Active Learning To Reduce Human Labeling For Automatic Psychological Text Classification

Author: Jahson O'Dwyer Wha Binda[1] , **Supervisors: Willem-Paul Brinkman**[1] , **Merijn Bruijnes**[1]
[1]TU Delft

Abstract

In recent years there has been an increase in the number of patients for issues relating to mental illness. To this effect to help with this increase, schema mode assessment through a conversation agent is being used to conduct schema therapy, a form of psychological treatment. To train such an agent, training data labeled by humans is necessary but can be very expensive to conduct. The question being researched is through the use of Active Machine learning is it possible to reduce the amount of required labeled data to do such classification. Three experiments on the use of active learning with currently available classifiers were performed where the active learner attempted to train the classifiers to an accuracy within +/- 3% of the same classifier trained with traditional machine learning on the full data set. The experimental results found that in all cases the use of active learning drastically decreased the number of necessary labeled data the classifier needed to achieve a similar accuracy. Consistently reducing the number by 98% and above answering the initial question. Though possible limitations of the data set and classifiers for such texts may be positively influencing the magnitude of the reduction.

1 Introduction

Mental illness has been a growing issue in recent times and accounts for one-third of the world's disability caused by adult health problems.[1] With this growing social issue new ways to detect and help possibly vulnerable people are being developed in the hopes of helping them with their issues.

Due to this David Allaart conducted research on schema mode assessment through a conversational agent, where he states: *"Schema therapy is psychotherapy for the treatment of personality disorders and other psychological disorders. An important element of schema therapy is determining a patient's schema modes, a concept central to deciding the treatment approach"*[1].[2] Where a schema can be said to be an unhealthy pattern of thoughts and behaviours often brought about through childhood trauma.

Allaart found that an agent can be used to predict a person's schema but then the problem of training the agent still remains. The agent uses text-based stories from multiple users to train and to try identify a schema. The number of stories is ever-increasing and labelling of this training data becomes more and more expensive and time-consuming.

1.1 Active Machine Learning

Active machine learning is a case in machine learning that allows for the agent to query the user (also known as wizard) to label data it finds necessary for its training. The algorithm works by initially training a classifier on a small sample of labeled data, and then train the classifier one data point at a time. Where the active learner selects a data point from a training pool it believes would have the greatest influence in improving the classifier predictions based on the chosen query strategy, to then query the wizard for its label.

This is where this research paper tries to evaluate a possible solution to the labeling problem mentioned before; if with the help of active machine learning can it aid in the labelling of training data for an psychological text agent.

This can possibly decrease the amount of necessary labelled training data. As the algorithm can query the user specifically for data points the model may not be certain about for their label and then update it's model. Which could potentially lead to the model needing a smaller set of labeled data to achieve similar performance to a traditional machine learning model, as only data the active learner deems necessary is used for training.

There are multiple strategies as to how the active learner decides to query the user to label a piece of data but it generally involves picking data points with high influence or high uncertainty for the classifier the active learner is training.

2 Problem Description

This section will be an overview of the problem of answering our research question.

2.1 How well can active machine learning be used to support human labelling of a data set?

This question is the main topic of this research, the goal in the end is to able to decide if active machine learning is a viable

[1]Allaart, Schema mode assessment through a conversational agent, 1

tool to support humans in labelling of a data set, by reducing the amount of required labelled data to train a classifier. With specific attention to the schema data Sets from Allaart's and Burger's research.

To tackle this question I have created 3 sub-questions to individually look at. Using the findings and results from exploring these sub-questions it gives us an entry point to answering the main question which can be used to draw our final conclusions.

2.2 Sub-Questions

1. What is the general performance/findings of active machine learning on schema data set and others?
Performance of active learning will be based on the number of labels needed to acquire similar accuracy on the same test set when compared to traditional machine learning. To do this we will be using three data sets:

1. David Allaart's Schema data set [2]
2. Franziska Burger's Schema data set [3]
3. Sentiment140's Tweet data set [4]

For example, if 1000 labeled training data using a Support Vector Machine (SVM) classifier achieves 80% accuracy on a test set and an active learner using the same classifier can achieve an accuracy of +/-2 of 80% while only using 30 labeled training samples.
In this case it can be said to be a positive performance for the active learner as significantly less labeled data was used to get a similar accuracy.

2. Which active machine learning strategy/s is most appropriate to use?
Active machine learning can have multiple strategies for picking which unlabeled item would be most appropriate to query the user for its label. Due to the multi-labeled nature of both schema data sets there are a few strategies available to use, due to time constraints we will not be experimenting with all or many of them. In some cases, the strategy to be used depends on the type of classifier the active learner is using and in some cases testing multiple strategies to see which performs best may be necessary.

3. Is the difference between the performance/accuracy and number of labeled data acceptable to be used over traditional machine learning?
This question explores after performing and testing active machine learning, on data like Allaart's data set and others is the reduction of labeled data worth possible trade offs?
For example is the possible reduction in accuracy worth it if significantly less labeled data is used during training? Or if to get to an acceptably accuracy on our model will it be the case that we will be using a similar sized set of labeled data defeating the purpose of using active learning?

Answering this section will allow us to make our final conclusions in the research paper.

2.3 Reasons For Multiple Data Sets

Classifier for the active learner
The reason for using multiple data sets in our research is due to the nature of active machine learning, at it is core to evaluate the performance of active learning a proven and acceptable classifier must be used by the learner, especially for querying and finding uncertainties the classifier may have for items in the data set.
Allaart's Schema data set has yet to have much research in a well performing classifier for user stories and possible schema. This is research that is currently being done by Han[5], Zhang[6] and Park[7] to which for the remainder of this paper I will refer to as my peers. Because of this early versions of such classifiers are being used and reviewed. Which may or may not be optimal, to show the magnitude of the decrease in training data.

Due to the possible limitation of a well preforming classifier for our active learner, we are including the Sentiment140's data set as they have research supporting a classifier for the twitter data set.[8] Allowing us to draw conclusions for our main research question on Active Machine Learning while not being limited by the problem of how best to build a model to classify the data. As a poorly optimized classifier may negatively impact our findings.

Origin of schema labels
Another possible limitation with the current schema therapy data is the labels the current classifiers are using. Allaart's data labels come from a Schema Mode Inventory (SMI) questionnaire, which labels schema for a person during a large time frame. This would mean that a collection of stories from a single person may fall under a single schema even though when a story is looked at individually it may not be indicative of that particular schema.

Burger's Data was labeled manually so there is still a possible flaw to how accurate the labels to each individual story being classified. Sentiment140's data is labeled on sentiment of the tweet, where a tweet with emoticons correlating to positive or negative sentiment are used for labeling. In general this form of labeling was found to be more indicative of the actual sentiment of the individual tweet, allowing us to compare the results on the labeling of the natural language of an individual story or tweet.

Sentiment140's tweet data set was chosen as tweets have a similar structure to stories like the ones in Allaart's and Burger's data which both involve classification of natural language and has a classifier to predict sentiment in tweets where sentiment is similar to schema. As well as having a large pool of data to possibly use (over 1.6 million labeled tweets).

This similarity between the sentiment of tweets and schema of user stories can help us with our final conclusions and comparisons with schema therapy related data. For example if it is found that on Sentiment140's Data and classifier the benefits of Active learning are substantially more than on Allaarts data. This may indicate that with the improvements to classifiers on Allaart's data currently being researched it may lead to similar or improved findings more inline with Sentiment140's in the future.

3 Applying Active Learning

To perform our research we will be applying active learning on different classifiers for each data set. The code used

to achieve our results will also be made available as Jupyter Notebooks.[9] (See Appendix B)

The programming language being used is Python.[10] Using multiple different libraries and their dependencies to help us create our learning agent. The main libraries used in creating our active learner include; Scikit-learn[11], to form the basis of our classification algorithms and to allow for easy swapping between different classifiers, as well as providing various helper functions for splitting of data and analysis of results. As well as ModAL[12], a modular and flexible active learning framework that works well with the classifiers of Scikit-learn.

3.1 Experimental work

Before beginning work on the data set it must be ensured that it is cleaned and ready to be processed. In the case of Allaart's data set this was initially done in tandem with my peers. The "stories", or messages that should be classified to specific schema required some changes to make it easier for the classifier to build relations between the words in a story and schema during training.

Pre-Processing of Allaart's data set

Allaart's data set is a comma-separated file with multiple columns, the ones useful to our research is the "Text" column which we plan to use as our training data as user stories, and the SMI questionnaire columns which are a group of columns usually between 7-10 columns with answers relating to the possible schema for the specific user story. Each schema has a range of columns dedicated to it. We will use these group of columns to create a final binary column to indicate if the user story is true for each schema, which will serve as our labels.

Together with my peers to clean the data set to allow it to be processed, it was decided the following rules will be applied to all "stories" in the data set.

- Lower-Case the story.
- Replace miss-spelling, contractions and numbers.
- Add missing sentence end marks, comma and spaces.
- Remove stop words and unnecessary white space.
  - Stop words are stories or messages directed to the chat bot collecting them which do not provide useful data to be applied to schema-therapy. These are words like: Ok, Quit, GoodBye, yes, no
- Finally adding a binary label column for each schema based on the questions that were answered in the questionnaire columns of the data. Using the same labeling rules as Allaart's did in his paper. [2]
  - If the average of questionnaire answers are higher than 3.5 then the story has that schema.
  - If any of the answers are 5 or more, the story has that schema.
- Final pre-processing methods inspired from Burger's paper[3] to go with her classifier. Especially her Tokenization method.

Pre-Processing of Sentiment140's data set

Pre-processing of Sentiment140's data will be taken from Ishan Kotian, a student from Ramrao Adik Institute of Technology Nerul, Bayes classifier which achieved an 85% accuracy on the data set.[13]

Some of the steps included in Kotian's pre-processing:

- Remove URL, Usernames, stop words and punctuation.
- Expand abbreviations to non abbreviated form.
- Tokenization
- Creating a Count Vector

Pre-Processing of Burger's data set

Burger's data had the majority of the pre-processing already complete. Due to time constraints of the research our active learner only works with binary labels while Burger's data set has labeled a story to a schema on a scale from 0 to 3 (inclusive).

To easily and quickly add her classifier to our active learner a reduction of possible labels will serve as the pre-processing.

- Stories labeled 0 and 1 are now a single label 0 for a low correlation to the specific schema.
- Stories labeled 2 and 3 are now a single label 1 for a high correlation to the specific schema.

Building The Active Learner

Once the data is ready for processing a classifier for each data set must be built, due to time constraints and for comparison a variant of a SVM type classifier will be used across all data sets. The classifiers used by the active learner will also be based on a previously made or researched classifier for the respective data set.

Every classifier used by the agent will need modification to some degree. This is to allow for the active learner to query the classifier on which data would benefit most from labeling, as well as allow for the wizard to add a new piece of training data to the classifier. The classifiers that will serve as the basis for the custom ones used by the active learner will be:

- **Allarrt's data**: Classifier based on the research currently being conducted by Park on the use of a SVM model for schema based classification will be used.[7]
- **Burger's data**: The SVM classifier used in her research paper; Natural language processing for cognitive therapy will be used.[3]
- **Sentiment140's data**: A SVM classifier similar to the one made by Park[7] and Kotian[13] will be created.

Each classifier will be used by the active learner agent built using the ModAL[12] framework, the steps involved in building our core active learning loop is as follows:

- Split available data to training and testing Sets.
  - 80% Training Pool & 20% Testing Pool
- Train and test initial classifier using traditional machine learning on the full training pool.
  - Fit and predict labels with all available training and testing data

  - Record accuracy
- Train a new instance of our initial classifier on a small subset of labeled data. And remove items from the training pool.
  - This is a requirement of the ModAL library, an initial set with an example of every class to be classified must be used so the query strategy is functional in the learning loop.
- Record our initial accuracy on the test set.
- Begin our active learning loop
  - Use the active learning query strategy to select which data item to ask the wizard to label.
  - Train the classifier with the chosen data item.
  - Remove data item from training pool.
  - Use classifier to predict on the test set and record accuracy.
  - End loop if active learner accuracy is greater or equal to traditional machine learning accuracy, or when all available training data is exhausted, or after a set number of iterations.
- Plot and analyse results.

4 Selecting Learning Strategy

There are multiple query strategies our active learner can use due to the multi-label nature of schema and user stories a strategy that works well in such a case would be beneficial.

For this, we will lean on Esuli and Sebastiani research on Active Learning Strategies for Multi-Label Text Classification.[14] Where they evaluated multiple active learning strategies and concluded in their results that a strategy involving minimum or maximum confidence in the specific training data should be the strategy of choice.

Based on their findings, the query strategy of choice among all our schema active learning experiments will be Minimum Confidence.

Minimum Confidence Strategy

In this strategy, the learner selects the data point in which it has the least confidence in its most likely predicted label. How this will be used by our learner is such that:

- The classifier is used to predict probability estimates for all items in the available training pool.
- The data point with the lowest confidence is selected.
- In case of a tie, the indices in question are shuffled and chosen at random.

5 Experimental Setup and Results

This section will include the experiment setup along with a brief analysis of each experiment result. A more detailed reflection will be discussed in Section 6.

5.1 Experimental Setup

1. Allaart's Data Set

For Allaart's data Park's Linear Support Vector Classifier (SVC) for schema will be used. [7]

Setup Steps

- A Binary SVC similar to Parks was created and trained on the full data set.
- A Binary SVC similar to Parks was created and trained on 8 items of labeled data.
- An Active Learner using the Minimum Confidence query strategy to find the index in the available training pool to query the wizard was created.
- An Active learning loop running 590 iterations with a target accuracy of +/- 3% of traditional accuracy.

2. Burger's Data Set

In Burger's[3] research paper she tested multiple different classifiers, for this paper her Linear SVC will be used which similar to the one experimented on with Allaart's data set.

Modification's to her original classifier to better suit the active learner will be necessary as she created a separate classifier for each schema in her data set for a total of 9 different classifiers while our active learner uses a single classifier for its predictions.

Setup Steps

- Convert multiple SVC's to a single Multi-Output classifier.
  - Now a single classifier for all schema can be used by the active learner.
- Create an instance of Multi-Output classifier and train on full data set.
- Create an instance of Multi-Output classifier and train on 18 items of labeled data.
- An Active Learner using the Minimum Confidence query strategy to find the index in the available training pool to query the wizard was created.
- An Active learning loop running 30 iterations, with a target accuracy within 3% of traditional accuracy.

3. Sentiment140's Data Set

For this experiment, a Linear SVC classifier similar to the one used in the previous experiments was created. This data set has not been used for research into Schema Therapy like the others but does share similarities, while Allaart's and Burger's data sets label user stories to specific schemas, Sentiment140's data set relates tweets to their sentiment, either a positive or negative tweet. Tweets can be seen to be similar to user stories and a positive and negative sentiment can be seen to be similar to schema.

Another important difference to keep in mind during this experiment is the difference in the data sets in terms of size and variance of labels. Sentiment140's data set is much larger than the other two with a total of 1.6 million labeled tweets available compared to the thousands in the others. Sentiment140's data set also has somewhat balanced examples of tweets in regards to their labels, Sentiment140's data set have an almost equal 50% examples of tweets labeled for positive and negative sentiment compared to the over 90% positive labeled examples for a particular schema in the other data sets.

Setup Steps

- Create an instance of Linear SVC and train on full training data set.
- Create an instance of Linear SVC and train on 10 items of labeled data.
- Create an Active Learner using the Minimum Confidence query strategy to find the index in the available training pool to query the wizard.
- An Active learning loop running until all available training data has been used or until an accuracy similar to the traditional machine learner is achieved. (Within a 3% margin)

5.2 Experimental Results

1. Allaart's Data Set

- Traditional Machine learning after training on the full data set (required 1100 labelled items). Achieved an accuracy of 37% for predicting all schema correctly for a given user story.
- The classifier trained on the initial set of 8 labeled items achieved an accuracy of 35%
- After 0 iterations of the active learning loop an accuracy of 35% is achieved, which was the target accuracy.
- Loop was continued until an accuracy of 36% was achieved at iteration 583.

Result Analysis And Graphing

Figure 1 is a graph of the accuracy of predicting the schema of user stories against query iteration. Where the blue line represents the accuracy of predicting all schema for a user story successfully and the others represent the accuracy of predicting the individual schema in the full test set.

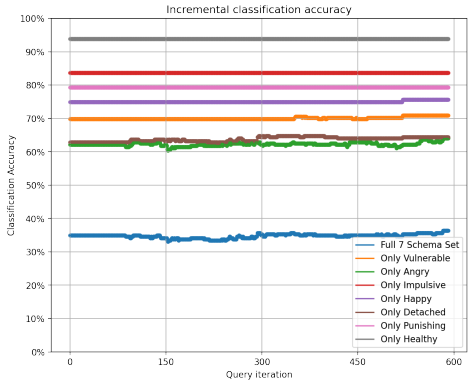

Figure 1: Accuracy against Active Learner Query Iteration For Allaart's Data Set.

As you can see from Figure 1 the target accuracy is achieved without a query from the wizard. This would mean that with only 8 labelled data we achieved an acceptable accuracy to traditional machine learning using 1100 labelled data.

This would mean that a reduction of over 99% was achieved, which is very impressive. Though these results seem promising, it may also point to a possible flaw in the current classifier for schema and the pre-processing of the stories. As well as a possible in-balance in Allaart's data set.

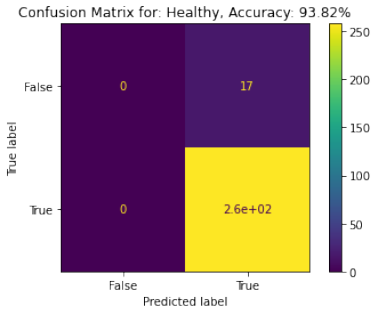

Figure 2: Confusion Matrix for Healthy Schema

A confusion matrix was generated for the Healthy schema and is shown in Figure 2.

A similar matrix is seen for most schema and is shared between the Traditional Learner and Active Learner. As you can see the agent in this case is always predicting *True* for Healthy to acquire an accuracy of 94%. This is the is the behaviour the Traditional Learner currently is doing and the active learner can mimic this with much less labeled data as seen in Figure 1.

A possible in-balance in Allaart's data set may lead to a classifier that behaves like this, as in all the available data there is an overabundance of single class examples for a single schema. For example over 90% of the stories in the data set are labeled as "Healthy" which can lead to a classifier that is trained on this data to always predicts healthy to achieve an accuracy of over 90% for that schema, as seen in Figure 1.

An exception to this are the Angry and Detached schema which have more balanced examples in the data set, since we continued running the learner even after achieving the target accuracy. Figure 1 it can be seen that more balanced schema a require grater number of labeled data to improve accuracy and avoid a confusion matrix that only predicts a single label.

2. Burger's Data Set

- Traditional Machine learning after training on the full data set (4151 labelled items). Achieved an accuracy of 21% for predicting all schema correctly for a given user story.
- The Active Learner trained on the initial set of 18 labeled items achieved an accuracy of 21%
- After 30 iterations of the active learning loop an accuracy of 21% was also achieved

Result Analysis And Graphing

From the results, we achieved the same accuracy on the test set when we trained on 4151 labeled data to the initial 18 labeled data set the active learner used. At this point, the active learning loop had no benefits as it already began with a classifier with similar accuracy, as seen in Figure 3.

This would indicate that with 18 labeled data we were able to reduce the amount of labeled data required by more than 99%. But similarly to the experiments run on Allaart's data

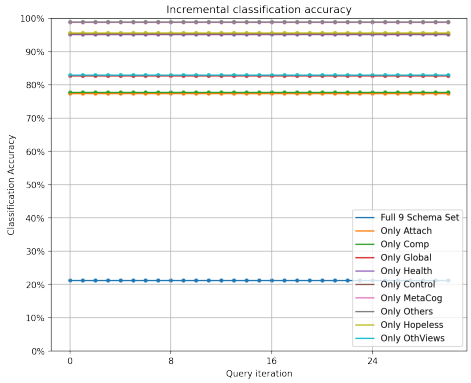

Figure 3: Accuracy against Active Learner Query Iteration For Burger's Data Set

set this reduction may be too good to be true for realistic psychological text classification.

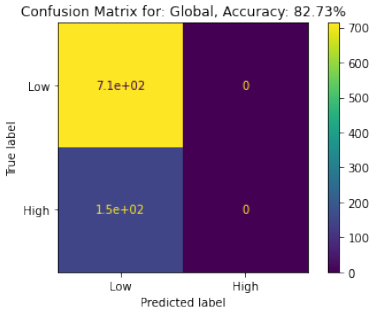

Figure 4: Confusion Matrix For Active Learner On Burger's Data

Taking a look at the confusion matrix of the Active learner for the Global schema, Figure 4, we see a similar behaviour to the one on Allaart's data. The agent predicts a single label for every story. We see the same confusion matrix for the traditional machine learner too.

As seen in Figure 4, there is not a single instance of the agent predicting a high correlation to the schema. This means that training an agent that predicts low at every instance requires minimal data, in theory for this particular schema a single low instance of training data would be enough to achieve an 82.75% accuracy in this test set. This behaviour may explain why our Active Learner had reached its accuracy goal even before any query to the wizard was made.

Figure 1 shows that there is a huge in-balance in the data set greater than with Allaart which may also indicate that the active learner can not improve a classifier that is already over fitted to an in-balanced data set.

As stated in the pre-processing for this experiment in Section 3 and the setup in this section, modifications to the classifier and labels of the data were made to better suit the active learning agent being used in this research paper. To verify that this behaviour is not only the case with our modified version of Burger's classifier an experiment on traditional machine learning was run with Burger's original classifier[3]. The difference from her research being, that accuracy is measured by correctly predicting all the labels for a given user story.

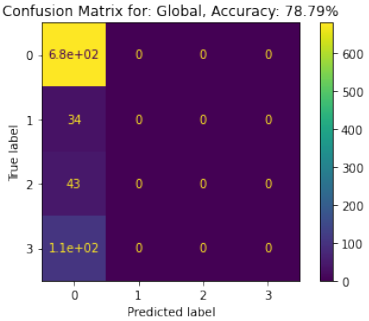

Figure 5: Confusion Matrix For Active Learner Using Burger's Classifier

Burger's classifier predicts a 0,1,2,3 level for each schema after training on the full train set, the accuracy for correctly predicting the level of all schema for a particular user story achieved was 12%. The confusion matrix as shown in Figure 5 looks to be similar to the one produced by the modified binary classifier, indicating that the behaviour of the modified classifier is shared by the original. The difference in accuracy percentage between can be argued to be because in the modified binary classifier the labels 0,1 and 2,3 have been merged to each be a single label.

So as with the experiment with Allaart's data, there has been a huge decrease in the required labeled data needed to achieve a similar performing classifier but the practical usefulness of such an agent is up for discussion.

3. Sentiment 140 Results

- Traditional Machine learning after training on the full training data set (1.28 Million labeled data). Achieved an accuracy of 76% for predicting the sentiment of a particular tweet.
- The Active Learner trained on the initial set of 10 labeled items achieved an accuracy of 51%
- After 20120 iterations of the active learning loop an accuracy of 73% was achieved

Result Analysis And Graphing

After running the experiment we can see we were able to train a classifier that performs similarly to the one using traditional machine learning while using 98.7% less labeled data. This like the previous experiments is a huge reduction in labeled data the difference here is that the active learner did not create an equivalent classifier almost instantly.

As seen in Figure 6, over time as the active learner queries the wizard for more labels there is a clear increase in accuracy on our test set. This curve appears to trend similar to a logarithmic curve.

While performance gain per query is small and seems to decrease as accuracy increases the reduction of the amount of labeled data required to train this classifier can be significantly increased with lower target accuracies as seen in Table 1. Table 1 shows even greater reductions in the amount of

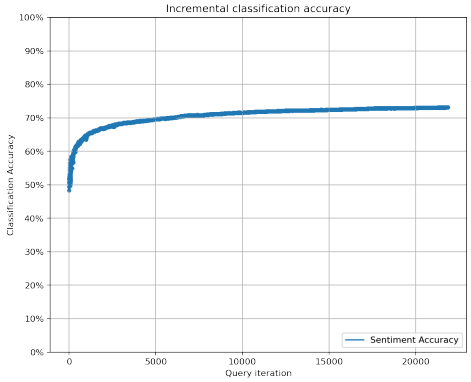

Figure 6: Accuracy against Active Learner Query Iteration For Sentiment140's Data Set

Accuracy	Query Iteration
55%	39
60%	265
65%	1034
70%	5859

Table 1: Query Iteration When Certain Accuracy Achieved

necessary labeled data can be achieved if a greater loss in accuracy is acceptable.

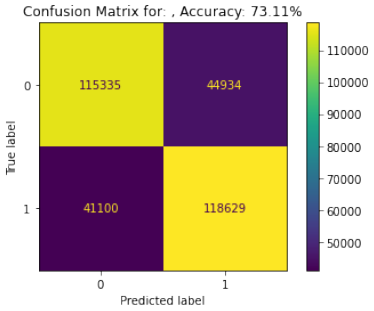

Figure 7: Confusion Matrix For Active Learner On The Sentiment140 Data Set

When looking at the confusion matrix for the active learner we may have some insight as to why we see such a large difference in the results of this experiment when compared to the other two.

Figure 7 indicates a behaviour not shared by the other classifiers, it is not predicting the same label at almost every instance.

5.3 Experimental Limitations

The experiments included in this paper do have limitations especially regarding the two involving schema therapy. As mentioned before the research on schema faced limitations regarding the in-balance of positive labels for many schema. Making the model from the classifier the active learner was trying to train to equivalence very simple as seen from the confusion matrices in the experimental results.

The schema labels for user stories as mentioned before in Section 2.2's "Reason for Multiple Data Sets" may also be limiting the research. For example with Allaarts data each story is labelled based on a SMI questionnaire, then it can be argued that this is not really indicative of an individual stories schema, as the schema for a period is given as the label to a single user story in isolation. Resulting in a model attempting to be made for data with a possibly low relation between story and label.

The binary labeling of schema used by the learner may limit our findings. Classification of individual stories may not be as simple as yes or no for a specific schema. Stories may have a range for particular schema, the use of binary labels can result in the relation between user stories which could be said to be very healthy and somewhat healthy to not be made for example. Which could be reasoned as to why Burger's research classified schema in a 0-3 range for a particular schema, though as seen in Figure 5 and Figure 4, the binary labeling in this case is still indicative of the range Burger used when it comes to the final classification.

6 Discussion

To achieve our results the first and second of our initial subquestions have been answered. Leaving the final third subquestion to be tackled before making our final conclusions.

Is the difference between the performance/accuracy and number of labeled data acceptable to be used over traditional Machine Learning?

Judging by our results we can say yes for the three tested classifiers. As in all cases of our experiments, over a 98% reduction in the amount of necessary labeled data was achieved to create a classifier with a similar accuracy. Especially in the case of using active learning on the data sets involved in schema-therapy as the active learner was able to achieve the target accuracy instantly.

How well can active machine learning be used to support human labelling of a data set?

Solely based on the results from the previous experiments, we can say that the use of active machine learning can greatly support human labelling of a data set. As an active learner has shown that it has the potential to drastically decrease the amount of necessary labelled data to train a model to an equivalent accuracy when compared to traditional machine learning.

Though in the case of schema-therapy and labeling of user stories, with the current classifiers and data sets available the real world usefulness of the classifiers trained by active learning is up for discussion. The same could possibly be said on a classifier trained with traditional learning.

As repeatedly seen in Section 5 of this paper, the classifiers for schema follow simple classification rules. From the confusion matrices, we see that the active learner for most schema only needs enough data to train on to make a single label prediction at every instance, as seen in Figure 5, Burger's classifier never predicted anything other than a 0 for the Global schema in the test set. This means that an active

learner potentially will not need any more than a single training example labeled 0 to achieve an equivalent classifier for this schema.

This difference in classifiers is most seen when comparing Figure 4 for schema therapy and Figure 7 for sentiment classification. Figure 7 shows that though the accuracy is lower in correctly guessing the sentiment, the matrix shows that the model used to create such a matrix has a more involved decision process when compared to Figure 4. This can be argued to be the reason why out of all the experiments the third sentiment classification experiment required the most labeled data to achieve an accuracy below the traditional machine learning accuracy.

What potentially caused Sentimenent140's results to differ from schema-therapy results

Since Sentiment140 only had to classify a single label sentiment, it should be only compared to a singular schema at a time and not correctly labeling all schema for a story.

In this regard, even though the pre-processing and classifier built for both data sets were somewhat similar the results achieved were very different. The reason for this could be because of two major differences in the set of Sentiment labels and the set of labels for individual schema.

- Sentiment140's data set is significantly larger.
- Sentiment140's data set has a near equal balance of positive and negative examples.

The second point can be argued to be the greatest influence, the schema data set seems to be very unbalanced in the case of some schema like "Healthy" the data set has over 90% positive examples which can lead to a machine learner training on such data to suffer from overfitting to these positive examples. Though it achieves a high accuracy this accuracy is not realistic if it were put into real world practice. The classifier trained on the sentiment database has a more varied training pool meaning that it not only learns from positive examples but from negative ones as well which lead to a classifier that may perform better outside the experiment environment. This idea is also supported in the experiments ran on Allaarts data as the more balanced schema required larger sample sizes to achieve greater accuracy which is in line with what is seen with Sentiment140's data.

There may be ways to improve the performance on a classifier trained on the data related to schema-therapy but for the purposes of this research paper it is out of scope and is research currently being done by my peers.

Attempting to be selective with items used in the data set to try balance it will significantly reduce the size of our training pool. Which in itself brings potential problems in the usefulness of a classifier trained on it.

7 Conclusions and Future Work

7.1 Conclusion

To conclude based on the findings of this paper we can say that Active Learning can reduce the amount of human labeling for automatic psychological text classification. Through the use of active learning a similar performing classifier was built with significantly less labeled data, reducing the amount of data a human will be required to label to train such a classifier.

The experiments from Sentiment140 data set has shown that current psychological text classification may be limited, due to the size and balance of the data set. As well as shedding light into possible flaws in the labeling of the data, while also giving an idea of what active learning can achieve with further improvements to psychological text classification.

7.2 Future Work

Improving Psychological Text Classification

In the future performing these experiments again on a schema-classifier with a more complex and accurate model may be useful to achieve results that may be more indicative to how it may work in the real world. This may involve looking into the possible flaws in the currently available data being used for this type of research and how it may be negatively impacting psychological text classification. Park's[7] classifier towards the end of her research paper has been able to achieve a 40% accuracy on Allaarts data which is greater than the one achieved by the model based on Burger's in this paper, possibly indicated a more complex model for the active learner to learn towards.

Relation between Accuracy and Required Training Data For Binary Natural Language Classification

When looking at Figure 6, which is the graph created by our active learner. An observation can be made, this graph looks to trend in a logarithmic manner. As can be visually seen when comparing Figure 8[2] (Appendix A) and Figure 6.

This could potentially point to a logarithmic relation between accuracy and the amount of necessary labeled data to train a classifier in the binary labeling natural language.

This logarithmic relation if seen to be consistent across various experimentation and not this single case could prove to be one of the most useful finding in this research paper.

In regards to schema therapy this could point to not only showing large reductions in the amount of human labeling needed to train a classifier that labels schema. It may also allow for estimates for how much labeled data is necessary, the possible accuracy reduction in regards to amount of training data and the point in which the accuracy gain of adding more labeled data is no longer useful, before the classifier is even created or any labeling is done.

This could also allow for modeling the use of Active Learning to classify schema by allowing a target accuracy for a classifier to be made without traditional machine learning on a full or large data set being performed first like in the experiments.

If the mathematical trend is known this could potentially lead to avoiding the situation of using humans to label more data than necessary to train the active learner as well as labeling less than what is required. Which is the main goal of the use of active learning for schema therapy and the driving force behind this research paper.

[2]Source: http://www.endmemo.com/r/log.php

8 Responsible Research

8.1 Scientific Integrity

Data used in this research is ethical and justified for the research. Data used is either open source and publicly available or if not the case precautions on limiting the access of the data has been made.

In the case of the non-public data set from Allaart's and Burger's research I indirectly worked with four other students namely Budi Han[5], Marijam Zhang[6], Jimmy Lam[15] and Jeongwoo Park[7]. We worked together in pre-processing of the data to be used to create Schema classifiers based on the data. Allaart and Burger have received ethical approval from the TU Delft's Human Research Ethics Committee using this data.

This researched refers to multiple literature and findings as well as code created by a third party, in the efforts to avoid plagiarism all literature and code inspirations have been properly cited and used in the context of the current research being conducted. Libraries and tools used to conduct the experiments have also been detailed and cited.

Furthermore to reduce bias in the data set all samples for the experiments were randomly chosen. The algorithms for randomness using a seed to allow for it to be reproduced reliably.

8.2 Reproducibility

This research is reproducible as details on the methods and tooling used to acquire results are detailed in the paper (See Section 5) as well as the reasoning behind using such tools. The data will be available in private TU Delft servers while the supporting code for the research paper will be made publicly available on an official website (See Appendix B). To allow for the experiments to be verified and reproduced all randomness in the code can be made consistent with the use of seeds which can be seen in the available code files.

References

[1] J.-L. E, A. P, S.-B. S, W. K, W. K, M. D, C. C, and L. P, *"Reducing the silent burden of impaired mental health," Journal of health communication, vol. 16, no. sup2, pp. 59–74*, 2011.

[2] A. David, "Schema mode assessment through a conversational agent," Master's thesis, Delft University of Technology, 2021.

[3] B. Franziska, "Natural language processing for cognitive therapy: extracting schemas from thought records," Master's thesis, Delft University of Technology, 2021.

[4] G. Alec, B. Richa, and H. Lei, *Sentiment140 Twitter Data Set*, 2009. Available at http://help.sentiment140.com/.

[5] H. Budi, "Automatic psychological text analysis using k-nearest neighbours," 2021.

[6] Z. Suo Xian "Mirijam", "Automatic psychological text analysis using recurrent neural networks," 2021.

[7] P. Jeongwoo, "Automatic psychological text analysis using support vector machine classification," 2021.

[8] G. Alec, B. Richa, and H. Lei, "Twitter sentiment classification using distant supervision," Master's thesis, Stanford University, 2009.

[9] K. Thomas, R.-K. Benjamin, P. Fernando, G. Brian, B. Matthias, F. Jonathan, K. Kyle, H. Jessica, G. Jason, C. Sylvain, I. Paul, A. Damián, A. Safia, W. Carol, and J. D. Team, "Jupyter notebooks – a publishing format for reproducible computational workflows," pp. 87–90, 2016.

[10] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[12] T. Danka and P. Horvath, "modAL: A modular active learning framework for Python," 2018. Available at https://arxiv.org/abs/1805.00979.

[13] K. Ishan, *Twitter Sentiment Analysis with Naive Bayes*, 2021. Available at https://www.kaggle.com/lykin22/twitter-sentiment-analysis-with-naive-bayes-85-acc.

[14] E. Andrea and S. Fabrizo, "Active learning strategies for multi-label text classification," Master's thesis, Istituto di Scienza e Tecnologia dell'Informazione, 2009.

[15] L. Wingho "Jimmy", "Generative algorithms to improve mental health issue detection," 2021.

A Logarithmic Curve Example

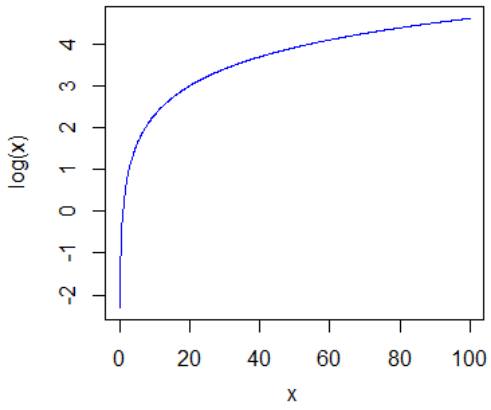

Figure 8: Example of a Logarithmic Curve

B Experimental Code Repo

Github Repo:
https://github.com/Jahb/ActiveMachineLearningResearch