**CO600 Terms and Conditions Analyzer**

Adam Preston (ap720), Euel Domingo (ejd29), Kester Gideon (kg327), Marcus Robinson-Bailey(mer25)

*School of Computing*

*University of Kent*

*Canterbury, United Kingdom*

**Supervised by**: Alex Freitas ([A.A.Freitas@kent.ac.uk](mailto:A.A.Freitas@kent.ac.uk))

**Abstract**

This report describes the stages and processes involved in creating the terms and conditions analyser. The system analyser is a user-friendly website that highlights any risky statements in the document. With the help of machine learning, the analyser will scan through the document inserted and compare with a trained set of words we deemed to be risky. The paper will begin by describing how the concept came about then move on to provide more information on the background and aims of the project. Further description will be given on the product development and testing of the final product. In addition to this the report we will provide a technical description of the classification algorithms used, quality assurance and challenges faced.

**1. Introduction**

There have been countless occasions where individuals sign documents or contracts, whether they be phone contracts or social media terms and conditions with no intention of reading the terms and conditions. A recent survey concerning social networking terms says that more than 30% of the respondents said they have never read the terms and conditions when signing up to social networks.[1] In relation to this current issue, we decided to create a terms and conditions analyser. The purpose of this would be to would allow users to insert their document in the form of plain text into the web interface which would then indicate the risky statements by displaying them in a display box. The risky words displayed allows the user to focus on the sentences and hence review the document before taking further action. The idea would instigate that machine learning would be involved and the use of Neural Networks. Machine learning is a method of data analysis to effectively perform tasks and it has recently received an increase in interest over the last few years. Machine learning helps in various aspects; whether it is predicting the next video on YouTube, to predicting when earthquakes could occur. The report highlights the process we took from the generation of the idea, to the completed phase. Research was undertaken in the initial stages into different terms and conditions, machine learning API, pre-processing scripts. Following this we planned our project by setting out aims for the project and a design plan.

In preparation for the project, we realized that this would entail additional knowledge to be acquired in order to produce a system that is able to read through sentences and highlight the risky statements. We quickly realized that in order to create a good machine learning system, we would have to follow some required steps. We would first take into consideration data preparation capabilities. Data would be collected and words we deemed risky would be placed into a document. Terms and conditions from a few top companies were used. A pre-processing script was created to accept the words we assumed were risky and show the number of occurrences of each risky word.

**2. Background**

Before any project, market research is an essential part to building a foundation on your project and gather information on your target market. Below include several companies that already exist in contract analyzing field.

**2.1 Market Research**

After conducting market research on our specific topic, we came across applications that held the same attributes as the ones we wanted to include in our software (EULAlyzer, Terms of Service Didn’t Read, Polisis). Terms of Service Didn’t Read was the one main application that mirrored features that we wanted to implement. The purpose of performing market research was to gather and calculate ideas for our university project. It also allowed us to brainstorm ideas but more importantly inspired us to make an application that first met user requirements and secondly push the boundaries of our data mining knowledge.

**2.1.1 Terms of Service Didn’t Read**

Terms of Service Didn’t Read [2] is our biggest competitor within the market we have chosen. This is so, because it both analyses the terms and conditions of a website but also provides browser add on compatibility. This feature would naturally within your browser rate and label website policies with class ratings ranging from Class A (good) – Class E (very bad). This would then inform the user about the rights they have on the specific website they’re signing up to. From Terms of Service Didn’t Read we acknowledged that there were aspects of the application that we didn’t want to implement for the shear reason that the feature would be very time consuming, expensive to implement and would slow down production of our main features.

**2.1.2 EULAlyzer**

EULAlyzer [3] is another software that held the same concepts as the ones we wanted to implement. Created by Brightfort the software prides itself on identifying important elements of terms and conditions. Even though EULAlyzer wasn’t our main competition like Terms of Service Didn’t Read it again helped with research and understanding what is essential for our project idea.

**2.1.2 Polisis**

A smaller advertised application was Polisis [4], created by an independent developer. An application that visualised privacy policies using artificial intelligence. It highlights information that a website is collecting from you and possibly sharing to external agencies. Researching both levels of application production shows the difference in quality and outlined how we had to be patient with implementation. Researching a spectrum of applications, reassured us that there would be limitations in our project, and we shouldn’t label the lack of features as a constraint.

**3. Aims**

According to a report published by the European Commission, 66.8% of consumers had problems relating to online purchases for having insufficient knowledge of the Terms and Conditions. Furthermore, 52.7% of consumers consider their problem as serious and 57.9% did not contact or take any action against the trader due to their lack of understanding of the Terms and Conditions.[5]

Our aim was to build a web application that will simplify the process of reviewing policies. The application will have a parameter where it takes a copy of a policy, extract the risky statements out from the policy, and then display these statements to the user for them to review. In addition, the display will include statements where users might want to have a second look of. This will speed up the process of reading through policies and reduce the risk of users getting into problems.

Our core components are a concise dataset and a suitable classification algorithm. To build the dataset, we would need a pre-processing application that will meticulously run through several different policies. If the dataset produces high precision and recall scores, our classification algorithm will be able classify the risky statements accurately.

In order to get a concise dataset, we will need to review a variety of policies and contracts. These will need to come from a variety of real companies. We will mark or specify statements that we deem are risky or statements that users might need a second look. We will start off by reviewing companies who fall into the same category first, such as mobile service providers. By reviewing these, we will be able to assimilate and see a pattern on their policies and contracts. As we review more and more polices, we will be able to find and mark risky statements consistently. After reviewing a significant number of contracts and policies on mobile service providers, we will extend on reviewing high profile companies.

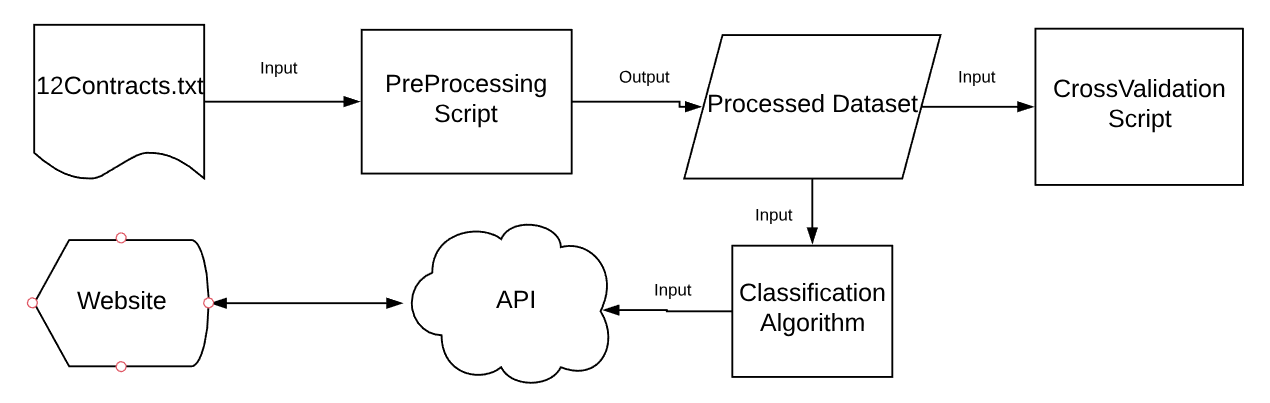
After meticulously finding the risky statements on a variety of contracts, we will need a pre-processing application that will filter, arrange and define our data. One of the main purposes of the application is to reduce 'noisy' data as much as possible. 'Noisy' data is simply data or text that is unnecessary to our data mining analysis. The application will prepare a dataset that will be used to feed in to our classification algorithm.

There are available classification algorithms that can be used for this project. In order to select a suitable algorithm, we will need to run our dataset on an algorithm and check whether it gives a high and consistent recall and precision scores. We are going to use a software called WEKA which is a collection of machine learning algorithms and data mining tasks. We will use this software to analyse different algorithms that we can use for our classification algorithm. The software's data mining task can also be used to pre-process our dataset in preparation for our classification algorithm.

**4. Development**

**4.1 General System Design**

We opted for a modular system design to our project. The modules are independent from each other and possess only the functionality required for the modules purpose.



The modular system meant that changes in the functionality of one module did not affect any of the other modules. This was especially useful in our project for two reasons.

Firstly, continual improvements were made to the ‘PreProcessing Script’ throughout the project in order to optimize the performance of the classification algorithm. If those changes forced modifications throughout the system, then resources would have been spent in rework instead of optimizing the classification algorithm.

We have 4 developers in the team, therefore the modular system allowed for an effective distribution of work. Developers could work on different areas of the system with the confidence that they will not be affecting each other.

**4.2 Language and Environment**

For the purposes of our ‘PreProcessing Script’, ‘Classification Algorithm’, ‘CrossValidation Script’ and ‘API’ we chose to use Python as our programming language. There are several reasons we chose to use this.

In terms of the system complexity, our application is relatively simple. Our application does not have to deal with threads, memory allocation issues or complex back-end server logic. Therefore, we believed that a lightweight, general purpose programming language such as Python would allow us to implement everything required but wouldn’t come at the cost of excluding group members from development because they were unfamiliar with a specific language.

Python (along with R) is quickly becoming the first-choice language for Machine Learning / Data Science enthusiasts. Python hosts a large collection of libraries that provide easy to implement and up to date machine learning algorithms. Considering the objectives of our project were focused on applying machine learning techniques rather than implementing our own, we believed Python would be a good choice for our project.

In addition, we also needed a language that would support access to our classification model through an API. We are aware that mathematical languages such as R, MATLAB and Octave do not have as many API packages available therefore we were apprehensive of choosing any of these languages for our project. A limitation of API packages means there is less fallback if issues arise with the current API package we are using. However, Python has many API packages available, providing greater security in the usability of our project.

**4.3 Preprocessing Script**

As humans we judge that attributes within a preprocessing script have no relevance to the model we wish to generate and therefore excluding them can improve both speed and accuracy results.

*See Figure A.1*

The columns (excluding ‘Is\_Risky’) represent words that the model would check the presence of, while the rows represent one sentence extracted from the privacy policies. A cell can contain either a 1 or a 0 and indicates the presence of a word in a sentence. The ‘Is\_Risky’ column labels the sentence as either Risky (‘Yes’) or Safe (‘No’).

To collect the data to be transformed by the preprocessing script, we examined 12 privacy policies and highlighted the risky sentences. For each risky sentence we would include a codeword within the sentence (0cf333), as presented below.

*See Figure A.2*

After the preprocessing script has used the nltk toolkit to extract the sentences from the provided documents, it looks for the codeword within each sentence. If the codeword is found, then the preprocessing script can store the corresponding sentence in a collection of risky sentences. If the codeword is not present, then the sentence is stored in a collection of safe sentences.

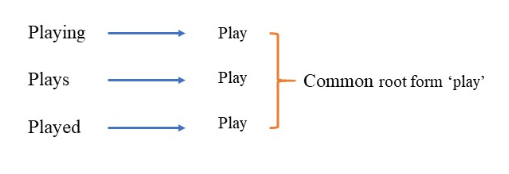
The preprocessing script then performs a filtering process on all words recorded when examining the documents to generate the attributes which will be included as the column headers. The script removes duplicates, punctuation, excluded words, numbers and stop words.

Stop words are words that are considered to add no extra value to a sentence/query when processed by a natural language algorithm. Examples of stop words include ‘your’, ‘that’, ‘the’, ‘when’. Non-stop words are usually keywords that help define the sentence, e.g. ‘information’, ‘payment’, ‘privacy’. Removing stop words can improve the accuracy of the model. By removing stop words, excess noise in the data is reduced and the feature space is focused on the most important attributes.

However, removing certain stop words can reduce the semantic information of sentences and have a detrimental impact on the model’s performance. For example, ‘not’ is commonly included as a stop word in most stop word datasets. However, consider the following example when removing

‘not’ from the sentence: *‘This website does not share your personal information with third parties.’* Removing ‘not’ from this sentence would result in a sentence which would be classified as ‘risky’ by the classifier. Therefore, in this example, ‘not’ needs to be included as an attribute so that the classifier learns during training that ‘not’ used in conjunction with a risky sentence usually results in a safe sentence.

Stemming is a technique employed by the preprocessing script to recognize the form of the word by reducing the word to its stem. An example is given below.



This improves the accuracy of the model because it reduces the number of words that may have been missed during training or testing. For example, the preprocessing script may record the word ‘playing’ 100 times when examining the documents, but only record ‘played’ a few times. Despite ‘playing’ and ‘played’ both possessing the same meaning, the preprocessing script might discount the word ‘played’ because it occurred too few times despite being relevant to the model. Suppose also, that a model is trained using a dataset where ‘playing’ occurs 100 times and ‘playing’ is an indicator of a risky sentence. However, ‘played’ does not occur once during training. A user then submits a document where the word ‘played’ occurs several times. The sentences where ‘played’ occur are misclassified as safe because the model is trained on recognizing the relationship between ‘playing’ and the presence of a risky sentence. Stemming fixes these issues by classifying all forms of the word ‘play’ under the same word/attribute. After filtering and stemming has been applied to the sentences in the privacy policies, the risky and safe sentences are re-examined and the presence of attributes within each sentence are recorded before being exported to a csv format to be interpretable by a classification algorithm.

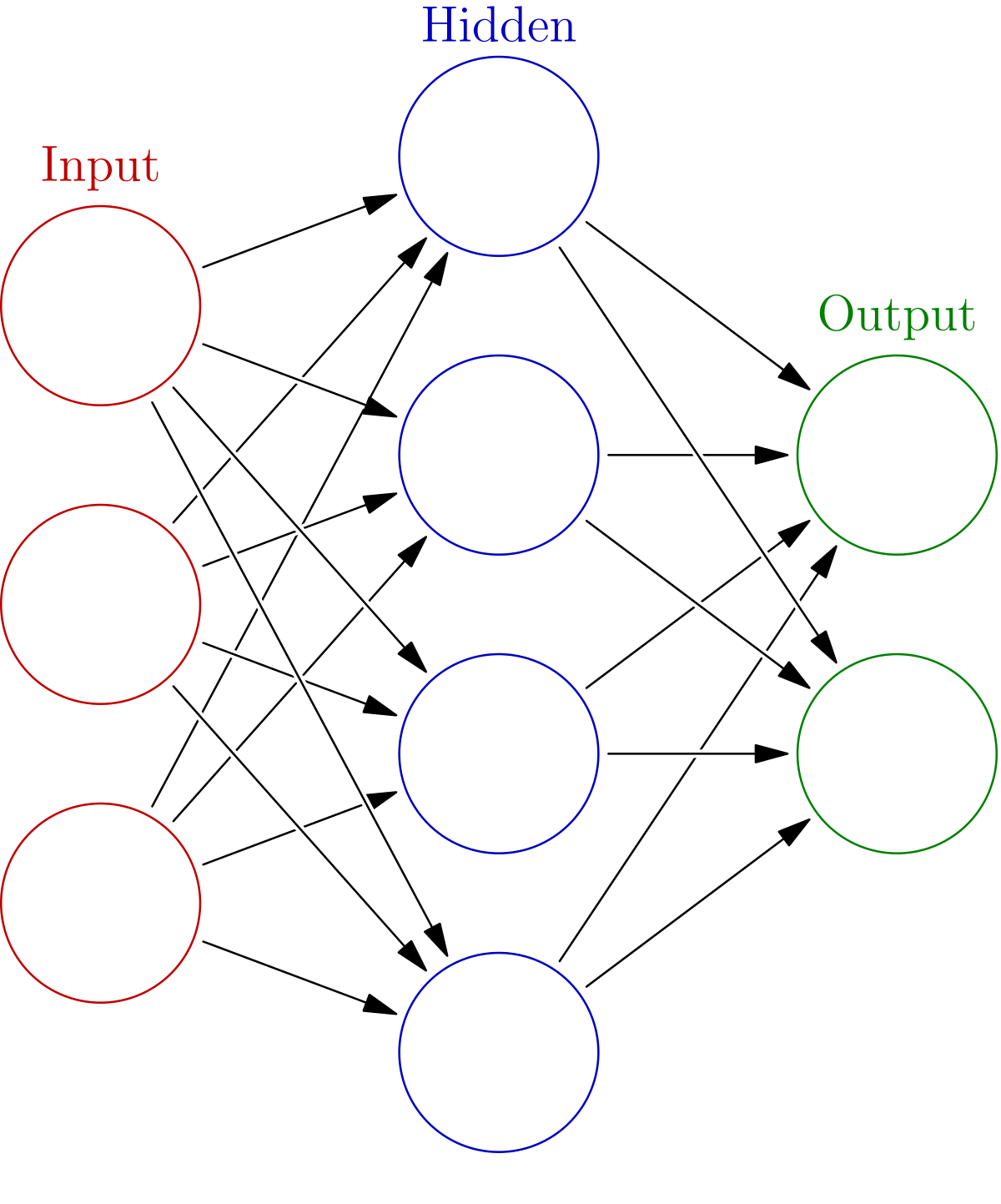
**4.3 Deciding on the classification algorithm**

Our first hurdle when deciding the classification algorithm was understanding which classification algorithms were applicable to our project. We decided against employing any unsupervised learning algorithms as we believed the differences may be too subtle for the classifier to distinguish between risky and safe sentences. As our project requires a classification problem, we excluded any algorithms that generate continuous output.

The remaining supervised learning algorithms available to us were:

* Neural Networks
* Support Vector Machines (SVM)
* Decision Trees
* Naïve Bayes
* Logistic Regression

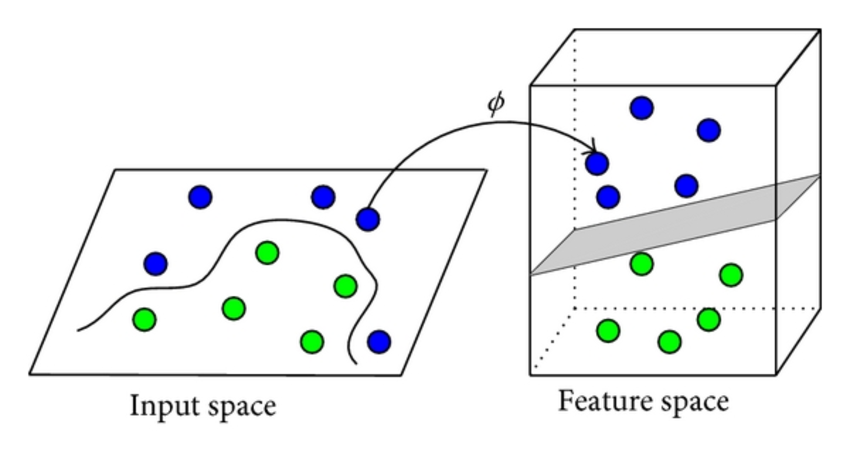
A Neural Network is a machine learning algorithm which consists of multiple units in multiple layers with each unit connecting to every other unit in the next layer. The structure of a neural network is always composed of one input layer, at least one hidden layer and an output layer. From a high-level perspective, the purpose of hidden layers is to combine the weighted inputs from the first layer in order to capture additional complexity from the data.



We believe that a neural network would be a good algorithm to evaluate more closely because often the distinction between a risky sentence and a safe sentence is very subtle. Therefore, a more complex model may be required to differentiate between the nuances in the data.

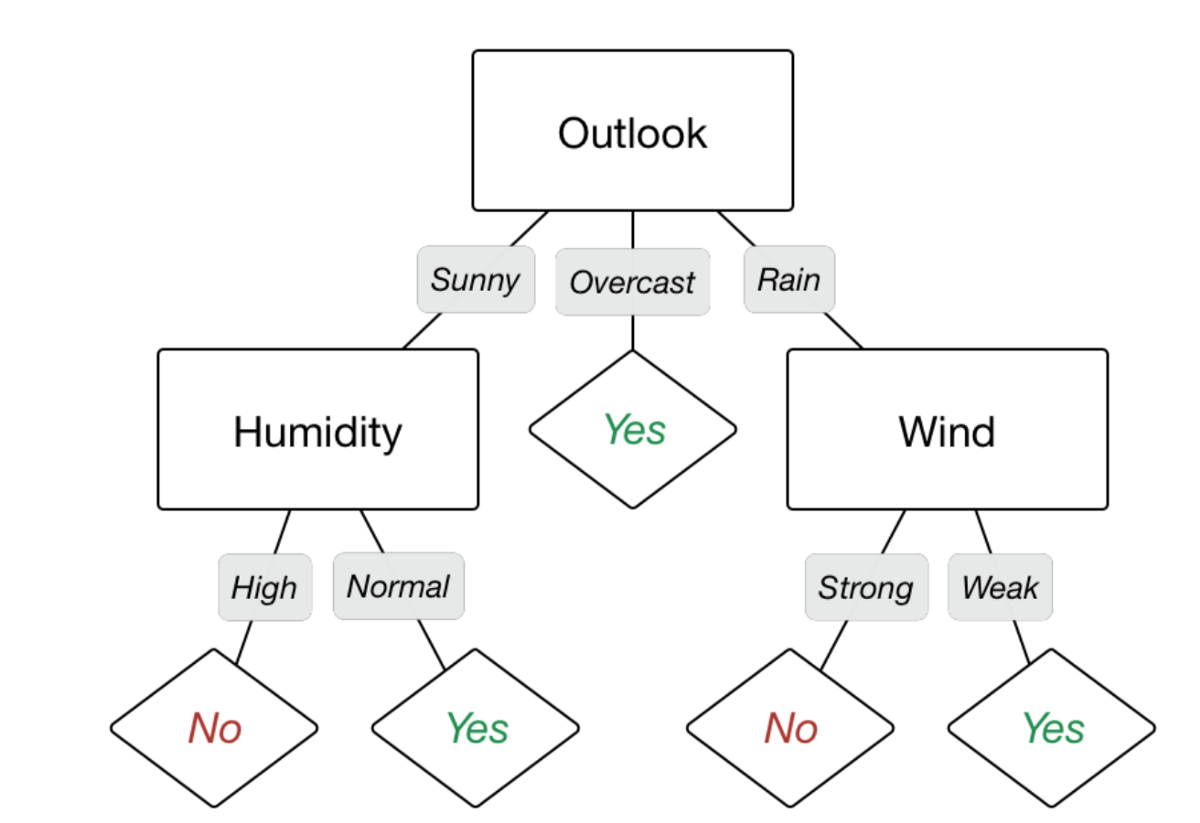
A Support Vector Machine separates two classes by finding the optimal hyper plane between them. The primary advantage of SVMs is that they can separate a non-linear function easily. By using a technique called the ‘kernel trick’, non-linear space can be transformed into linear space, enabling the two classes to be separated easily. For example, suppose there are two classes which are governed by the function f(x^2).

Currently these classes are non-linearly separable. However, when the input space is transformed by the sqrt(X), the separability between the two classes becomes linear.



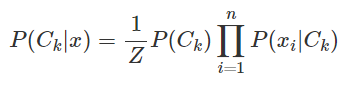
Our task may require a non-linear boundary between the two classes due to the complexity of the classification problem therefore we considered this algorithm for further evaluation.

Decision trees in machine learning are constructed by examining the entropy of the attributes. Entropy is a measure of the homogeneity of the samples. An entropy of 0 indicates high homogeneity while an entropy of 1 indicates low homogeneity. Attributes of low entropy are more favorable to a decision tree because they offer a more attractive split to help distinguish between the two sets of classes. For example, an attribute of ‘gender’ which has 4 examples labelled to the positive class and 4 examples labelled to the negative class does not provide us anymore insight regarding our ability to distinguish between the two classes.



Although Decision Trees struggle to match the complexity of Neural Networks or SVMs, the interpretability of the model appealed to us. The model’s decision process can be revealed to users of the website to support a more transparent application.

Naïve Bayes is a simpler learning algorithm that focuses on utilizing bayes theorem to calculate the probability of a class. A new example is classified by multiplying the calculated probabilities of discrete attribute values occurring for a certain class, then comparing the returned probabilities of the different classes occurring given the example.



The Naïve Bayes classifier assumes independence between its attributes. Therefore, each attribute’s contribution to the model can be interpreted by examining the calculated conditional probabilities. This makes the model very interpretable for a human examiner. In addition, a Naïve Bayes classifier is inexpensive to train. For these reasons, we decided to investigate this algorithm further and evaluate its performance using our dataset.

Logistic Regression uses a sigmoidal function on the attribute values to differentiate between the two sets of classes. However, we did not decide to perform further evaluation on this algorithm. As a group we decided that its decision boundary was too close to that of SVMs, while SVMs provided better classification results in nearly all cases.

**4.4.1 Classification API**

The motivation for developing an API is so that our website can communicate with our classifier and is essential for enabling normal users to analyse privacy policies.

When the API is started, it runs the classification algorithm to generate a predictive model. Once a request is received, the API parses the given document and extracts the sentences. Each sentence is then run through the classification algorithm to obtain a prediction (risky/safe). All risky sentences are returned to the website in JSON format so they can be displayed to the user.

**5. Testing**

**5.1 Web Testing**

**5.1.1 Functionality Testing**

We made sure that the website is valid by running it through <https://validator.w3.org/>. This also checks for HTML syntax errors. All buttons on the website have been tested to make sure they correctly perform their tasks. Since the website is only one page, we made sure that the appropriate buttons navigate to the correct section of the page.

When the ‘Analyse’ button is pressed, it checks whether the text area is empty or there is enough text/document to be analysed. This prevents the website from sending empty or insufficient data to the server. In addition, it is of capable of catching error and displaying them to the user if it arises.

**5.1.2 Usability Testing**

The content of the website is proofread. We used a web application that strips the html tags and proofreads the content. The colours on the website are plain and do not contrast with one another. In Addition, a site map for the website is unnecessary.

**5.1.3 Compatibility Testing**

The website can run on different browsers, mainly Internet Explorer, Google Chrome, Mozilla Firefox and Safari with the most recent versions. It is also compatible with tablet and mobile devices. The @media is a CSS feature where it detects the size of the browser window and automatically adjusts the content and style of the website to adapt with the browser maintaining high quality content throughout the website.

**5.2 Quality of the Dataset**

It was vital for the project that the dataset used to feed in to our classification algorithm is meticulously checked for every changes and additions we make. In order to keep track of the quality of our dataset, we used WEKA to check for precision, recall and F-measure scores. Our classification algorithm (i.e. our Neural Network that we implemented) can also check these scores for our dataset. These scores will be our indication on the progress and quality of our dataset. The higher these scores the better quality our dataset. For every change we make on the dataset, we check the scores. As a result, we found out that the more contracts we evaluate and add to our dataset, the better the scores.

**6. Challenges**

**6.1 Accuracy**

One major challenge that we experienced was producing accurate data. This involved accuracy in our pre-processing data and making sure that all the data collected followed the same guidelines. Originally, we had an issue with contrasting data where we all had conflicting views on what was considered “risky”. So, when going through our data collection process we gathered a lot of data that wasn’t consistent. This was a problem for us but in order to resolve this problem we created Terms and Conditions guidelines to keep us focus on what should be considered “risky” and prevent us from swaying in opinions.

**6.2 Expectation exceeding reality**

No matter how much we planned, gathered more data or time spent on our project, we always came with the overwhelming feeling of have with set the bar too high for what we expect to complete in the year project. As the idea was based around machine learning, a quite new aspect of computer science, we generally came across the hurdles of either the knowledge scope of development for the analyser was way over what we knew as students or the time we had to implement such algorithms exceeded the time we had for the project. Even smaller features of the project like collecting data became a longer process than we first imagined. Looking back ideas we brainstormed at the start of the project could’ve been refined to help us progress further into the analyser and expand its means of functionality.

**6.3 Algorithms**

Another challenge that we faced was choosing the right classification algorithm to apply to our model. With such a wide range of classification algorithms within machine learning e.g. Linear regression, decision trees, neural network, naïve bayes etc. we had a real problem with choosing an algorithm that met our model requirement.

To have a model that easily could distinguish between “risky” and “non-risky” sentences we needed an algorithm that produced high accuracy and recall results. Accuracy being total number of risky/safe statements correctly identified by the classification model. Recall being the number of examples that have been classified correctly. We as a group had a real issue with low results from a lot of the chosen algorithms and were struggling to find an algorithm that produced a high enough accuracy and recall results to meet requirements. One of the solutions to this issue was to collect more training data i.e. review more company privacy policies but due to time constraints this action wasn’t feasible.



**Conclusion**

**7.1 Project Reflection**

To conclude we built a system that was able to highlight risky statements within documents using a neural network and some classification algorithms. Overall as a group we feel as though we were able to hit the targets that we envisioned. The final product produced was able to run in the way we expected it to. Documents would be analysed by giving the user feedback which entails all risky statements in the document.

The website was aimed to be user-friendly which as a group we feel it served its purpose. In relation to the risky statements, the accuracy and recall scores we finally improved, were able to produce relevant statements that the group deemed risky enough to fulfil the purpose of the product.

The project idea was one we realized was unique, as the idea of checking through documents was not common across other projects. The main aim of the product was to allow users to check any documents they may feel contains risky statements. The novelty of the idea for our product formulated from the general misconception giving to reading through tedious number of documents such as contracts to identify risky statements. Within the duration of the project, the most fundamental thing was to ensure that we understood how the classification algorithm works and data mining. The group would advise if anyone was to undertake a product like this, it is important to review the topics that are going to be highlighted and gain understanding of them.

**7.2 Future Improvements**

The possibility of having extra time with the project enables us to complete certain tasks that time permit to improve unambiguousness within risky statements generated. Below briefly highlights some adjustments and improvements that we could consider implementing in future modifications.

**7.2.1 User Feedback**

Another future improvement could include sending the highlighted risky statements found to the users email so they can further evaluate it. Further information of the results obtained can be sent to the users via email or social media. Making registration compulsory would mean users would have an account which would further enable us to send them updates and new features added to the analyser.

**7.2.2 Browser Add-Ons**

Future enhancements would include making the analyser a web browser add-on. The browser ad-on would be supported by Internet Explorer, Mozilla/Firefox, Opera, Google Chrome. Its purpose would be to provide quick access to users and alternatively enable them to rank a This idea was something we contemplated undergoing as a project after the termination of the project.

**7.2.3 Cost sensitive classification**

If time permitted, for us to improve the results of our classification result, we would have used cost sensitive learning. All our misclassifications are treated equally meaning there is no reward for identifying the minority class over the majority class. Cost sensitive learning would enable us to penalise misclassifications of the minority class hoping to increase True Positive for the majority class. This would be done by changing the default weights of both the False Negative and the False Positive. Changing the default weight (known as punishing) to a higher rate for False Negatives, then we would be able to generate better recall. Higher rates given to False Positive would also help in improving Precision.

**7.2.4 Testing and Results**

Additionally, if time permitted, we would have liked to work on testing many other different classification algorithms in order to find the best suited one. In addition, if we had amplitude amount of time, it would be relevant to improve the accuracy, recall and F-measure scores to make the analyser be more specific and critical in its analyzation of risky statements. Although the final scores we received were somewhat reasonable we feel improvement can be made.

**Final Words**

To sum up, our project prompted a deeper inquire into data mining and neural networks skills and we were able to achieve the core aims of our project; a product that would be able to identify risky statements in documents. We as a group accept there is a wide area that could not be explored surrounding neural networks and data mining techniques and we hope this project ignites a greater interest in this field.

**Acknowledgements**

The team would like to thank Alex for his support and showing great interest in the success of the project. The team would also like to thank all the companies whose policies were used.

**8. References**

[1] <https://www.adweek.com/digital/survey-many-users-never-read-social-networking-terms-of-service-agreements/>

(Viewed 15/2/2019)

[2] <https://tosdr.org/>

(Viewed 6/2/2019)

[3]<https://www.brightfort.com/eulalyzer.html>

(Viewed 6/2/2019)

[4] <https://pribot.org/polisis>

(Viewed 6/2/2019)

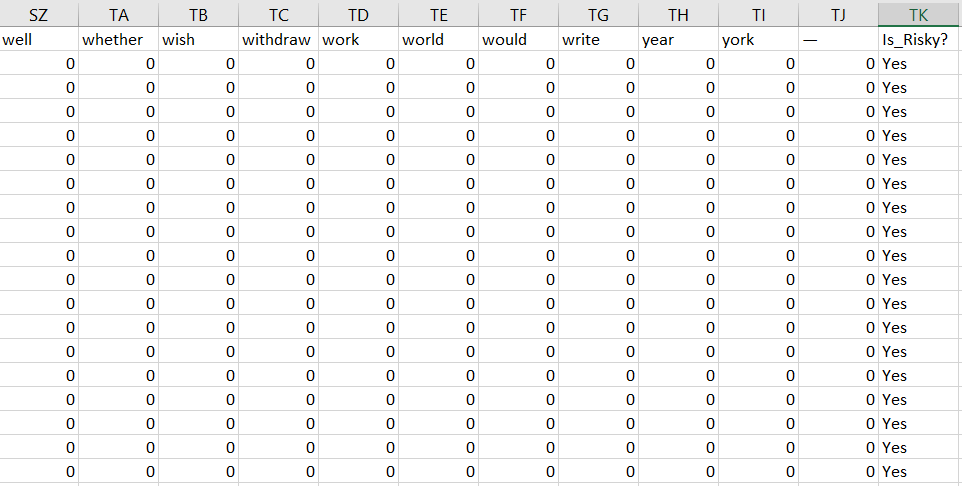
[5]<https://ec.europa.eu/info/sites/info/files/terms_and_conditions_final_report_en.pdf>

(Viewed 20/3/2019)

<https://machinelearningmastery.com/improve-deep-learning-performance/>

**Appendices**

**Figure A.1**



**Figure A.2**

