjsw/bcx094

Derczynski, L. (2019). Simple Natural Language Processing Tools for Danish. *ArXiv:1906.11608 [Cs]*. http://arxiv.org/abs/1906.11608

Dey, L., & Haque, S. K. M. (2009). Studying the Effects of Noisy Text on Text Mining Applications. *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, 107–114. https://doi.org/10.1145/1568296.1568314

Ding, F. (2019, April 29). How to AI: navigating the buzzwords of artificial intelligence. *Retresco*. https://www.retresco.de/en/how-to-ai-natural-language-processing/

Eriksson, M., Marschik, P. B., Tulviste, T., Almgren, M., Pereira, M. P., Wehberg, S., Marjanovič-Umek, L., Gayraud, F., Kovacevic, M., & Gallego, C. (2012). Differences between girls and boys in emerging language skills: Evidence from 10 language communities. *British Journal of Developmental Psychology*, *30*(2), 326–343. https://doi.org/10.1111/j.2044-835X.2011.02042.x

Fergusson, D. M., Horwood, L. J., & Ridder, E. M. (2005). Show me the child at seven: The consequences of conduct problems in childhood for psychosocial functioning in adulthood. *Journal of Child Psychology and Psychiatry*, *46*(8), 837–849. https://doi.org/10.1111/j.1469-7610.2004.00387.x

Flegal, K. M., Keyl, P. M., & Nieto, F. J. (1991). Differential Misclassification Arising from Nondifferential Errors in Exposure Measurement. *American Journal of Epidemiology*, *134*(10), 1233–1246. https://doi.org/10.1093/oxfordjournals.aje.a116026

Fletcher-Tomenius, L., & Vossler, A. (2009). Trust in Online Therapeutic Relationships: The Therapist's Experience. *Counselling Psychology Review*, *24*(2), 24–34.

Frank, E., & Bouckaert, R. (2006, September). *Naive Bayes for Text Classification with Unbalanced Classes*. European Conference on Principles of Data Mining and Knowledge Discovery. https://doi.org/10.1007/11871637_49

Fukkink, R. G., Bruns, S., & Ligtvoet, R. (2016). Voices of Children from Around the Globe; An International Analysis of Children's Issues at Child Helplines. *Children & Society*, *30*(6), 510–519. https://doi.org/10.1111/chso.12150

Fukkink, R. G., & Hermanns, J. M. A. (2009). Children's experiences with chat support and telephone support. *Journal of Child Psychology and Psychiatry*, *50*(6), 759–766. https://doi.org/10.1111/j.1469-7610.2008.02024.x

GDPR.eu. (2020). *Complete guide to GDPR compliance*. Horizon 2020 framework program of the European Union.

Gilbert, N., Parton, N., & Skivenes, M. (2011). *Child Protection Systems: International Trends and Orientations*. Oxford University Press, USA.

Glasheen, K., & Campbell, M. (2009). The use of online counselling within an Australian secondary school setting: A practitioner's viewpoint. *British Psychological Society, Counselling Psychology Section: Counselling Psychology Review*, *24*(2), 42–51.

Golinkoff, R. M., Hoff, E., Rowe, M. L., Tamis-LeMonda, C. S., & Hirsh-Pasek, K. (2019). Language Matters: Denying the Existence of the 30-Million-Word Gap Has Serious Consequences. *Child Development*, *90*(3), 985–992. https://doi.org/10.1111/cdev.13128

Gong, M., Shou, L., Lin, W., Sang, Z., Yan, Q., Yang, Z., Cheng, F., & Jiang, D. (2019). NeuronBlocks: Building Your NLP DNN Models Like Playing Lego. *ArXiv:1904.09535 [Cs]*. http://arxiv.org/abs/1904.09535

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 900–903. https://doi.org/10.1109/UKRCON.2017.8100379

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

Hagendorff, T. (2020). The Ethics of AI Ethics—An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *ArXiv:1508.01991 [Cs]*. http://arxiv.org/abs/1508.01991

Jakobson, R. (1961). *Structure of Language and Its Mathematical Aspects*. American Mathematical Soc.

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*, *39*(1), 1503–1509. https://doi.org/10.1016/j.eswa.2011.08.040

Joachims, T. (1996). *A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization.* (CMU-CS-96-118). Carnegie-Mellon Univ Pittsburgh PA Dept of computer science. https://apps.dtic.mil/docs/citations/ADA307731

Jones, L. S., Anderson, E., Loades, M., Barnes, R., & Crawley, E. (2019). Can linguistic analysis be used to identify whether adolescents with a chronic illness are depressed? *Clinical Psychology & Psychotherapy*. https://doi.org/10.1002/cpp.2417

Katalin, K. (2010). The suicidal linguistic behaviour (conclusions based on the multiple-standpoint analysis of the farewell letters written by 152 suicides). *Вестник Пермского Университета. Российская и Зарубежная Филология*, *3*. https://cyberleninka.ru/article/n/the-suicidal-linguistic-behaviour-conclusions-based-on-the-multiple-standpoint-analysis-of-the-farewell-letters-written-by-152-suicides

Kaufmann, G. M., & Beehr, T. A. (1986). *Interactions between job stressors and social support: Some counterintuitive results. 71*(3), 522.

Kemmis, S., & McTaggart, R. (2005). Participatory Action Research: Communicative Action and the Public Sphere. In *The Sage handbook of qualitative research, 3rd ed* (pp. 559–603). Sage Publications Ltd.

Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of Advances in Information Technology*, *1*(1), 4–20.

Kirkedal, A., Plank, B., Derczynski, L., & Schluter, N. (2019). *The Lacunae of Danish Natural Language Processing*. *167*, 356–362.

Kleinbaum, D. G., Klein, M., & Pryor, E. R. (2002). *Logistic regression: A self-learning text* (2nd ed). Springer.

Kotov, M. (2017). *NLP resources for a rare language morphological analyzer: Danish case*. Computational linguistics andintelligent systems (COLINS 2017). http://ena.lp.edu.ua:8080/handle/ntb/39456

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., Forshee, R., Walderhaug, M., & Botsis, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics*, *73*, 14–29. https://doi.org/10.1016/j.jbi.2017.07.012

Kriz, K., & Skivenes, M. (2010). Lost in Translation: How Child Welfare Workers in Norway and England Experience Language Difficulties when Working with Minority Ethnic Families. *The British Journal of Social Work*, *40*(5), 1353–1367. https://doi.org/10.1093/bjsw/bcp036

Kubat, M. (2017). *An introduction to machine learning* (M. Kubat, Ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-63913-0_14

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2018). Counterfactual Fairness. *ArXiv:1703.06856 [Cs, Stat]*. http://arxiv.org/abs/1703.06856

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification. *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Twenty-Ninth AAAI Conference on Artificial Intelligence. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9745

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22.

Lison, P. (2015). *An introduction to machine learning*. Edinburgh, UK: Language Technology Group.

Lopez, M. M., & Kalita, J. (2017). Deep Learning applied to NLP. *ArXiv:1703.03091 [Cs]*. http://arxiv.org/abs/1703.03091

Ludwig, S. A., Picek, S., & Jakobovic, D. (2018). Classification of Cancer Data: Analyzing Gene Expression Data Using a Fuzzy Decision Tree Algorithm. In C. Kahraman & Y. I. Topcu (Eds.), *Operations Research Applications in Health Care Management* (Vol. 262, pp. 327–347). Springer International Publishing. https://doi.org/10.1007/978-3-319-65455-3_13

Maglogiannis, I. G. (2007). *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press.

Magnuson, K. A., Sexton, H. R., Davis-Kean, P. E., & Huston, A. C. (2009). Increases in Maternal Education and Young Children's Language Skills. *Merrill-Palmer Quarterly*, *55*(3), 319–350. https://doi.org/10.1353/mpq.0.0024

Mandal, A. K., & Sen, R. (2014). Supervised learning Methods for Bangla Web Document Categorization. *International Journal of Artificial Intelligence & Applications*, *5*(5). http://arxiv.org/abs/1410.2045

Martinčić-Ipšić, S., Miličić, T., & Todorovski, L. (2019). The Influence of Feature Representation of Text on the Performance of Document Classification. *Applied Sciences*, *9*(4), 743. https://doi.org/10.3390/app9040743

Mavroforakis, M. E., & Theodoridis, S. (2006). A geometric approach to Support Vector Machine (SVM) classification. *IEEE Transactions on Neural Networks*, *17*(3), 671–682. https://doi.org/10.1109/TNN.2006.873281

Mertens, D. M. (2014). *Research and Evaluation in Education and Psychology: Integrating Diversity With Quantitative, Qualitative, and Mixed Methods*. SAGE Publications.

Morchid, M. (2018). Parsimonious memory unit for recurrent neural networks with application to natural language processing. *Neurocomputing*, *314*, 48–64. https://doi.org/10.1016/j.neucom.2018.05.081

O'Connor, D., & Conway, M. (2016). Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology*, *9*, 77–82.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., & Smith, N. A. (2013). Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 380–390. https://www.aclweb.org/anthology/N13-1039

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Patten, S. (2015). *Epidemiology for Canadian Students: Principles, Methods and Critical Appraisal*. Brush Education.

Perron, B. E., Victor, B. G., Bushman, G., Moore, A., Ryan, J. P., Lu, A. J., & Piellusch, E. K. (2019). Detecting substance-related problems in narrative investigation summaries of child abuse and neglect using text mining and machine learning. *Child Abuse & Neglect*, *1–13*, 104180. https://doi.org/10.1016/j.chiabu.2019.104180

Pestian, J. P., Grupp-Phelan, J., Cohen, K. B., Meyers, G., Richey, L. A., Matykiewicz, P., & Sorter, M. T. (2016). A Controlled Trial Using Natural Language Processing to Examine the Language of Suicidal Adolescents in the Emergency Department. *Suicide and Life-Threatening Behavior*, *46*(2), 154–159. https://doi.org/10.1111/sltb.12180

Raisch, S., & Krakowski, S. (2020). Artificial Intelligence and Management: The Automation-Augmentation Paradox. *Academy of Management Review*, 2018.0072. https://doi.org/10.5465/2018.0072

Rini, C., Emmerling, D., Austin, J., Wu, L. M., Valdimarsdottir, H., Redd, W. H., Woodruff, R., & Warbet, R. (2015). The effectiveness of caregiver social support is associated with cancer survivors' memories of stem cell transplantation: A linguistic analysis of survivor narratives. *Palliative & Supportive Care*, *13*(6), 1735–1744. https://doi.org/10.1017/S1478951515000681

Russell, S. J., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Malaysia; Pearson Education Limited. http://thuvienso.thanglong.edu.vn/handle/DHTL_123456789/4010

Sahakian, S., & Snyder, B. (2012). Automatically Learning Measures of Child Language Development. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, *2*, 95–99.

Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research Methods for Business Students*. Pearson Education.

Schnoebelen, T. (2017). Goal-Oriented Design for Ethical Machine Learning and NLP. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 88–93.

Schoenherr, T., & Speier-Pero, C. (2015). Data Science, Predictive Analytics, and Big Data in Supply Chain Management: Current State and Future Potential. *Journal of Business Logistics*, *36*(1), 120–132. https://doi.org/10.1111/jbl.12082

Seedall, M., MacFarlane, K., & Holmes, V. (2019, May 7). *SafeChat System with Natural Language Processing and Deep Neural Networks*. EMIT 2019. https://sure.sunderland.ac.uk/id/eprint/10968/

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Silipo, R. (2019, August 2). *From a Single Decision Tree to a Random Forest*. Dataversity. https://www.dataversity.net/from-a-single-decision-tree-to-a-random-forest/#

Sindahl, T. N. (2011). *Chat Counselling for Children and Youth: A Handbook.* Børns Vilkår.

Sindahl, T. N., Côte, L.-P., Dargis, L., Mishara, B. L., & Bechmann Jensen, T. (2018). Texting for Help: Processes and Impact of Text Counseling with Children and Youth with Suicide Ideation. *Suicide and Life-Threatening Behavior*. https://doi.org/10.1111/sltb.12531

Sindahl, T. N., Fukkink, R. G., & Helles, R. (2019). SMS counselling at a child helpline: Counsellor strategies, children's stressors and well-being. *British Journal of Guidance & Counselling*, *48*(2), 263–275. https://doi.org/10.1080/03069885.2019.1580676

Smith, B. D., & Donovan, S. E. E. (2003). *Child Welfare Practice in Organizational and Institutional Context*. *77*(4), 541–563.

Sperry, D. E., Sperry, L. L., & Miller, P. J. (2019). Language Does Matter: But There is More to Language Than Vocabulary and Directed Speech. *Child Development*, *90*(3), 993–997. https://doi.org/10.1111/cdev.13125

Stringer, E. T. (2013). *Action Research*. SAGE Publications.

Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). Predicting Dark Triad Personality Traits from Twitter Usage and a Linguistic Analysis of Tweets. *2012 11th International Conference on Machine Learning and Applications*, *2*, 386–393. https://doi.org/10.1109/ICMLA.2012.218

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *ArXiv:1503.00075 [Cs]*. http://arxiv.org/abs/1503.00075

Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications*, *34*(4), 2622–2629. https://doi.org/10.1016/j.eswa.2007.05.028

van Dolen, W., & Weinberg, C. B. (2019). An Empirical Investigation of Factors Affecting Perceived Quality and Well-Being of Children Using an Online Child Helpline. *International Journal of Environmental Research and Public Health*, *16*(12), 2193. https://doi.org/10.3390/ijerph16122193

Watson, S. (1994). An exploratory study into a methodology for the examination of decision making by nurses in the clinical area. *Journal of Advanced Nursing*, *20*(2), 351–360. https://doi.org/10.1046/j.1365-2648.1994.20020351.x

Wisse, A., & de Meij, J. (2015). Procesevaluatie van Jij en je gezondheid: Een nieuwe werkwijze voor in het Voortgezet Onderwijs. *GGD Amsterdam*, 47.

Wright, R. E. (1995). Logistic regression. In *Reading and understanding multivariate statistics* (pp. 217–244). American Psychological Association.

Wu, Y., Wu, W., Xing, C., Zhou, M., & Li, Z. (2016). Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots. *ArXiv:1612.01627 [Cs]*. http://arxiv.org/abs/1612.01627

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative Study of CNN and RNN for Natural Language Processing. *ArXiv:1702.01923 [Cs]*. http://arxiv.org/abs/1702.01923

Yun-tao, Z., Ling, G., & Yong-cheng, W. (2005). An improved TF-IDF approach for text classification. *Journal of Zhejiang University-Science A*, *6*(1), 49–55. https://doi.org/10.1007/BF02842477

Zacharis, N. Z. (2018). Classification and Regression Trees (CART) for Predictive Modeling in Blended Learning. *International Journal of Intelligent Systems and Applications*, *10*(3), 1–9. https://doi.org/10.5815/ijisa.2018.03.01

Zhu, X. (Jerry). (2005). *Semi-Supervised Learning Literature Survey* [Technical Report]. University of Wisconsin-Madison Department of Computer Sciences. https://minds.wisconsin.edu/handle/1793/60444