Classifying Urgency

A Study in Machine Learning for Classifying the Level of Medical Emergency of an Animal’s Situation

DANIEL STRALLHOFER

JONATAN AHLQVIST
Abstract

Classifying Urgency: A Study in Machine Learning for Classifying the Level of Medical Emergency of an Animals Situation

Daniel Strallhofer and Jonatan Ahlqvist

Abstract—This paper explores the use of Naive Bayes as well as Linear Support Vector Machines in order to classify a text based on the level of medical emergency. The primary source of testing will be an online veterinarian service's customer data. The aspects explored are whether a single text gives enough information for a medical decision to be made and if there are alternative data gathering processes that would be preferred. Past research has proven that text classifiers based on Naive Bayes and SVMs can often give good results. We show how to optimize the results so that important decisions can be made with these classifications as a basis. Optimal data gathering procedures will be a part of this optimization process. The business applications of such a venture will also be discussed since implementing such a system in an online medical service will possibly affect customer flow, goodwill, cost/revenue, and online competitiveness.

Index Terms—Medical Urgency, Veterinarian, Text Classification, Machine Learning, Multinomial Naive Bayes, Linear Support Vector Classification, Edge cases, Data gathering process.

1 INTRODUCTION

In a constantly evolving digital society the need for personalized hospital visits have created an opportunity for new forms of health care and health care counseling. Attentive subway travelers in Stockholm might have noticed the advertisement for services such as Doktor24 and KRY appearing [1]. These companies advertise a digital service in which the customer can schedule a video-call appointment with a medical professional. The health care professional then observes the patient and asks them questions. Thereafter the health care professional gives some sort of recommendation and further treatment plan. Recommendations can range from advice as simple as avoiding unnecessary physical movement to writing out prescription drugs or even referring the patient to a physical meeting with a designated specialist or a regular healthcare center.

This paper will be done together with the company FirstVet, a "digital veterinary clinic" that offers video meetings with veterinarians, similar to the health care services described above. They advertise a quick diagnosis and if needed they can write out a prescription or refer the customer to a qualified specialist.

2 BACKGROUND

When a user wants to schedule a video meeting with a FirstVet veterinarian they must first sign up on their homepage. When scheduling the actual meeting the user is requested for personal information about their pet such as age, breed, and weight. The user is also asked to write about the symptoms they are noticing in their pet. A video call meeting with a veterinarian is then booked for a later date. Shortly before the booked meeting a licensed veterinarian takes a quick look at the data submitted by the user, usually about 15 minutes before the meeting takes place. The veterinarian then proceeds with the video meeting where they talk with the owner, visually inspect the pet, and advance to a recommendation or diagnosis. Therefore, the veterinarian is mostly unaware at how urgent the matter is until just before the video meeting. At this point in time every case is handled in a FIFO-order (First In First Out). This has previously led to issues where urgent matters that require a physical check up are handled later than they should be.

FirstVet are interested in some sort of automated program that will, given information from the user, try and classify the issue as either urgent or non-urgent. This will then create an opportunity for the program to recommend the more pressing cases go straight to a clinic. Implementing such a system could be very helpful if a potentially urgent case has an extended waiting period until the scheduled meeting.

The goal of the project is therefore to implement a program that is able to successfully pinpoint FirstVet cases that have a high risk of being urgent. This is to be done without the input from a veterinarian and will therefore be based on the data submitted by the customer when scheduling a meeting.

2.1 Scientific Question

Is it viable to determine an animal’s state of medical urgency, based on the client data submitted to FirstVet, with a machine learning algorithm?

2.2 Scientific Relevance

This work will test whether it is possible to accurately classify pieces of text based on their content. This is of scientific relevance because it will evaluate the best methods...
for determining the severity of a medical situation through text. It will see if it is possible to get an idea of a problem before it is presented to a medical professional, both saving time and effort. No previous work of text classification on veterinarian data for medical classification has been found. Although similar text classification exists and is widely used, such as sentiment analysis in blogs [2], an important aspect of this is the fact that there has to be almost zero percent possibility for failure due to the potential seriousness of the customers situations. Or baring the possibility of a 100% success rate, putting measures in place to ensure that wrongfull classification of a situation does not result in more serious problems. This is interesting because it will analyze the possibility of handling mistakes in machine learning programs.

2.3 FirstVets Business Model

One important thing to note is that when looking at previous statistics in this field of study the focus will be on larger pets, disregarding animals such as fish and hamsters, and using cats and dogs as a baseline for pet insurance trends in Sweden.

A study conducted in 2012 showed that 76.5 percent of all dogs and 35.6 percent of all cats in Sweden are insured [3]. Estimates show that in 2018, 90 percent of all dogs and 50 percent of all cats are insured [4]. Because of this, FirstVets business plan is built upon their relationship with insurance companies. Using FirstVets services requires no initial payment or any form of subscription, but rather implements a pay-per-use business scheme, where users “pay” every time they contact a veterinarian for support. The reason “pay” is emphasized is because of how FirstVets work together with insurance companies.

Sweden’s largest pet insurance company, Agria Djurförsäkring, works closely together with FirstVet, offering anyone with Agria as their insurance company unlimited free calls to FirstVet. Along with this, they also pay a premium to FirstVet for every user who schedules a meeting with a FirstVet veterinarian. The reasoning behind this relationship is that when a customer schedules a meeting with FirstVet, most do this instead of taking their animal to a clinic immediately, and as the data shows, only about 30 percent of all FirstVet appointments are referred to a clinic [5]. This means that rather than a pet insurance company having to cover the medical costs of these visits, that money is saved by FirstVet intervening and helping the less serious cases.

In Sweden, FirstVet has similar business relationships with Dina Försäkringar, Folksam, ICA Försäkring, IF Försäkringar, Moderna Djurförsäkringar, Svedea, and Svealand Djurförsäkringar [5].

2.4 Business Applications

The first question posed in this paper is whether it is possible determine a customers urgency via machine learning, but the more important question for FirstVet is why they should implement this. This section analyses the issue from a business standpoint and will take under consideration how this project can possibly increase revenue and goodwill, improve customer flow and decrease costs. There are also many possible applications of this concept within other industries and the implications of this will also be discussed.

2.4.1 The risk

The concept of determining a customers urgency and giving a recommendation based on their input can be a risky one from a business standpoint. A faulty implementation can lead to a wrongful diagnosis which in turn could lead to increased customer dissatisfaction. If your animal is sick and a machine learning algorithm determines otherwise before a veterinarian even takes a look at the situation, it can have dire consequences. The main issue that this can cause is a decrease in customer retention. That is why it was quickly established that if such a program where implemented, then a fail-safe would have to be implemented along with it.

2.4.2 Customer Retention

As of right now, FirstVet boasts a 99 percent customer satisfaction rate [6]. This has multiple positive revenue effects. The direct effect is that it causes an increased rate of repeat customers, which is a pivotal aspect of many online businesses such as FirstVet. The value of this can be calculated in a customer lifetime value, which is the measurement of the net profit earned during the relationship between a customer and the business [7]. Because of how FirstVets business plan works, retained customers will be instrumental as every scheduled meeting increases revenue. An analysis found that that three-quarters of people who download an app stop using it within 90-days [8].

Due to the characteristics of the medical field being such that the customers main focus is the quality of the visit, having high customer satisfaction is the building block of a successful business. High satisfaction rates will not only lead to customer retention, but also attracting new customers.

This is where a successful implementation of the machine learning algorithm could be instrumental. If it were possible to gauge a customers urgency, and based on that, give a suggestion to take the pet immediately to a clinic, the chances of an animal who is in dire need of medical attention being left waiting until it is to late is drastically decreased. This can lead to an increase in customer satisfaction as well as lowering the chances of bad PR and a decrease in goodwill due an animal being diagnosed too late.

Just a couple instances of an animal being diagnosed to late could be enough to sway users from joining FirstVets clientele. Having a machine learning algorithm as a backup for cases where the user is unaware that their animal is really sick could help prevent future bad PR. It is important to stress that FirstVet should not rely fully on the machine learning algorithm to handle the customer flow because in its early stages it will not have more insight than a trained medical professional. Rather, FirstVet would be able to use it as an extra level of precaution, not altering the way they schedule clients, but rather telling users whose animals may be in need of urgent medical attention that they should go immediately to a veterinarian. The only change to FirstVet’s customer interaction in this case would be that some animals who dont need medical attention are referred to a clinic, which is something that FirstVet will have to consider when pitching to insurance companies.
2.4.3 Selling Point to Insurance Companies

FirstVets main selling point when dealing with insurance companies is the money that they save the insurance companies by eliminating the clinic visits for animals who don’t need it. A machine learning algorithm that makes sure that urgent cases are quickly sent to clinics would be another cost-saving selling point for the insurance companies.

An article called “Veterinary Introduction to Business and Enterprise” from the University of Adelaide details the costs of the different services of veterinary visits [9]. In this article it is shown that compared to a basic clinical visit, more extensive measures such as surgery ensue a large amount of new costs. If an animal is referred to clinic faster and therefore does not need the extensive treatment that would have been needed if the animal had to wait for a scheduled meeting, the costs decrease significantly. The value of early detection in FirstVet’s case is something that would have to be further researched in order to determine a specific cost estimate, but could be a point of leverage for FirstVet when starting relationships with new pet insurance companies or renewing existing contracts. Proof of similar cost reduction in humans is evident in cancer treatment [10].

3 Theory

3.1 Text classification

Text classification or document classification is an important problem in computer science that tries to evaluate plain texts algorithmically or statistically and assign them to a predefined type or class [11]. Usually a classification task involves separating data into training and testing sets. Each datapoint in the training set contains one “target value” (e.g. what class or type is this data assigned to) and multiple “attributes” (e.g. features, where in text classification every word can be deemed a feature). While many different techniques can be used, the ones implemented in this study are Multinomial Naive Bayes and Linear Support Vector Classification (SVC). The reasons for choosing these particular algorithms are described in their respective section. These two algorithms are also recommended by scikit-learns guidelines for our particular issue (see fig. 1, page 4).

3.2 Multinomial Naive Bayes

The Multinomial Naive Bayes model is chosen for multiple reasons. It is simple, easy to implement and often produces impressive accuracy despite its name. The naive part of the name refers to the “assumption” that all the attributes of the features are independent of each other given the context of the category [12].

Multinomial Naive Bayes captures word frequency information in documents. In the multinomial model, a document is an ordered sequence of word events, drawn from the same vocabulary \( V \). We assume that the lengths of documents are independent of class [13]. A “bag of words” representation is created for each document \( d_i \). Define \( N_{it} \) to be the count of the number of times word \( w_t \) occurs in document \( d_i \). Then, the probability of a document given its class from Equation (1) is the multinomial distribution:

\[
P(d_i|c_j; \theta) = P(|d_i| | d_i!] \prod_{t=1}^{V} \frac{P(w_t|c_j; \theta)^{N_{it}}}{N_{it}!} \tag{1}
\]

The parameters of the generative component for each class are the probabilities for each word written \( \theta_{w_t|c_j} = P(w_t|c_j; \theta) \), where \( 0 \leq \theta_{w_t|c_j} \geq 1 \) and \( \sum_j \theta_{w_t|c_j} = 1 \) [13].

Lidstone smoothing is implemented to allow the assignment of non-zero probabilities to words which do not occur in the sample. The more common term, Laplace smoothing, is a unique case of Lidstone smoothing where the additive constant is exactly 1. Lidstone smoothing is a more over-arching method that can be altered to influence results.

Although a multi-variate Bernoulli model can be used here, it does not perform as well on larger vocabulary sizes as a multinomial model. Overall on average the multi-variate model performs worse than the multinomial model [13]. This was confirmed for this study as well during testing although to a minor degree.

3.3 Linear Support Vector Classification

Linear SVC is simply one of many techniques using Support Vector Machines (SVM). SVMs are a popular classification technique where the datapoints are mapped on a dimensional space. When it comes to text classification the dimensional space can move towards infinity as the number of features (words in this case) move towards infinity. The SVM tries to find a separating hyperplane with the maximal possible margin between the different classes in this higher dimensional space. For the purpose of binary text classification, it applicable to use a Linear SVM. The reasoning for this is that since the documents are to be classified to just two classes we can separate them in the hyperspace using just a linear hyperplane. A non-linear SVM was used in testing and showed marginally worse results than the linear SVM. In the case of Linear SVM the hyperplane separating the datapoints is linear. Fig. 2 (page 4) shows a simple version of 2-dimensional Support Vector Classification.

4 Method

4.1 Data

A machine learning approach is only as good as the data it is built upon. This project was done with the help of FirstVets dataset from customer bookings over a 17-month time period. The dataset is a 17MB .csv file. Each datapoint in the dataset contains a description of the medical problem written by the costumer. There is more information about the pet and the issue such as race, weight and age. To classify and set up labels we used the information regarding if the appointment had been referred to a clinic as a determiner between urgent and non-urgent issues. This is a simplification but still the best practice until cases are individually graded on an urgency scale by the FirstVet veterinarians. There is also a journal attached to every datapoint written by the veterinarian after the issue was handled describing the problem as well as suggested solutions.

Not only must the data be reasonable for an effective machine learning approach to be successful, there should be a lot of it as well. This dataset contains a total of 14,423...
Fig. 1. Scikit-learn cheat sheet for choosing an algorithm. We have more than 50 samples but less than 100,000 samples, predicting a category using labeled text data.

Fig. 2. 2-dimensional example of SVM

data points. However, due to missing information, such as certain fields having not been filled out, only a total of 8,436 data points are usable, the rest either contain an empty customer description or no information regarding how the issue was handled after consultation with the veterinarian. Out of the valid data points 6,009 (71.2%) are labeled non-urgent and 2,427 (28.8%) are labeled urgent. It is important to remember this difference in class distribution as it directly affects the results of the model [14].

4.1.1 Consumer written texts

Although there are multiple fields of information to consider in this study the main focus will always be on descriptions of the issue written by the consumers. When filling out a booking with a veterinarian the consumer is asked specifically “Describe the issue” (Swedish: “Beskriv ärendet”). For privacy reasons we can not show examples of user texts, but analyzing the metadata for this datafield give some insight into the issue. On average consumers write 43 words in this field. After preprocessing the user text the average shrinks down to an average of 25 words. It is also important to note that is an informal text not designed to be used for machine learning. Therefore misspellings and abbreviations are very common. This will enlarge the word vector space unnecessarily, thereby reducing the success of the algorithm and increase the computational power requirement [15]. To counteract this the Linear Support Vector Machine is trained to only recognize words that have shown up at least 3 number of times in the dataset. This ensures that common misspellings/abbreviations are still treated as meaningful. An example of this would be the common abbreviation of the Swedish word “eller” to “lr”. They share a semantic meaning and removing “lr” could hurt the success of the algorithm.

4.2 Implementation

The code has been written in the following way. The machine learning techniques are implemented through scikit-learn. Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems [16].
Also imported from scikit-learn are evaluation metrics such as k-fold cross validation and confusion matrix. The code is divided into three main parts, a reader, a pre-processor and a model builder. The reader simply reads the .csv file and sends the relevant user texts into to the preprocessor. After getting a cleaned string back it appends it to the data matrix that holds all the texts while also appending the corresponding class label to the target matrix. The preprocessor is more intricate and unique and is described in further detail. The model builder receives two large matrices (the data and its corresponding target). Both matrices are split up using the scikit-learn library into 80/20 train/test batches. The train sets are used to build a model using either Multinomial Naive Bayes or Linear Support Vector Machine. After the model has been built (which takes an average of 6 seconds) it is then evaluated on different scores using the remaining test set. These are precision, recall and f1-score set up in a confusion matrix. For added information 5-fold cross validation is used on the entire dataset to get a accuracy score as well as recall score for the urgent class since these metrics are especially noteworthy. K-fold cross validation works by training and testing on the same dataset while constantly changing what is used as training/testing data. This gives a more accurate presentation of the results.

### 4.2.1 Pre-processor

When implementing machine learning there are certain procedures with regards to how the data should be handled before it is used as training/test data that, although not mandatory, are necessary for optimal performance. These procedures are what is called the preprocessing stage of machine learning and the implementation of this completely depends on the data available. With regards to text classification, there are certain preprocessing steps that are more trivial, and it is standard practice to implement them, while others are more data-specific and require further research. As previously seen, FirstVets data was available in the form of a .csv file and had multiple features for every recorded booking. One preprocessing step has already been discussed and involved removing all datapoints with missing information. This step determines which datapoints can be used for machine learning and leaves us with only data that includes all the required information. However, although this data is now "usable" for machine learning, it is highly unoptimized. Because the user text will be the main point for classification, stemming and the removal of stopwords is still required.

Stopwords are words in any language that give no insight to the meaning of the text, such as "is" and "mine" and will have to be removed in order to not confuse the classifier. In order to accomplish this, a Swedish stopword list was imported from nltk.corpus [17] and every user text was parsed, removing any instance of the stopwords.

A stemmer is an algorithm that derives the normalized versions of words by removing the morphological and inflexional endings of words. In this project the SnowBall stemmer [18] was used due to it being one of the few stemmers available that relatively accurately stems Swedish words. The snowball stemmer is not a perfect stemmer as it does not find the original "root" word for every instance and can therefore not associate words where the inflexion causes the core of the word to change. However, since it does remove morphological and inflexional endings it allows for the grouping of words that would have otherwise been seen as different by the machine learning algorithms.

Preprocessing also includes making everything lowercase, separating words by symbols such as "-" and "/", as well as removing punctuation. See fig. 3 for an example of preprocessed sentences (Swedish and English).

![Text written by authors, not derived from FirstVet data](image)

An article written about the profound effect of preprocessing in order to obtain relevant data in machine learning argues that for optimal accuracy, one should match ones preprocessing choices with ones knowledge [19]. What this entails is not blindly allowing all available data to be used at training data, but rather making informed decision as to what data may be relevant. In the project, this meant interviewing a FirstVet veterinarian as to what data may be relevant in determining an animal’s state of emergency. For example, the veterinarian interviewed mentioned that many times an infant animal is at a higher risk of being referred to clinic than a fully-grown animal. Such insights build a foundation as to which features could possibly contain relevant information and which features give no relevant data. Along with the user text being the main point of analysis, this narrowed the features that would be taken under consideration to animal_species, problem_earlier, problem_duration, animal_birthdate, and appointment_type. These data samples will then be added as features in the Naive Bayes algorithm and the Support Vector Machine algorithms and analyzed based on their relevance in obtaining a higher precision rate when determining urgency in the appointments.

## 5 Results

Although there are many different ways of evaluating the success of a machine learning algorithm the confusion matrix is almost always a good baseline metric [20]. Table 1 show the results of the Multinomial Naive Bayes algorithm. Table 2 includes 5-fold cross validated accuracy and recall scores for added information.

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-urgent</td>
<td>72.7%</td>
<td>79.5%</td>
<td>75.9%</td>
<td>1173</td>
</tr>
<tr>
<td>Urgent</td>
<td>40.6%</td>
<td>32.0%</td>
<td>35.8%</td>
<td>515</td>
</tr>
<tr>
<td>Weighted average</td>
<td>62.9%</td>
<td>65.0%</td>
<td>63.7%</td>
<td>1688</td>
</tr>
</tbody>
</table>

The confusion matrix is based on a 80/20 split (the training data is 80% of the dataset with the testing done on the remaining 20%). The Multinomial Naive Bayes algorithm uses a Lidstone smoothing parameter (alpha parameter) of 0.2. The classes are weighted to their presence in dataset and are not treated as equally likely. For transparency all relevant hyper-parameters are included.
Table 2
5-fold cross validated score - Multinomial Naive Bayes

<table>
<thead>
<tr>
<th>Metric</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>64.7%</td>
</tr>
<tr>
<td>Recall Urgent-class</td>
<td>32.6%</td>
</tr>
</tbody>
</table>

MNB_parameters: {'alpha': 0.2, 'class_prior': None, 'fit_prior': True}

Table 3 and 4 show the results of the Linear Support Vector Machine algorithm.

Table 3
Results - Linear Support Vector Classification

<table>
<thead>
<tr>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-urgent</td>
<td>71.4%</td>
<td>84.8%</td>
<td>77.5%</td>
<td>1173</td>
</tr>
<tr>
<td>Urgent</td>
<td>39.5%</td>
<td>22.5%</td>
<td>28.7%</td>
<td>515</td>
</tr>
<tr>
<td>Weighted average</td>
<td>61.6%</td>
<td>65.8%</td>
<td>62.6%</td>
<td>1688</td>
</tr>
</tbody>
</table>

Table 4
5-fold cross validated score - Linear Support Vector Classification

<table>
<thead>
<tr>
<th>Metric</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>65.7%</td>
</tr>
<tr>
<td>Recall Urgent-class</td>
<td>21.4%</td>
</tr>
</tbody>
</table>

Table 5 show the top 10 most important features (or in this case stemmed words) for each of the classes according to the Linear SVM.

Table 5
Top 10 most important features (words) according to the linear SVM with their respective weight

<table>
<thead>
<tr>
<th>weight</th>
<th>Non-urgent</th>
<th>Urgent</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.7291</td>
<td>mjölk</td>
<td>2.0851</td>
</tr>
<tr>
<td>-1.5477</td>
<td>läs</td>
<td>1.9429</td>
</tr>
<tr>
<td>-1.4598</td>
<td>hornmöllan</td>
<td>1.8874</td>
</tr>
<tr>
<td>-1.4569</td>
<td>ras</td>
<td>1.8837</td>
</tr>
<tr>
<td>-1.4429</td>
<td>sek</td>
<td>1.8126</td>
</tr>
<tr>
<td>-1.4415</td>
<td>framlöv</td>
<td>1.8047</td>
</tr>
<tr>
<td>-1.4272</td>
<td>liksom</td>
<td>1.7836</td>
</tr>
<tr>
<td>-1.4167</td>
<td>skavstå</td>
<td>1.7746</td>
</tr>
<tr>
<td>-1.4166</td>
<td>darning</td>
<td>1.7718</td>
</tr>
<tr>
<td>-1.3993</td>
<td>sannolik</td>
<td>1.7697</td>
</tr>
</tbody>
</table>

Although the words are stemmed they give a perspective into how the customers are describing their animals conditions. This will be further discussed and analyzed in the Discussion.

Hyper-parameters for the Linear Support Vector Machine:

LinSVC_parameters: {'C': 1.0, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'loss': 'squared_hinge', 'max_iter': 3000, 'multi_class': 'ovr', 'penalty': 'l2', 'random_state': None, 'tol': 0.0001, 'verbose': 0}

6 DISCUSSION

With the results presented the question arises as to how the results can be evaluated, adjusted and improved. This will be explored throughout this section both from a computer science and a business perspective. The report presents mediocre results. To compare the results we set up a major class statistical baseline at 71.2% [21]. Both accuracy scores are slightly below this baseline at 64.6% (Multinomial Naive Bayes) and 65.6% (Linear Support Vector Classification).

Both algorithms seem to overestimate the Non-Urgent class over the Urgent one. This can be seen in their relative F1-scores. The urgent class shows a much lower F1-score (see fig. 4). This is likely due to an overweighting of the Non-Urgent class since it is more prevalent in the training data.

It is noteworthy that both algorithms receive very similar results.

6.1 Weighting

For this specific task the goal will most likely be to avoid letting through any Urgent cases as Non-Urgent. Therefore the focus lies on maximizing the Urgent recall score. This can be done by manually weighting the classes. If this is done the algorithm will be more inclined to classify a datapoint as Urgent. Fig. 5 (page 7) demonstrates how increasing the weight of the Urgent class will increase its recall but lower the overall accuracy. The goal of every machine learning algorithm should be to fit the question asked. Thereby the optimal solution might be to set a high weight for the Urgent-class even though it lowers the overall accuracy. The actual implications of this is however further discussed below in 6.8.1.

6.2 Effect of preprocessing

Also noteworthy is the very minor difference between the algorithms with and without preprocessing. Fig. 6 demonstrates these differences and the minor impact that preprocessing seems to result in. It is difficult to speculate why this is the case. One possible answer could be that although preprocessing is helpful, for the success of the algorithm, when the applied data is too homogeneous then the preprocessing is not capable of making each datapoint more distinguishable. Another possible reason for the lack of improvement could be that the Swedish stemmer is not as advanced and useful as its English counterpart. Testings show that its rules are rather naive and prone to mistakes.
For example: "liter", "liten", "lita" are all stemmed to "lit" despite their semantic differences.

### 6.3 Evaluation using another dataset

To evaluate the success of the algorithms we decide to show the results of using it on another set of datapoints. Included in the dataset is the journals written by the veterinarians after each meeting. This is a summary of the video meeting with the customer and contains textual data about the issue as well as suggested solutions. The results shown from this set proves that mediocre results are due to a problematic dataset. With the journal dataset the baseline result is reached and a high overall accuracy is possible (see fig. 7 and 8).

### 6.4 Top 10 words

As seen previously in table 5, the 10 most significant words for both the urgent and the non-urgent category were printed in order to give an insight as to which terms the algorithms weighed most heavily when deciding which category a text belonged to. This information can then be used for numerous different goals, both in terms of optimizing the machine learning algorithm and also in FirstVets business in general.

When looking at these words from a purely linguistic standpoint one might question whether some of these words really give an insight as to the urgency of a situation. Using the urgent category as an example, words such as "ryckning"(twitch) and "kortisontablet"(cortisone tablet) seem like they would give information about the medical severity while a word like "lr"(Swedish abbreviation for "eller"(or)) seems like it would give no information at all. This is however an incorrect way of viewing these words because although "lr" may appear to have very little medical relevance, it may have been weighed so heavily for the "urgent" category because it is a pattern among urgent cases that a user shortens "eller" to "lr" because of stress regarding the animals situations. On the other hand, it may have been common in urgent cases purely by coincidence. This is a question that would require a much more in-depth look at all 8,436 texts and even then might not give much more of an insight as to why these words are important.

Another interesting aspect to analyze is the presence of the words "ryckning"(twitch) and "darrning"(shiver). While both words seem to be semantically related they appear on the different sides of the list. How can this be explained? One answer could be that despite their semantic similarity the words could have underlying medical differences. It is however unlikely that laymen could make a qualified distinction between the two words. Further study in to the data itself showed no obvious correlation between the words and their medical severity. This along with further discussion with our supervisors trained in machine learning has pointed to the fact that the occurrence of these words in the "top 10" list is most likely due to random correlation between the words and the corresponding urgency level.

Such an issue can easily occur within machine learning containing smaller datasets and points to problems with the machine learning algorithm.

It is however possible that the words have some sort of connection to the urgency level if one were to examine the
context of the words by looking at the nearby words or even

the entire word vector instead of the individual words. For example, while the word "bleeding" may not contain much information regarding urgency, the combination of "bleed-
ing paw" is most likely much less urgent then "bleeding eye". This goes to show that further word analysis should be done on multiple context levels for optimal insight.

Still, this goes to show that further use of the algorithm should most likely not follow the top 10 words list as a rulebook for what words can be deemed as Urgent/Non-Urgent triggers. Instead the list can be used as a more general guideline into which further study is required.

A more practical use of these words would be in FirstVets data collection process. For almost every digital company in the world, data collection is an integral part of the business. This can be seen in the rising trend of websites using cookies and optimizing the way that users fill out data. In FirstVets case, what data is required by the user when applying for an appointment is something that is constantly being updated and improved. By looking at these words, FirstVet could gain an insight as to which words or categories of words are more relevant and familiar with their users and could shape how a user fills out information when scheduling an appointment. Using the "kortisontablet" as an example, FirstVet could ask users if the animal has taken any medication recently.

6.5 Alternative baseline

Although the statistical baseline can be set to a specific percentage (in this case 71.2%) it might be necessary to consider another baseline for this specific problem. Clinical studies estimate that approximately 90% of all medical diagnoses are correct [22]. This is for general medical diagnosis and not specifically veterinarian diagnosis. Although it varies greatly from physician to physician this estimate can be used as a realistic baseline for how good a machine learning algorithm must be for implementation. This baseline is never achieved throughout testing.

Although the “medical baseline” is more relevant in this situation, it is important to recognize that the algorithm is not a replacement for a veterinarian. Rather, the main goal of the algorithm will be to lower the amount of cases that are not referred to a clinic in time. Because of the real life implications of this project, it is almost more important to consider the consequences of the different accuracy levels. A case that is not labeled as urgent will still go through the current FirstVet procedure with no changes whatsoever. Therefore a small percent of algorithmically referred cases, for example through edge case analysis, could also be acceptable.

6.6 Developing Edge Cases

To prevent low statistical results an alternative method of classification could be used: “edge case” analysis. In this context, an edge case would be a situation where a customer appointment is weighed so heavily towards either urgent or non urgent that the possibility of the classification being correct is drastically increased. In medical cases such as these, that would mean either finding a statistical cut-off point where the probability of the classification being correct is extremely high or it could mean determining, together with a veterinarian, certain cases that are almost always urgent and weighting the machine so that these are always classified as urgent.

The first example of finding a cutoff point can be illustrated via both algorithms used. In Naive Bayes it would be when the total probability of document d belonging to class X is much larger than the probability of d belonging to class Y. In Linear Support Vector Classification it would be when the document d is sufficiently far away from the hyperplane. These limits would have to the determined so that the edge case analysis is still applicable to a reasonable amount of datapoints without sacrificing accuracy.

The result of this would be that not all documents could be classified. But the documents that are classified would be so with a very high degree of accuracy.

However, despite this possibly being an effective solution that would ultimately increase recall scores and only refer cases with high probability of urgency, the main point of improvement at this stage in time is still the data collection process. Because of how random the texts are and the large amount of possible responses from customers with low to no medical knowledge, classifying these texts with an acceptable accuracy is still a tall task. Before implementing an algorithm that only classifies "edge cases," FirstVet will have to change their data collection process.

6.7 Future Data Collection Possibilities

When considering how this project leads to a relatively low accuracy score with regards to the baseline, it may seem like implementing this type of program would be a questionable business decision. However, as shown above, the possibility of weighting the data to fit the business structure is possible and with an optimized data collection system this project is highly relevant. One of the main recommendations from this project would be improving the way in which a client describes their case. This could lead to both increased accuracy when classifying urgency, but also open up avenues for a plethora of different machine learning and business opportunities.

6.7.1 Manual Tagging by Veterinarians

To improve future results, it would useful to label the data in a way more optimal for machine learning algorithms. Instead of setting up the data as a binary classification issue it would be desirable with a grading scale between 1-5 where 5 is a very urgent case and 1 is a non-urgent case. Every case would then be graded by veterinarians while writing the case journal. To prevent grading bias FirstVet could set up clear guidelines with examples of how severe a case must be to achieve a certain grade. An example of this is shown in table 6. This would make sure that there is a clearer distinction between different cases that currently would be labeled as equally severe.

6.7.2 Tree Structure

The data collection process could be set up using a tree structure. This structure would be modeled after decision trees that are already an integral part of classification theory. Decision tree are large models that by first answering
TABLE 6
Example of new grading scale

<table>
<thead>
<tr>
<th>Severity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No action needed. Problem will pass by itself.</td>
</tr>
<tr>
<td>2</td>
<td>Non-prescription medicine needed.</td>
</tr>
<tr>
<td>3</td>
<td>Prescription based medicine needed.</td>
</tr>
<tr>
<td>4</td>
<td>In addition to medicine a physical check-up with a veterinarian is needed.</td>
</tr>
<tr>
<td>5</td>
<td>A physical check-up with a veterinarian is needed urgently.</td>
</tr>
</tbody>
</table>

Fig. 9. Rudimentary decision tree [23]

generic questions and then more and more specific ones eventually will end up on a node that classifies the issue (see fig. 9). The problem with this model would be designing it and setting up relevant questions for it, since it would in practice be extremely large and would require the input of both trained veterinarians and FirstVet employees in order to determine the important nodes. It is however completely in the realm of possibility and would allow for a very efficient and specific data collection process.

6.7.3 Chatbot

Another solution would be a knowledge-based conversational chatbot that, based on a users initial text, inquiries for more specific information. Similar to the decision tree model the chatbot would make inquiries based on pre-defined decision points, or nodes. Similar bots have recently been developed within the medical field such as https://www.your.md/ but this process is very difficult and the existing platforms are still lackluster [24]. The main reason being that unlike a decision tree, a chatbot will have to derive relevant information from ambiguous user texts, navigate spelling errors, and accurately press the user for more relevant information.

Out of the data collection processes this on has the possibility of being the most effective and accurate one due to the fact that it would be able to collect more data than the decision tree. It will however take the longest and the most manpower to implement.

6.7.4 More Text Boxes

As previously discussed, the way that FirstVet currently collects data regarding meetings is by having the user fill out one large text field where they describe the situation and then filling out information about the animal such as age, breed, and how long the issue has been relevant. The recommendation here would be to instead of having a user describe the situation in one single field, it could be split up into smaller, more specific textboxes. These text boxes could then have predefined answers in a dropdown menu. Although this option would limit the information that a user would divulge, it would eliminate ambiguity and allow for only “relevant” information. It would also require the least amount of time and manpower to implement.

6.8 Business Discussion

Many of the topics of discussed above are primarily analyzed from a computer science standpoint. For FirstVet it will be important to understand what business implications they will have and what effects they could result in on the business side of the company.

6.8.1 Business Implication of Weighting

The weighting process is something that will have to be looked into from a business standpoint due to its overarching effect on FirstVets business model, mainly their negotiating powers with pet insurance companies. As stated previously, urgent cases being referred to clinics earlier has a cost benefit for the insurance companies because it will allow certain cases to be treated before expensive procedures are necessary. However, a decrease in overall accuracy will increase the amount of trivial cases being referred to clinics when no such action was necessary, causing a cost increase for insurance companies. Because of how FirstVets business plan is built on the fact that they are an extension of pet insurance companies services, this is something that has to be analyzed before being implemented.

An effective way to analyze the effects of such a venture would be to create a cost-benefit analysis (CBA) where the positive effects of increased recall in emergency cases are weighed against the negative effects of an increase of non-emergency cases. This can be done by calculating the average cost-saving of early treatment and multiplying it by the number of cases that received early treatment because of the algorithm and then calculating the average cost of trivial errands visiting a clinic and multiplying it by the number of those cases that were referred to a clinic because of the algorithm. These two numbers should then be compared at different stages of the weighting process in order to establish an optimal level of accuracy vs. recall for the urgent class.

6.8.2 Business Implication of Edge Case Analysis

Implementing edge case analysis is something that will not only be relevant in FirstVets business model, but could become relevant in all businesses where machine learning is implemented within an area where incorrect decisions are much more serious. It will lower the chances of machine
learning algorithms making mistakes that are harmful to the company.

In FirstVet’s case, wrongful classification of urgency could have severe consequences and overall accuracy is less important than urgent cases being properly classified. Edge case analysis becomes a type of guard against bad PR and decreased goodwill that could come as a result of wrongfully classifying an urgent case.

The downside of edge case analysis is that only a subset of user texts will be able to be classified. The overall effect of edge case analysis is thereby small compared to other alterations to the current algorithm such as Altering Data Collection Processes.

6.8.3 Business Implication of Altering Data Collection Processes

When a company considers undertaking a project such as altering the way in which they collect data, many monetary and resource aspects have to be considered. As discussed previously, the main issue when considering implementing machine learning for FirstVet is the data available. In order for accurate machine learning to be viable, the data should be collected with machine learning in mind. Three possible improvements to FirstVet’s data collection process have been discussed and will be further analyzed with regards to their viability in terms of what resources they would require. FirstVet will then have to determine whether the benefits of this outweighs the commitments necessary to implement them.

Already from this point in time, it is clear that implementing a chatbot might not be the optimal solution for FirstVet. This is due to the large capital commitment it would require along with the fact that previous studies show that even when such a project is undertaken, positive results are not guaranteed. Considering the fact that FirstVet’s current customer process is very effective, adopting a chatbot would not allow for improved customer response and the increased data collection would be underwhelming.

Similar to a chatbot, implementing a tree structure would require significant resource allocation, although on a lower scale. A tree structure would also allow for a much more structured approach as to what data is collected and it would allow FirstVet to tailor which information they find important and guide users into entering relevant information. Despite it requiring the input from both veterinarians and developers in order to implement a function tree structure, it is a much more viable option than a chatbot.

The final data collection system recommended was expanding the number of text boxes a user filled out. This might be the most viable option for FirstVet at this point in time due the minimal resource allocation required and the unsure future of machine learning within FirstVet. The new data collection process would be able to function on FirstVet’s existing architecture and would only require the input of trained veterinarians in order to determine which information is relevant. This could be changed depending on which type of machine learning implementation FirstVet deems most relevant. The previous two options might be a future possibility that become relevant when FirstVet want to implement machine learning with a certain goal in mind.

6.8.4 SWOT Analysis

The implications of implementing such an algorithm are summarized in a SWOT analysis in Fig. 10. Here the aspects brought up throughout the report are discussed.

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Fast customer response</td>
<td>- Accuracy not high enough</td>
</tr>
<tr>
<td>- Improved efficiency by removing urgent cases wait</td>
<td>- No personal contact with algorithmically referred cases</td>
</tr>
<tr>
<td>- Improved customer flow leads to more revenue</td>
<td>- Recall for urgent cases is low</td>
</tr>
<tr>
<td>- Reduces risk for urgent cases being treated to late</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Opportunities</th>
<th>Threats</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Reduced veterinary costs due to early treatment gives FirstVet bargaining power among insurance companies</td>
<td>- Increased veterinary costs due to mislabelled non-urgent cases being referred to clinics could lower FirstVet’s bargaining power with insurance companies</td>
</tr>
<tr>
<td>- Extra fast for urgent cases can lead to increased public image and goodwill</td>
<td>- Mislabeling an urgent error can cause customer dissatisfaction due to animal harm</td>
</tr>
<tr>
<td>- Data from algorithms can improve FirstVet’s data collection process</td>
<td>- Mislabeling a non-urgent case can lead to customer dissatisfaction due unnecessary visit to clinic</td>
</tr>
</tbody>
</table>

Fig. 10. SWOT analysis for the implementation of a classification algorithm

7 CONCLUSION

In this paper we have shown the implementation of two different algorithms for text classification of medical data. We have discussed the methodology of working with machine learning algorithms such as preprocessing and data analysis. We have compared results and discussed issues attributed to the used dataset and shown possible parameter switching solutions to addressing these issues. We have also theorized in alternative ways of collecting data in the future to achieve greater results.

Although the results are promising and show some sort of correlation between text and medical emergency, raising the accuracy and recall to an acceptable baseline will be difficult with the data currently collected by FirstVet. This is at least true when discussing the implementation of a fully functional classifier. The possibility of only classifying cases that reach a certain level of accuracy, or “edge cases” would be possible at this stage in time. Despite that, the overarching conclusion would be for improved data collection, due to the superior effect it would have on the end result as compared to only classifying “edge cases.”

ACKNOWLEDGMENTS

The authors would like to thank everyone at FirstVet for their support and interest throughout this project. The authors would also like to thank Joakim Gustafson for his assistance and helpful advice as supervisor.

AUTHOR PRESENTATION

Daniel Strallhofer, born 1996, is currently enrolled at KTH Royal Institute of Technology pursuing a Master of Science in Industrial Engineering and Management, minoring in Software Engineering.

Jonatan Ahlqvist, born 1995, has an IB Diploma from Sigtuna Skolan Humanistiska Läroverket. He is currently enrolled at KTH Royal Institute of Technology pursuing a Master of Science in Industrial Engineering and Management, minoring in Software Engineering.
References


