Detecting User Engagement Using Mouse Tracking Data

Prifysgol Abertawe
Swansea University

David Saunders (910995)

Department of Computer Science
Swansea University

This dissertation is submitted for the degree of
Master of Science

September 2020

Declaration

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

David Saunders
September 2020

Statement 1

This dissertation is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by giving explicit references. A bibliography is appended.

David Saunders
September 2020

Statement 2

I hereby give consent for my dissertation, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

David Saunders
September 2020

Summary

This project explores the use of Hidden Markov Models to determine if users are paying attention during a crowdsourced study. The users are split into 2 distinct groups, online crowdsourced users, and lab study users. We propose that the lab study users were paying attention, and the crowdsourced users may or may not be paying attention. By using observed data recorded from the users mouse cursor we are able to model both groups interaction with the system with separate Hidden Markov Models. Some users interaction seem misplaced compared to their groups. These are reclassified with the aim of changing the groups of user from lab study or crowdsourced, to users paying attention and not paying attention. It was found that potentially 9.8% of turk users were paying attention, and 43% of lab study users may not have been paying attention.

Contents

1 Motivation

Crowd-sourcing marketplaces like Amazon's Mechanical Turk are popular services that provide a method for researchers to get participants to complete human intelligence tasks [1]. A common use for this technology is to label data for use in training for machine learning algorithms [2]. They can also provide a cheap, scalable method for scientists to gather responses in research. The level of user engagement, attention, and low quality responses can all be issues when gathering data from participants with distributed approaches [3]. The primary motivation of this project is to develop a system of potentially identifying if a crowdsourced user is paying attention during a task.

The use of gathering responses using Amazon's Mechanical Turk and other crowdsourcing alternatives are becoming prevalent across disciplines. It is commonly used in conducting clinical research [4] and it is estimated that almost half of all cognitive science research involves the use of crowdsourcing services to collect data samples [5]. The creation of an easy to implement method to measure user engagement would massively help researchers to increase reliability of their research.

Research has been conducted on how the accuracy and attention of crowdsourced tasks can be increased. Methods such as offering financial incentives [6] and engaging a users curiosity [7] have been found to motivate workers into performing better at crowdsourced tasks. Despite the research there is still debate as to which method is superior. If user engagement can be effectively identified then the best method of ensuring user engagement could be found using this method.

Measuring user engagement is well studied in the field of web analytics [8]. However the existing methods of reporting and analysing website data cannot be easily applied to crowdsourced tasks. Characteristics such as session duration and customer satisfaction are used as proxies for engagement. However a longer session doesn't necessarily mean a more engagement user and customer satisfaction is not applicable for crowdsourced tasks. Therefore existing solutions will not work and now methods must be evolved.

1.1 Background

The work in this project is build on data gathered from a previous study; *"A comparison of user interactions in lab and crowdsourced studies"*[9].

The study was centred in the field of visualisation. The paper explored whether an alternative more detailed interface could be used help improve a users effectiveness at completing a task. The task simulated an investment scenario, the aim was to maximise the return of a collection of stocks over a number of iterations.

Figure 1 shows the main interface of the lab study, with the risk and return of stocks to the right and allocation sliders to the left. Users were asked to invest their money by putting it into varying amounts of stocks. The 5 stocks were labelled from A to E. Depending on the makeup of the users stock portfolio,

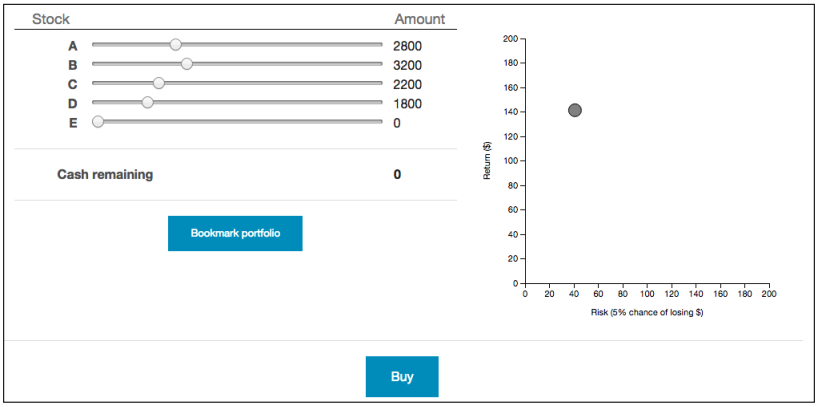


Figure 1: The user interface of the lab study.

the plot would update showing the return and risk of their portfolio. They were then asked to confirm their choice by buying these stocks. This process was repeated for 5 steps, where after each step the users portfolio value with increase based on their previous investment decisions.

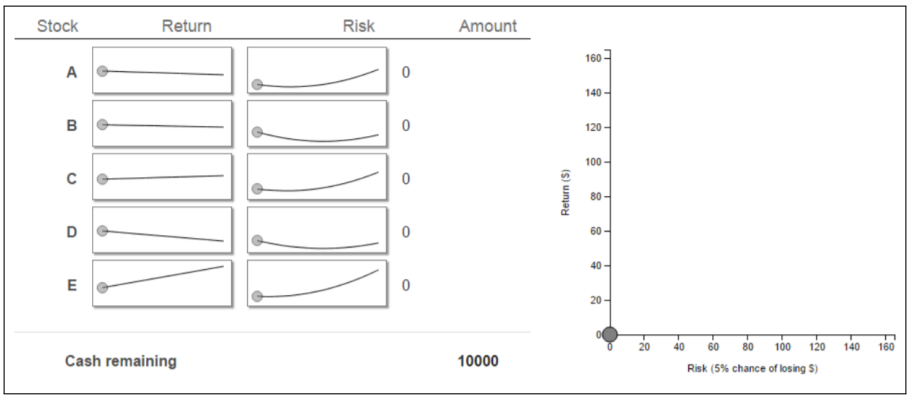


Figure 2: The alternative user interface of the lab study.

Figure 2 shows the alternative interface design for the study. The slider representing each stock was split into 2 separate sliders, showing the return and the risk of the stock. Additionally the sliders are no longer 1 dimensional, both sliders are 2 dimensional line charts. The assumption of the study was that these extra visualisations would help users to maximise the value of their portfolio.

Participants from the lab study were heavily monitored to ensure they were engaged, focusing on the task. The crowdsourced participants were not monitor so we're unable to say that they were engaged with the task, and not looking at another screen.

Table 1: Data collection methods used in the study [10].

Data collection method	Number of participants	Were participants paying attention?
Lab study	18	Yes
Crowdsourced task	370	Unknown

Table 1 shows the two different ways in which data was collected. Only 4.6% of the data was gathered in person, with the vast majority of responses being crowdsourced. The difference in number of participants shows one of the reasons crowdsourcing is popular. It is much easier to crowd source responses than it is to organise an in person lab study.

For the purpose of this project the distinction of which interface a user was using ignored. The study was inconclusive in proving that the alternative interface design had an meaningful effect on the success of a users portfolio. Additionally rationalise of this is that regardless of which variant of the interface a lab study user was using they would be paying attention. There is likely to have crowdsourced users who are engaged and some that are not engaged regardless of which interface is used. While there will be differences in users mouse data depending on the interface , I hypothesise that the difference will not be as large as the difference between users that are engaged or not engaged.

1.2 Initial attempt

One naive assumption might be that a users attention level might be inferred from the time taken to complete the tasks. Exploration of the data showed that the division of users may not be so straightforward, shown in Figure 3.

The lab and turk user classes cannot be separated by a simple metric of their data, such as length of events sequence and time to complete task. This rules out distance based machine learning approaches such as K Nearest Neighbours or a Support Vector Machine as there is no relationship there. It should be noted that length of events sequence and time are correlated with a pearson's correlation of 0.344.

As a rule of thumb we can say that correlations below 0.3 have no relationship, and values between 0.3 and 0.5 have a weak relationship [11]. This means the 2 features are weakly correlated, but correlated nonetheless. This means that one of these features may be a good substitute for the other and that having both may be redundant.

1.3 Contributions

In this project my contributions to fix it are, -a system to classify users, a way of visualising their mouse paths, ways to directly and quantitatively compare different users, and a multistage semi-supervised based binary classification output to answer the question of 'are users engaged'.

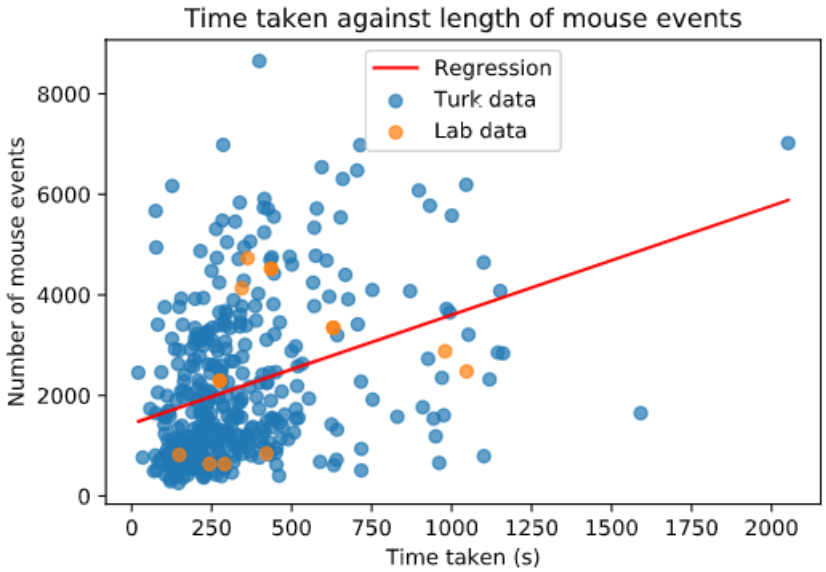


Figure 3: Scatterplot of users.

2 Related Work

All related work about actual other attempts to detect user engagement from mouse tracking data. Other subsections will be like miniature literature reviews or something?

2.1 Eye tracking

Non-verbal information such as eye tracking may be used to detect user's level of engagement [12]. Vision is one of the most powerful human senses so it has the potential to give a good measure of user engagement. The methodology of eye tracking is that we move our eyes to focus on particular areas that we want to see in more detail, and divert our attention to that area [13]. Thus tracking a user's gaze can provide insight into which part of a system they're engaged with, and how much so.

Eye tracking data can be used to show user interface elements that users focus their attention on as shown in Fig 4. From this researchers were able to predict the amount of attention elements of the page would receive. By observing what parts of an interface users are interacting with we can determine what a user is engaging with [14]. Eye-tracking has been used, and found success novel applications such as recording the engagement of users when playing a game. Tracking users eye movements helped game designers understand how users can recognise interactable game objects and could be used to investigate problematic game design issues [15].

Figure 4: Heatmap showing popular locations of users eyes on a webpage [14].

Eivazi and Bednarik extracted features from a users eye tracking data to determine their cognitive state in a problem solving exercise [16]. Features such as "mean fixation duration" and "total path distances" were engineered, and users were split into classes based on their performance. Given a user feature set and their performance class it was able to classify their cognitive state during the task with a 87.5% accuracy with a support vector machine.

Szafir and Mutlu identified a plethora of verbal and non-verbal behavioural cues used by teachers in an educational setting, with gaze being identified as one of them [17]. The behavioural cues could not be recorded directly by a computer, instead EEG signals measured from a headset were used to measure engagement.

Eye tracking is not however a perfect solution and its limitations have been well documented. Track subjects eyes with a good degree of accuracy requires the use of expensive, intrusive equipment that frequently needed recalibrating [18].

2.2 Mouse cursor tracking

Research has also found that there is a correlation between a user's gaze and their cursor position. The position can be considered a "poor man's eye tracker" as it has been found that eye gaze match mouse position 69% of the time [19]. Mouse movement data can be collected without the drawbacks of eye tracking and with more automatic methods, meaning more data can be collected, and on a larger scale [20]. Therefore it can be said that mouse data can be used as

a good alternative to eye tracking data.

By using mouse data it is possible to unobtrusively record a user's normal use of a web browser without disturbing their experience [21]. It has also been found that users tend to follow the text they are reading with the mouse cursor [22]. It can be determine what paragraph of a page was being read with an accuracy of 79% by using mouse cursor data [23].

Other methodologies explored ways of classifying user engagement from eye and cursor data, however it is also possible to predict users attention and user frustration in complex webpages [24]. Not all studies agree that mouse cursor is always a good approximation for eye data. Hauger et al found that distinct cursor behaviour exists depending on the task, and that the relationship between eye gaze and mouse position is more nuanced than measuring only mouse data [25].

2.3 Crowdsourcing Cheating

Cheating within crowdsourcing system is a well established issue with the concept REF. Users use a variety of methods to maximise their HITs. These are a few examples.

Crowdflower has an integrated method of detecting cheating. This method was used effectively when using human workers to help label datasets [26]. Tasks for which the answer is known are periodicity asked to the user, and the accuracy is measured. Once a user has completed a number of tasks over a threshold value then their accuracy on the known tasks is assessed. Of this accuracy is deemed to be unacceptable then the results from that user are ignored, and they are prohibited from contributing any more. This system has the drawback of taking more time and being more expensive than with no cheat detection method. Out of the 15 unique users that completed some of the tasks 9 of them where labelled as cheaters, presumably using scripts to randomly select answers. This allowed them to create a much larger volume of responses than any legitimate users, with 91% of the results coming from cheaters.

Crowdsourced tasks can be classified as Closed Class tasks and Open Class tasks, with each having unique cheating approaches [27]. Closed class questions are more common, requiring workers to chose an answer from a predefined list. This includes check boxes, buttons, multiple choice questions, and importantly to this project, sliders. Randomly picking answers is the most common cheating attempt, this can be detected by comparing a suspected users responses with a user known to not be cheating or comparing with all other workers answers. Open class questions are much easier to cheat, these involve giving a user much more freedom to complete a task. Typically workers may be given blank text fields to fill in, or blank canvas' to draw on. Standard cheating approaches involve leaving the fields empty, or repeatedly entering the same text. These approaches can be easily detected. More advanced attacks copy domain specific text copied from the internet, which is difficult to detect.

2.4 Spam detection

2.5 Semi Supervised Learning

This will be a key aspect of the project as we only have definitive labels for part of our data. This reflects the challenges of real world data, by some estimates only 2% (made up number) of all data is structured and labelled, the rest is unstructured [find reference].

Semi-supervised learning is the study of combining both labelled and unlabelled data to improve machine learning techniques [28]. One of the most popular semi-supervised learning techniques are Semi-Supervised Support Vector Machines (S3VMs).

Mention semi-supervised SVMs here. And the abbreviation semi supervised learning (SSL)

Reference [29] as it looks to be massively influential with 4,000 google scholar citations.

The data used in this dissertation is labelled. All data belongs to 2 classes, a crowdsourced turk user class, or a lab study user class. However on the other hand the data is not labelled for the task I would like to explore. The goal of this project is to identity which users were paying attention. We have assumed that we can infer that lab study users will be paying attention, so we can say that those samples are labelled. However the rest of the samples we have which are all of the online data are unlabelled. Therefore this is a kind of semi supervised learning problem where we only have a small percentage of our samples labelled, and only confident labels for one class. (See if a paper on this exists, semi supervised binary classification with only labels for one class.)

Semi supervised learning has already been applied successfully to the field of user engagement. During a 2017 study video data of secondary school children was recorded during a set of tasks, with the aim of classifying pupils between the states of 'Engaged' and 'Disengaged' [30]. Rather than annotate all data by hand which would be very time intensive, 1000 frames of video were selected at random and hand labelled. This led to a split of data labelled 563 engaged and 437 disengaged. AU facial recognition features were extracted from the pupils faces in the video frames, and important features decided with a Principal Component Analysis method to keep features totalling 95% of the variance.

Their chosen method of semi supervised learning is a safe semi-supervised support vector machine. This variant of a semi supervised support vector machine finds a better separator when there is a small volume of labelled data, but a large volume of unlabelled data. The SSL model was trained with numbers of labelled data from 50 to 500, and and equal number of samples from both classes. The AUC metric ranged from 0.65 to 0.733 depending on the quantity of labelled dta used, outperforming a traditional SVM.

Unfortunately as shown in figure 3 the data here cannot be spatially separated and therefore any SVM would not be successful on the data. It is still useful however to see how semi-supervised learning has been applied to the field of detecting user engagement and how it has been successful.

2.6 Hidden Markov Models

A Hidden Markov Model will be the main statistical modelling technique used in this paper.

In the case of this project, the hidden states X would be the users thought processes including whether or not they're paying attention. The observations, Y, are the data that has been collected from the users such as mouse location over time and sequence of interface elements.

Paper Tom send me about HMM for text classification that I might be able to use [31].

Paper claims to use HMM for spam detection, I think they actually just use it to detect misspellings of words or something which is used by spam to hide from filters [32]. Probably could find a better spam detecting HMM, I just like how this almost does something different from the title, which my diss will end up doing.

This paper was recommended by someone on stack overflow as an old influential paper in the field with tens of thousands of references [33]. Called a tutorial on HMM and its so old so original source. Would definitely be good to reference if I include any of the mathematics behind HMMs.

Someone's dissertation on the topic of generating synthetic data with HMMs. This can be a good way to create synthetic data [34].

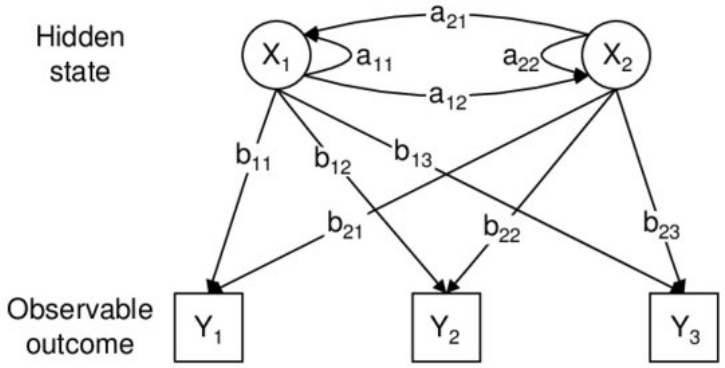


Figure 5: Diagram of an example Markov Chain model [35].

Figure 5 shows the states of a markov model of my system with states A-E representing sliders 1-5 and state F representing an html element. In the top right we can see a possible transition matrix of our system.

Cut from repeated experiments section.

HMMs are typically used with time series data. Here I just used the series of mouse events and ignored the timestamps of the event. It was found earlier that the length of mouse events and time are positively correlatied with a pearsons correlation of ?? 0.6, and so just the sequence of mouse events should still hold the detail that time data would.

With the bag of words model we ignored all order to the sequences looking only at the frequency of each item. This is a different approach, here the order is the most important thing.

3 Methodology

The assumption of this project is that labs are paying attention and turks are not. Therefore if we get any outliers from the turk data, but that are similar to the lab data, then we will say that that online turk user was paying more attention than his peers, and that they were paying attention.

This study (ref) says that only 10% of all people / turk users pay attention during a task. We will look at the 10% (30ish) of the turk data that looks like it is lab data and day that they were paying attention. This is just the assumption we have made for the project, unfortunately the dataset isn't extensive enough for us to fully test this hypothesis.

4 Data Pipeline

When planning and completing this project many decisions were made about the steps taken to convert the raw data to a finished product/classification. This section may act as an overview of the project, detailing the different sections of work, what they may contain, and the order in which they will be completed.

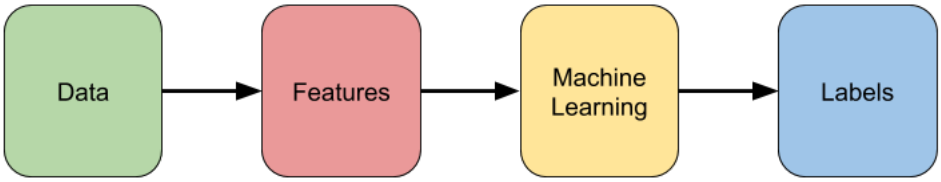


Figure 6: Diagram of the Data Pipeline of the project.

4.1 Data

The first component of the project has been completed, and data has been extracted form JSON format to a csv format.

Table 2 shows us the features of the data. The lab study data and the crowdsourced data have the same schema, except lab results have a different ID field.

The target field shows us which element in Fig 1 a participant is interacting with and event_type details the type of interaction. Time field shows the time taken in seconds since the first recorded mouse event. We can hypothesise that

Table 2: The first 5 records of results of the crowdsourced task.

event_type	target	time	x	y	step	turkId
mousedown	alloc-slider-1	0	477	405	1	A35YFAFWP33C70
mouseup	alloc-slider-1	0.111	478	405	1	A35YFAFWP33C70
click	alloc-slider-1	0.111	478	405	1	A35YFAFWP33C70
mousedown	alloc-slider-1	1.516	479	405	1	A35YFAFWP33C70
mousedirchange	alloc-slider-1	2.395	543	403	1	A35YFAFWP33C70
mousedirchange	alloc-slider-1	3.161	594	402	1	A35YFAFWP33C70
mouseup	alloc-slider-1	5.048	514	407	1	A35YFAFWP33C70
click	alloc-slider-1	5.048	514	407	1	A35YFAFWP33C70
mousedown	alloc-slider-2	5.461	494	441	1	A35YFAFWP33C70
mouseup	alloc-slider-2	5.513	494	441	1	A35YFAFWP33C70

participants with a shorter time paid less attention than a participant who took much longer, thinking about their actions more. The x and y fields show the location of the mouse and step shows which stage of the task, from 1 to 5, a participant was in.

4.1.1 Data Manipulation

Here I will summarise what I have done to the data. As previously mentioned the data was gathered through a lab study and an online crowdsourced study. The purpose of the online study was to gather a larger amount of data that was possible to do so in person. Data was recorded in a JSON formate with lots of irrelevant and duplicate data relating to the users background and not their mouse movements. JSON is a form of unstructured data, meaning it is not fully structures but has some organisation to it [36]. To use the data in this project it first had to be processed into structured data. This was challenging and time consuming as Python's JSON library was unable to directly convert the data. This was because the mouse events were stored as a nested JSON dictionary and there were additionally errors with the way the data was logged causing it to be invalid JSON. After these challenges were addressed the data was converted and saved to a Pandas DataFrame, and then to a csv with over 100,000 lines. Errors were found in some of the data and so the final number of usable data is less than the figures shown in 2. Example errors include missing mouse events, or time to complete task being recorded incorrectly as 1.4×10^{12} seconds or $31,688$ years, obviously incorrect.

This leads to the problem of imbalanced data samples. After erroneous data was removed there were only 14 lab data samples and 461 online data samples. This means that there were over 30x as many data samples from one class compared to the other. As stated in my assumptions, we can say that the lab participants were paying attention, where as the online participants may or may not have been paying attention.

If the classes were balanced then simple approach may be to treat this prob-

lem as binary classification problem. Using something like a Support Vector Machine we could classify a given point as lab or online / paying attention or possibly paying attention based on their proximity to other data points. To do so we would need to have balanced classes otherwise the algorithm would have a high accuracy from just classifying everything as possibly paying attention as that is the most frequent class. However due to the imbalances, alternative methods are required.

4.2 Features

This part of the pipeline refers to what features I am going to extract from the data. Features of data can be defined as 'attributes or interesting things from the data'. [reference] These will consist of both raw and created features, but what do I mean by this? Raw features will consist of the the number of mouse events recorded, while a created feature could be comparing the trace of users cursor data when using the program.

One such feature recorded in the data is mouse target. This refers to the HTML element that a users mouse was over at any given time. As shown in Figure 1 the most prominent part of the interface that users will interact with is the 5 stock allocation sliders. It is hypothesised that users who are paying attention may spend more time fine tuning their stock allocation and therefore would have more mouse events with the sliders as the target. An attempt to pick up on different parts of the interface was achieved, however the html elements were not clearly named and thus this aim was unachievable.

The features chosen to be used by the HMM is the target of

4.3 Machine Learning

Once we have insightful features from the data we can consider what machine learning algorithm would be most appropriate to use on the data. This will obviously be highly dependent on what form the final features are in. For example if the features are numerical values such as time taken to complete task and number of mouse events then an algorithm such as a Support Vector Machine could be a good choice. If the data is in the form of sequential data such as a list of all mouse events then alternative techniques such as LSTM or RNN networks or HMMs would be more appropriate choices. If the features output was an image such as a trace of mouse position over time then a CNN could be a good choice as they're designed for image data. It is likely that text classification algorithms will be used when comparing the targets of mouse events. Comparing n-grams can be done with algorithms such as XYZ [reference]. Other text classification algorithms such as cosine similarity or sentiment analysis could also be used.

4.4 Labels

Lastly an important section of the pipeline, as it reflects the final outputs of the system. Labels will refer to which users are classified as paying attention and which users are not. A key aspect of this project will be semi-supervised learning. That is we have some data points we can confirm were paying attention, and others where they're level of attention was questionable. Once we have an algorithm that can classify some users as paying attention or not we can rerun the algorithm with these preliminary outputs as new training data. If this is done recursively then we can end up with a system that can split all data points into the 2 classes, perhaps with a degree of confidence given as a percentage.

5 Final Implementation

To implement a Hidden Markov Model the python library HMMLearn was used [37]. The sequence data used to train the models was the order of mouse targets. The temporal aspect of the data was ignored, but earlier research showed that number of mouse events and time were heavily correlated. Therefore using just the order of mouse events should capture the temporal aspect of the data.

When implementing a Hidden Markov model on the actual data there are lots of parameters that must be considered. Given a sequence of input data the HMM model can calculate the parameters of the model such as the transition matrix, emission probabilities, and initial probability distributions by using the Viterbi algorithm [37]. The most prominent parameter that must be given is the number of hidden states to use.

The paper [38] addresses this issue on similar data. The authors use a HMM to model animal movement behaviour, where they hope the hidden states would roughly relate to the behaviour of the animal. For example a 2 state HMM on this model may relate to "foraging/resting" and "travelling". Humans can be more complex than animals but we can think of potential states of being "inquisitive" and "bored" They trained models with 2, 3, 4, and 5 hidden states and compared the criterion's of AIC, BIC, ICL. Similarity to evaluate Lab and Turk models models were trained with number of hidden states ranging from 1 to 11. Generally an increase in states should increase the effectiveness, however it will also take longer to train. All models used for this comparison was trained to 50 iterations for a fair comparison.

Figure 7 shows a different criterion of the total log likelihood of the training data of the model. This value is normalised by dividing by the number of data samples, to give a per-sample total likelihood. The evaluation of each HMM model does fluctuate based on the random state used, which is why there are drops in the evaluation criterion even when more hidden states are used. We can see that the criterion for both models stop improving after 9 states. Therefore a 10 state model was used.

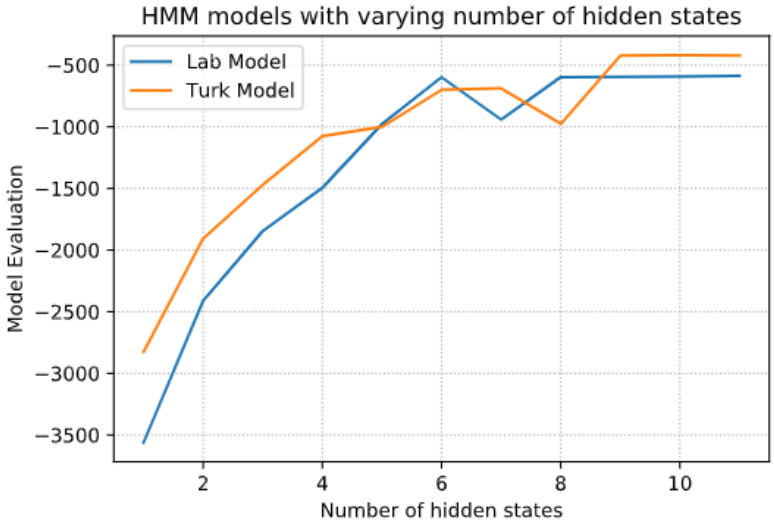


Figure 7: Line plot of evaluation of different HMM architectures.

6 Repeated Experiments

This section will give an overview of the work I have completed during this project. Each sub section offers a unique approach to tackling the problem. The different methods were tried sequentially, and the pipeline had to be amended for each one as the input requirements changed.

Approach of this project was to try many different methods of looking through, classifying, and predicting attentiveness of this data. These experiments enabled me to get a better understanding of the data, and to explore which methods would be most appropriate. While these results did not end up directly contributing to the main goal of this paper, they were important stepping stones of the project.

Before any users were reclassified based on their engagement an alternative aim was established. This was to create a method of identifying whether a user was from the lab study or if they were a crowdsourced turk user based on some input data. While this is not the ambition of this dissertation, it would reveal interesting information about the two user groups. If it was trivial to tell these groups apart then it would suggest that there may have been an error in how the data was recorded, or some other unintentional feature a classifier may be identifying. If any machine learning method was unable to classify a given user then it may suggest that there is little to no difference between the users data. This could additionally indicate that there is no link between a users mouse data and their engagement, which would be a huge and surprising discovery.

6.1 SVM separation method

This section provided the initial exploration into the problem. These experiments attempted to separate the data samples for crowdsourced turker users, and lab participants by plotting samples in space. The aim is that this might reveal information about the features of different users. If there was a clear distinction between the classes then we could potentially reclassify some points based on their euclidean distance to one another.

A Support Vector Machine (SVM) is a common supervised machine learning technique [39]. Here we look at if a simple hyperplane could separate the data when looking at the simple features of the number of mouse events and the total time taken to do the task.

It was found that a linear, or non-linear classification with a kernel trick was unable to separate the data. However as shown in Figure 3, the data does not appear to be linearly separable. Regardless this method was still attempted.

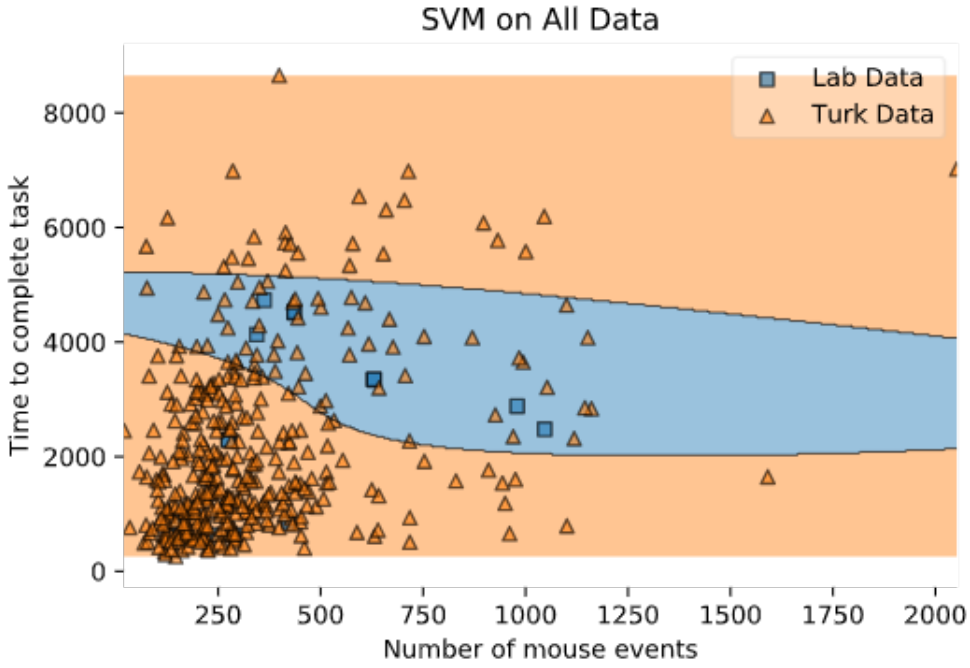


Figure 8: Plot of the SVM, showing the decision region boundary.

Figure 8 shows the unsuccessful attempts of this method, and shows the sheer imbalance of the dataset. Only 53% of the lab data was within the relating decision boundary, where as 86% of the turk data was in its respective region. This highlights how this method will be ineffective on the dataset. While the SVM has created this central band of lab data, this region doesn't accurately model all of the data.

Additionally this method highlights a reoccurring problem of this project , the unbalanced data classes. This can make the accuracy of the model high, even if it misclassified most data points.

6.2 Text Classification

This section of experiments is inspired by text classification algorithms. Text mining is becoming increasingly popular, due to the large increases in available text data from the web and social media. Such data is typically unstructured, and contains very large amounts of information. Such data is also sparse and high dimensional [40]. The high dimensionality of text data stems from the large lexicon of words a document may contain, just the English language has over 170,000 words [41]. Text data is usually sparse because a given document is unlikely to contain anywhere near that many unique words.

The users mouse data used in this project can be represented as an abnormal word document. We can consider a mouse target as a "word", and a "text document" as a single users sequence of mouse targets. This data will have a very limited lexicon of only 6 "words", the sliders from 1 to 5 and other. Therefore this data will have much lower dimensionality than typical text data. Additionally every user has interacted with every target, therefore the data is not sparse but dense.

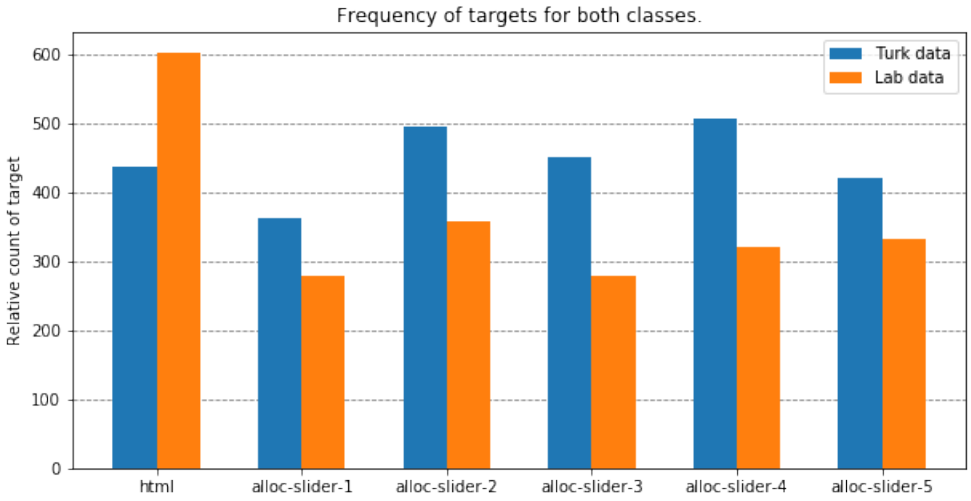


Figure 9: Histogram of targets.

Figure 9 shows the relative count of each of the targets for the user groups. The relative count is calculated by dividing the count of the targets by the number of users, this is so the numbers are comparable otherwise the crowdsourced data would be of a much larger quantity. We can see that lab study users seem to perform more mouse events over the other html elements and not over one

of the sliders. For lab users the most common sliders are 2, 5, then 4. For turk users they are 4, 2, then 3. The causes of these differences are unknown. Importantly they show that there are differences in the data from the classes, something that a classifier should be able to exploit.

One popular text classification method is with a Bag-Of-Words model, called such because the grammar and order of the words is ignored and only the counts of each word is used as a feature [42]. Grammar is meaningless in the context of mouse data, however the sequence of the words may hold crucial information regarding what class a user is in. Disregarding the sequence may be a big limitation of this method as we can imagine a user switching between sliders frequently may be very engaged in the task. Therefore disregarding the order of targets could be detrimental to the model. The Bag-Of-Words representation of the data can be used as input to a Naïve Bayes classifier [42]. A Naïve Bayes classifier predicts a class based on the probability of seeing the counts of words in a document. It can potentially be used here to classify users as lab study or crowdsourced based on their counts of targets. For example it may be found that crowdsourced turk users may have a higher lower counts for the later targets, potentially showing how they got tired of the task and stopped putting effort into it.

$$\begin{pmatrix} 322 & 430 & 464 & 308 & 230 & 1317 \\ 332 & 120 & 112 & 123 & 59 & 164 \\ 41 & 367 & 173 & 565 & 278 & 449 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

This matrix shows the first 3 rows of the bag of words representations of each users data. The matrix has 6 columns representing each of the mouse targets or "words".

This attempt of classification was a failure. The accuracy was 29.6% and F1 score of 22.8%.

6.2.1 N-Grams

The previous Bag-Of-Words method would be considered an unigram model of natural language processing. An n-grams is a series of n words, with unigrams being one word, and bigrams being two word pairs [43].

Different n-grams can be combined together to better understand the complexities of a text document. A mixture of unigrams, bigrams, and trigrams are used extract different levels of text complexity and perform well with document classification [44]. The same technique can be applied to this mouse dataset. Bigrams will be of particular interest as they show the transitions from one target to another. However higher n n-grams could reveal information about a users approach to the tasks. A user who has many n-grams of different sliders would mean they often changed between sliders, potentially looking for the optimum slider values.

TODO: Plot of ngrams for both classes. Create side by side bar plot.

With a combination of 1-grams, 2-grams, and 3-grams the matrix now has a total number of 222 columns.

However due to the aforementioned problems there is still an accuracy of 94.9% and an F1 score of 48.6.

6.3 Imbalanced classes

HMMs can be used for classification as shown above, but they are also known as generative models [34]. This means they can be used to create new data. The above methods of SVM, KNN, and Naive Bayes failed due to the imbalances of classes. Now the trained HMMs could be used to generate new data points, to reduce the imbalances, and make these methods effective.

The HMM can generate a probable sequence of n observed states, however to generate new samples we must decide what length n must be.

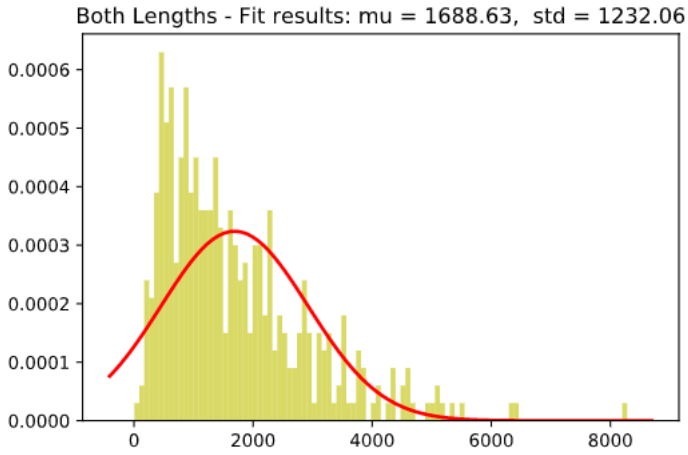


Figure 10: Histogram of length of mouse events for turk and lab data.

Figure 10 shows a the frequency of different mouse events lengths. In red we can see a positively skewed normal distribution which matches the distribution fairly well. When determining n for the new sequence data, we will sample from a gaussian distribution with a mean of 1688 and a standard deviation of 1232. This should generate a realistic distribution of lengths, and when combined with the trained HMM model, a realistic sequence of data.

6.3.1 Revisiting methods

With the generated lab study data we can now revisited some of the methods that failed due to imbalanced classes. With the naive bayes of counts of different mouse targets, there was little improvement, with an accuracy and f1 score of

49% and 46% respectively. However there was a big improvement when the bigrams of events were fed into the model. The accuracy and f1 scores increased to 94%, a dramatic increase. This shows that there is enough difference in the data for a classification algorithm to distinguish the points.

7 Results

Hidden Markov Models were trained on both lab study users mouse data and online crowdsourced mechanical turk mouse data. Then each users mouse data was evaluated by each HMM, and the log Likelihood of that sequence belonging to that model was recorded. This will tell use the likelihood that that sequence of mouse events belongs to either of the classes. If there are any points that appear to be outliers, then we can say that they were doing something differently to the majority of their peers. Outliers with a higher likelihood of belonging to the HMM trained on data from the other group then we can say that that outlier sample appears to belong to the wrong group. Or at least that outlier is doing something different to their peers, which may be a higher level of attention given to the task. These points were labelled Reclassified Lab data and Reclassified Turk Data.

7.1 Table or graph of results

One way of attempting to classify users is to identify any users with a higher lab likelihood as being more similar to a lab user than a turk user. Figure 11 shows this result graphically. Any data point above the line $y = x$ has a higher likelihood of belonging to the lab model than the turk model. Therefore any turk data that is above this line can be reclassified as seeming to belong more with the lab data than the fellow turk datapoints. An interesting and surprising result is shown from the lab data. Not all of the lab data is shown to be belonging more strongly to the lab model than the turk model. 6 out of 14 (43%) of the lab data samples has been reclassified by the algorithm as being more similar to the turk data.

Figure 12 shows which of the turk users are more similar to the lab users , than other users from the same group. The plot shows the same datapoints and axis as in Figure 11 , but the points above the line are recoloured green to show their difference from the actual lab data and normal turk data. 45 of the 461 (9.8%) turk users are above this threshold. Which seems like a reasonable percentage of the population to be paying attention.

The exact distribution of these points change based on the initial random state when training the model.

Figure 3 showed a naive simple attempt to separate the classes of turk users and lab users. It was concluded that a spacial based technique such as a SVM would be unsuitable as the data didn't seem to form any patterns or clusters. Figure 13 shows how the reclassified datapoints are not linearly separable. This proves that my initial hypothesis was correct and that data could not be re-

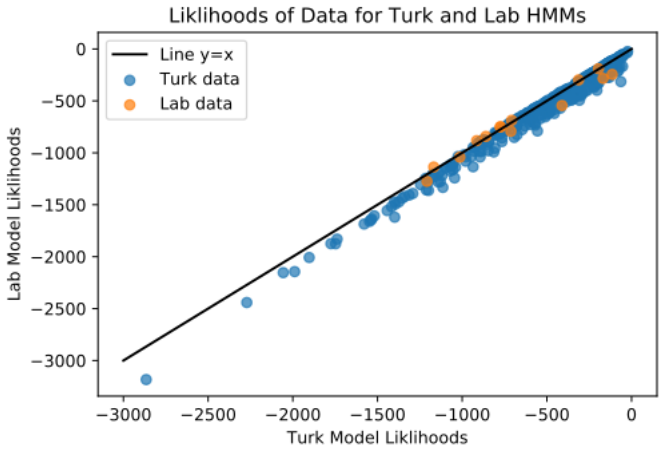


Figure 11: Scatterplot of users likelihood of belonging to either group.

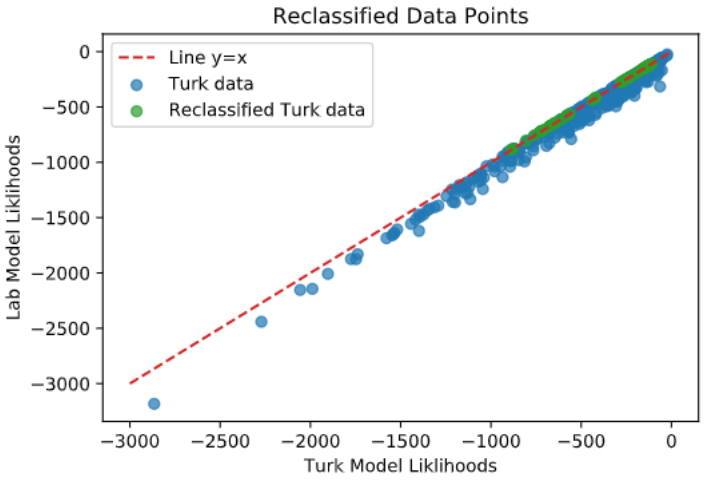


Figure 12: Scatterplot showing the turk users that have been reclassified as behaving more like lab users.

classified purely by looking at simple feature of time taken and just number of mouse events.

7.2 Likelihood length correlation

Examination of the extreme points with the highest and lowest likelihoods revealed a potential correlation likelihoods and length of mouse events sequences. I decided to plot this data to understand if there was actually a link between

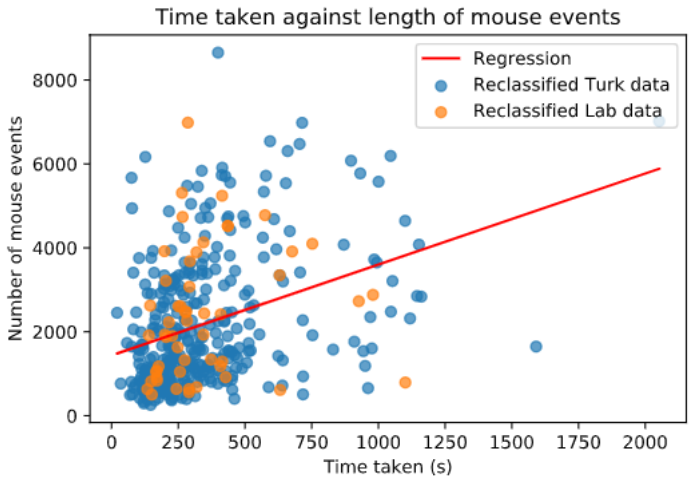
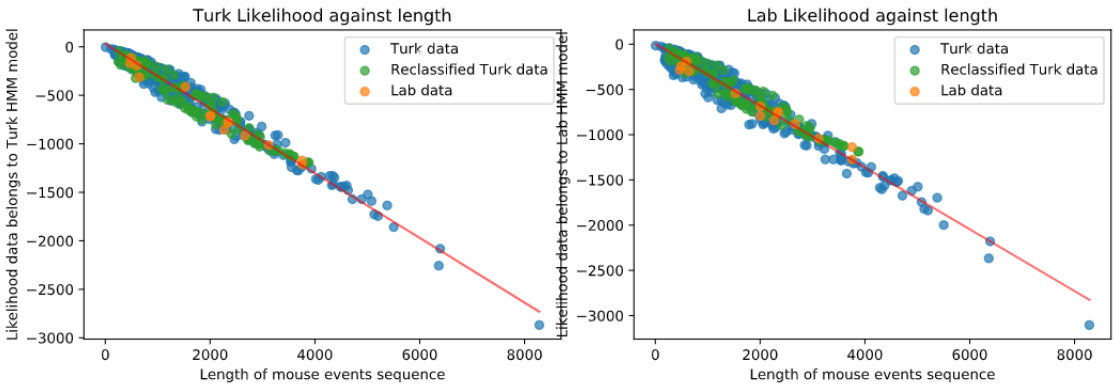


Figure 13: Plot showing features of the reclassified points.

these features.



(a) Turk HMM Model

(b) Lab HMM Model

Figure 14: Plots showing correlation between likelihood of different models and the length of mouse events.

Figure 14 show that the log likelihood from both models and length of a users mouse sequence are highly inversely correlated. The Lab model likelihood and length have a correlation of -0.982, and similarly the Turk model has a correlation of -0.976. A value less than -0.7 would indicate a strong negative correlation, therefore Figure 14 shows an extremely strong negative correlation [11]. Such a strong correlation could indicate the models are doing nothing, other that looking at the length of a sequence. However looking at the reclas-

sified Turk data we can see that there doesn't appear to be any correlation between length, and whether the data has been reclassified. Therefore such a strong correlation between the length and likelihood is of no concern.

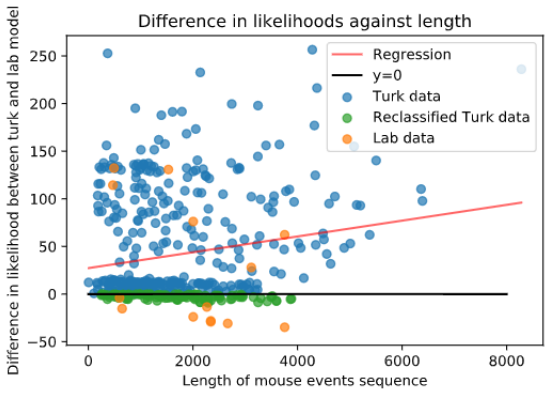


Figure 15: Plot showing difference in likelihood between the models against length of mouse events sequence.

Figure 15 removes the concern that length of mouse events sequence is the only feature being used in the models. The difference in likelihoods and length of mouse events sequence has a correlation of 0.176, meaning the relationship between the features is non-existant.

8 Further Results

In order to confirm any findings it was decided to attempt another experiment using similarly designed HMMs. The key difference is that these models will be trained with different data than the previous models. If models are created with different data, but they identify the same users as potentially belonging to the wrong group then it would be a confirmation that those users are indeed outliers. If the models predict a completely different group of users then it would indicate there is no consistency within the data and that the models may not be reflecting the actual processes.

These confirmation models will be trained on spatial data, of which HMMs are a known classification method. The exact makeup of data will be a sequence of tuples of users mouse cursor location, (x, y).

There were concerns with the mouse location data, mainly being that the data does not seem to be calibrated correctly. Figure 16 shows the issues with the mouse path data.

Ideally the start and end points of all users should be identical, with the majority of events happening with the same y coordinates because of the horizontal sliders. However we can see that there is a massive range in mouse data, some users have a small range of x and y coordinates, others have a range much larger

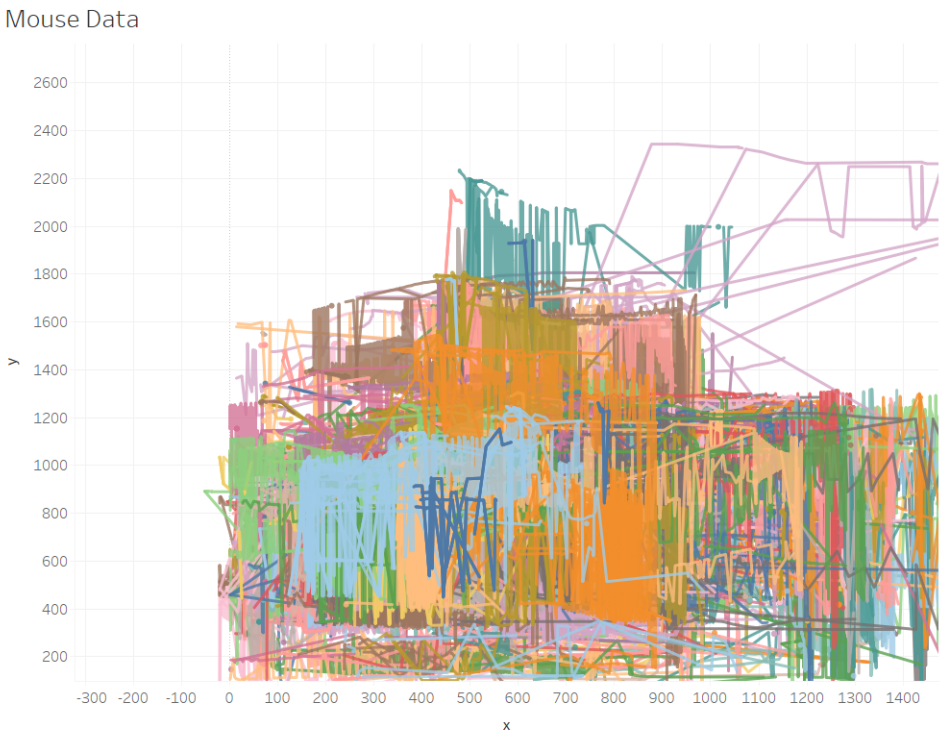


Figure 16: Graph of spatial coordinates of all users mouse path data.

up to a y value of over 1800. This difference is mouse data was not addressed, this may cause some of the users data to be unlikely to belong to any model as they're such a large outlier.

8.1 Results

Figure 17 shows how the new HMMs reclassified 80 of the crowdsourced turk users as belonging more to the lab study. It is worth noting that the likelihoods of each user belonging to the model is much lower when the spatial data is used, compared to when the mouse trget data was used. This is because there is much more variance in the continuous spatial data rather than the categorical target data. The original previous HMM classified 45 of the crowdsourced users. Taking an intersection of the sets of users will show us how consistent the results are. If there is a large intersection of users then the results are consistent with before, however if there is little to no intersection then it would indicate that there is no consistency between models.

$PreviousReclassifiedUsers \cap NewReclassifiedUsers$

There is an intersection of 12 users between the two sets of users. This indi-

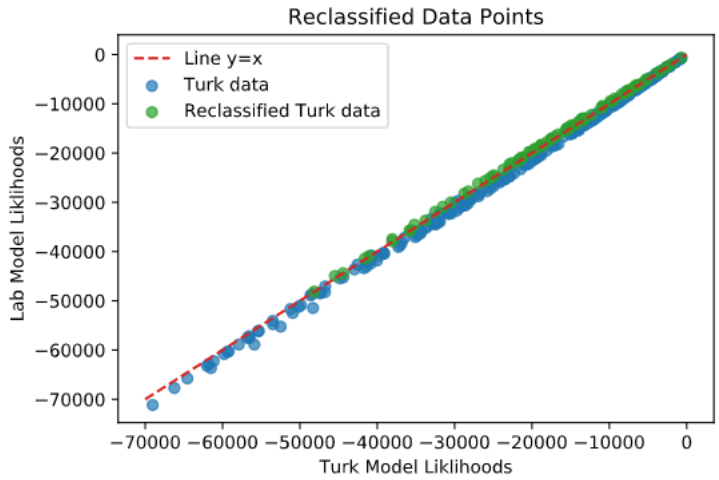


Figure 17: Graph of reclassified datapoints from spatial data HMMs.

cates either that there is some consistency in the results. 27% of the originally reclassified users were identified by the secondary model. This is promising as it shows that a number of the reclassified users are shared between both methods. The inconsistency in the mouse location data identified could be causing the differences in reclassified users between the methods. Therefore I would predict that the results from the spatial models to be less accurate and representative of the real world processes that the models trained with mouse target data.

If we take the union of both sets of reclassified users we get a set of 200 reclassified users. This allows us to have a few different sets of users that we may reclassify as belonging to the wrong group, with different confidence levels. I am most confident in saying that the 12 users reclassified by both versions of the models seem more similar to the lab study users than the other crowdsourced users. I am next most confident in the 120 users reclassified by the HMM trained with mouse target data. This is because data used in those models was more properly cleaned and had less variation in data. The users I am the least most confident in reclassifying is the union of users from both sets of models. This consists of 200 crowdsourced users which is 43% of the total population.

Alternatively inverting this gives us users which we can assume are not engaged in the tasks, to different degrees. I am most confident in saying that 261 of the crowdsourced users not reclassified form the union of the models belong with their class and should not be reclassified. I have next most confidence in the 341 users that were identified with the mouse target data models. Lastly we have the 449 users not identified by the intersection of the models. I am least confident in saying that these users were not engaged in the task.

9 Conclusion

To evaluate the results I would not be too confident in my findings. The data was incorrectly labelled for the task I wanted to perform. Additionally the data was heavily imbalanced. All of the conclusions I have made from the lab data are relying on only a small numver of datapoints which is not statistically significant. Additionally there was not a lot of data from the other class of online data either. Typically data science and machine learning uses big data, where as this project used only 388 records in total which came to only 330 MB of data (MAYBE). Any bias in any of these original samples will be magnified as this was used to classify more points, which would again spread this bias. This would become a self fulfilling prophecy as more similar points would be labelled and spread the belief.

Any concern that the models have learned of a simple feature such as sequence event has been disproven , showing the reclassification algorithm has nonesuch simple relationship.

Biggest flaw of the project is the first assumption, that the lab people were all paying attention. While they were monitored and we can be sure they were not distracted by phones or televisions, many of them may have 'zoned out' and not have been given it their full effort and attention.

10 Future work

All the problems addressed in the conclusion have potential to be addressed in future work. As stated the main flaw of this project is that the label is not labelled for the task I am trying to perform. Therefore to fully evaluate any of these results would require further research with a properly labelled dataset.

It was hypothesised that a users mouse data would not be significantly different depending on which version of the study interface they were using. This hypothesis was based on the findings of the previous study, where the different interfaces had little to no impact on a users success at the investment scenario [9]. Perhaps an investigation into users mouse data might show more of a difference between the behaviours of users using both interfaces.

Another area of future work would be to see if any of the methods developed here could be applied to other similar datasets. A kaggle crowdflower dataset seems like an idea candidate [45]. While it is still not labelled with attention and non-attention, it does contain mouse data of crowdsourced users. Additionally the dataset contains the users results from 3 different tasks, a users success at a task may prove to be a reasonable proxy for engagement.

References

[1] Gabriele Paolacci, Jesse Chandler and Panagiotis G Ipeirotis. "Running experiments on amazon mechanical turk". In: *Judgment and Decision making* 5.5 (2010), pp. 411–419.

[2] Joseph Chee Chang, Saleema Amershi and Ece Kamar. "Revolt: Collaborative crowdsourcing for labeling machine learning datasets". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017, pp. 2334–2346.

[3] Panagiotis G Ipeirotis, Foster Provost and Jing Wang. "Quality management on amazon mechanical turk". In: *Proceedings of the ACM SIGKDD workshop on human computation*. 2010, pp. 64–67.

[4] Jesse Chandler and Danielle Shapiro. "Conducting clinical research using crowdsourced convenience samples". In: *Annual review of clinical psychology* 12 (2016).

[5] Neil Stewart, Jesse Chandler and Gabriele Paolacci. "Crowdsourcing samples in cognitive science". In: *Trends in cognitive sciences* 21.10 (2017), pp. 736–748.

[6] Chien-Ju Ho et al. "Incentivizing high quality crowdwork". In: *Proceedings of the 24th International Conference on World Wide Web*. 2015, pp. 419–429.

[7] Edith Law et al. "Curiosity killed the cat, but makes crowdwork better". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016, pp. 4098–4110.

[8] Eric T Peterson and Joseph Carrabis. "Measuring the immeasurable: Visitor engagement". In: *Web Analytics Demystified* 14 (2008), p. 16.

[9] Gruber Julian. "A comparison of user interactions in lab and crowdsourced studies". University of Vienna, 2017.

[10] Thomas Torsney-Weir et al. "Risk fixers and sweet spotters: A study of the different approaches to using visual sensitivity analysis in an investment scenario". In: (2018).

[11] Diana Mindrila and Phoebe Balentyne. "Scatterplots and correlation". In: *Retrieved from* (2017).

[12] Divesh Lala et al. "Detection of social signals for recognizing engagement in human-robot interaction". In: *arXiv preprint arXiv:1709.10257* (2017).

[13] Andrew T Duchowski. "Eye tracking methodology". In: *Theory and practice* 328.614 (2007), pp. 2–3.

[14] Georg Buscher, Edward Cutrell and Meredith Ringel Morris. "What do you see when you're surfing? Using eye tracking to predict salient regions of web pages". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2009, pp. 21–30.

[15] Tony Renshaw, Richard Stevens and Paul D Denton. "Towards understanding engagement in games: an eye-tracking study". In: *On the Horizon* (2009).

[16] Shahram Eivazi and Roman Bednarik. "Predicting problem-solving behavior and performance levels from visual attention data". In: *Proc. Workshop on Eye Gaze in Intelligent Human Machine Interaction at IUI*. 2011, pp. 9–16.

[17] Daniel Szafir and Bilge Mutlu. "Pay attention! Designing adaptive agents that monitor and improve user engagement". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2012, pp. 11–20.

[18] Daniel C Richardson and Michael J Spivey. "Eye tracking: Characteristics and methods". In: *Encyclopedia of biomaterials and biomedical engineering* 3 (2004), pp. 1028–1042.

[19] Lynne Cooke. "Is the Mouse a" Poor Man's Eye Tracker"?" In: *Annual Conference-Society for Technical Communication*. Vol. 53. 2006, p. 252.

[20] Urška Demšar and Arzu Çöltekin. "Quantifying gaze and mouse interactions on spatial visual interfaces with a new movement analytics methodology". In: *PloS one* 12.8 (2017).

[21] Jeremy Goecks and Jude Shavlik. "Learning users' interests by unobtrusively observing their normal behavior". In: *Proceedings of the 5th international conference on Intelligent user interfaces*. 2000, pp. 129–132.

[22] Chen-Chung Liu and Chen-Wei Chung. "Detecting mouse movement with repeated visit patterns for retrieving noticed knowledge components on web pages". In: *IEICE transactions on information and systems* 90.10 (2007), pp. 1687–1696.

[23] David Hauger, Alexandros Paramythis and Stephan Weibelzahl. "Using browser interaction data to determine page reading behavior". In: *International Conference on User Modeling, Adaptation, and Personalization*. Springer. 2011, pp. 147–158.

[24] Vidhya Navalpakkam and Elizabeth Churchill. "Mouse tracking: measuring and predicting users' experience of web-based content". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2012, pp. 2963–2972.

[25] Jeff Huang, Ryen White and Georg Buscher. "User see, user point: gaze and cursor alignment in web search". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2012, pp. 1341–1350.

[26] Alexander J Quinn et al. "Crowdflow: Integrating machine learning with mechanical turk for speed-cost-quality flexibility". In: *Better performance over iterations* (2010).

[27] Carsten Eickhoff and Arjen P de Vries. "Increasing cheat robustness of crowdsourcing tasks". In: *Information retrieval* 16.2 (2013), pp. 121–137.

[28] Xiaojin Zhu and Andrew B Goldberg. "Introduction to semi-supervised learning". In: *Synthesis lectures on artificial intelligence and machine learning* 3.1 (2009), pp. 1–130.

[29] Xiaojin Jerry Zhu. *Semi-supervised learning literature survey*. Tech. rep. University of Wisconsin-Madison Department of Computer Sciences, 2005.

[30] Omid Mohamad Nezami, Debbie Richards and Len Hamey. "Semi-Supervised Detection of Student Engagement." In: *PACIS*. 2017, p. 157.

[31] Michael Collins. "Tagging with Hidden Markov Models". In: (2016).

[32] José Gordillo and Eduardo Conde. "An HMM for detecting spam mail". In: *Expert systems with applications* 33.3 (2007), pp. 667–682.

[33] Lawrence R Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[34] Jaime Ferrando Huertas. *Generating synthetic data through Hidden Markov Models*. 2018.

[35] Charisma Choudhury et al. *State dependence in lane changing models*. 2007.

[36] Jeremy Ronk. "Structured, semi structured and unstructured data". In: *Retrieved November* 21 (2014), p. 2015.

[37] Hmmlearn developers. *hmmlearn*. `https://github.com/hmmlearn/hmmlearn`. 2020.

[38] Jennifer Pohle et al. "Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement". In: *Journal of Agricultural, Biological and Environmental Statistics* 22.3 (2017), pp. 270–293.

[39] William S Noble. "What is a support vector machine?" In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567.

[40] Charu C Aggarwal and ChengXiang Zhai. "An introduction to text mining". In: *Mining text data*. Springer, 2012, pp. 1–10.

[41] Beth Sagar-Fenton and Lizzy McNeil. *How many words do you need to speak a language?* 2018. URL: `https://www.bbc.co.uk/news/world-44569277` (visited on 12/09/2020).

[42] Dan Jurafsky and C Manning. "Text Classification and Naive Bayes". In: *Vorlesungsskript http://www.stanford.edu/class/cs124/lec/naivebayes.pdf* (2015).

[43] James Martin Daniel Jurafsky. *Speech and Language Processing*. 2009.

[44] Charu C Aggarwal and ChengXiang Zhai. "A survey of text classification algorithms". In: *Mining text data*. Springer, 2012, pp. 163–222.

[45] Human Computation. *Workers Browser Activity in CrowdFlower Tasks*. 2017. URL: `https://www.kaggle.com/humancomp/worker-activity-crowdflower` (visited on 03/05/2020).