Intelligent Systems

24678805 | January 2022

Coursework two

CALLUM MCLAUGHLIN

2022

Contents

[Introduction 2](#_Toc92300620)

[Literature Review 2](#_Toc92300621)

[Methodology 4](#_Toc92300622)

[Naïve Bayes 4](#_Toc92300623)

[Decision Tree 4](#_Toc92300624)

[Stacking 5](#_Toc92300625)

[K-Nearest Neighbour (K-NN) 5](#_Toc92300626)

[Majority Vote 6](#_Toc92300627)

[Experiments 7](#_Toc92300628)

[Classification Reports 8](#_Toc92300629)

[Decision Tree 8](#_Toc92300630)

[Naïve Bayes 9](#_Toc92300631)

[Stacking 9](#_Toc92300632)

[K-NN 9](#_Toc92300633)

[Majority Vote 9](#_Toc92300634)

[Conclusion 9](#_Toc92300635)

[Bibliography 10](#_Toc92300636)

# Introduction

The main focus of this task was on sentiment analysis which is an important field of AI which the main aim is to automatically categorise opinions expressed in text. The task was to design, develop and evaluate different AI algorithms in order to automatically analyse sentiment expressed by customer reviews of video games on amazon webstore.

The training dataset consisted of 5000 customer reviews while the test dataset also contained 5000 reviews. It is also important to note that between the two datasets there is no overlap. All algorithms tested in this task will be applied to these data sets that are provided.

In total there will be five algorithms tested, the two best performing will be modified further with parameters later. There will be two algorithms chosen that fit the task best and that have justifiable results when looking at their F-Score and Accuracy etc and they will be explored in more detail.

The five algorithms that will be used are: K-NN, Decision Tree, Stacking, Naïve Bayes and Majority Vote. As previously stated, two out of these five will be chosen as the best fit for the task.

# Literature Review

There are a huge number of existing AI systems on the market for sentiment analysis and they are a huge help to large companies and corporations around the world, especially companies that are driven by customer reviews and customer satisfaction.

Brand Health can be tracked by a type of AI system that will allow companies to understand how their customers feel about their brand. They can identify “Pain Points” and gather feedback on their products directly from the customers which will also help in meeting their demands, increase customer retention and improve their target audience methods and product offerings.

One of the most popular tools on the market for monitoring things like Brand Health is an AI system called HubSpot. HubSpot is a powerful and easy to use Customer Relationship Platform (CRM) which has all the tools and integrations needed for marketing, sales, content management and customer service (HubSpot, 2022). Each product on the platform is powerful alone but when it works together with each module of the system it is one of the best AI tools on the market.

There are five different Hubs within HubSpot, Marketing Hub, Sales Hub, Service Hub, CMS Hub and the operations Hub. Marketing Hub helps grow traffic and convert most visitors of websites to customers. It also runs complete inbound marketing campaigns at scale (HubSpot, 2022). Sales Hub is CRM software to help the company get deeper insights into prospects, automate the menial tasks and close more deals and sales faster (HubSpot, 2022).

Service Hub allows the company to connect with customers and this is the part of the whole system that mainly uses sentiment analysis on customer reviews of the company and their products to determine whether that review is positive or not. This allow for this statistic to be published in real time to a company’s website, for example, there could be a real-time statistic that updates on the footer of a website showing the percentage of customer revies that are positive (HubSpot, 2022).

The CMS (Content Management Hub) is software which uses AI to give customers personalized experiences. An example of what this could be is that a customer could be browsing the internet for a new phone, more specifically a Samsung phone. The customer then clicks on the company’s website and let’s say the company is a Mobile Phone Shop, the customer will get personalised advertisements on the front page for Samsung phones (HubSpot, 2022). The last Hub is the operations hub which manages the connections between all of the other 4 hubs and provides statistical data and predictions (HubSpot, 2022).

Another popular tool for sentiment analysis is Repustate which is the self-proclaimed “Smartest, fastest way to analyse customers and employee sentiments in any language.” (Repustate, 2022). Repustate’s AI-Powered technology is an all-in-one sentiment analysis platform that gives customer and employee insights with relatively no training required. It is not a complicated system and has ease of use design in mind.

Repustate can give companies advanced sentiment analysis and semantic search technology as well as multilingual analysis for 23 languages with no translation needed. It also provides data insights from social media, YouTube, TikTok and more. It offers real time reporting with actionable, visual data that is easily understood. The machine learning technology they employ monitors changes in local dialogue, slang and industry jargon to ensure the data is always current, this is a very impressive feature to have and especially important for having up to date data.

An example of the Sentiment Analysis API is given on the company website which allows you to generate a sample review and it shows the sentiment analysis summary to the right of it. An example of this is show below:

Graphical user interface, application

Description automatically generated

Image Reference: https://www.repustate.com/sentiment-analysis/

Although Repustate boasts a range of features and an easy-to-use UI, it only offers analysis across 23 languages, meaning that most of the sentiment analysis is only able to be done across 23 languages. If they could improve the number of languages, they can apply the sentiment analysis AI to, they would surely become a lot more successful.

Another very popular form of sentiment analysis is used on social media platforms which give people the chance to express their emotions and opinions publicly. The algorithms used on these applications work on real-time data, Twitter, Facebook and Instagram to name a few. During the COVID-19 pandemic, people from all around the worlds share their thoughts and emotions though these platforms, especially their thoughts on the pandemic.

Twitter being one of the world’s largest social media sites, using real-time data, storing the data into an understandable format and further analysing these datasets using NLP (Natural Processing Language) and Machine learning. By using sentiment analysis, the algorithm will detect the text polarity which will help to arrange assumptions onto one of three groups, Positive, Neutral or Negative (Rijwan Khan, 2020).

These social media sights that use sentiment analysis, as far as research has revealed, do not have sufficient systems in place to deal with slang terms used on the sites and sentiment analysis is not as good for things such as images and videos.

One thing that AI cannot factor in is irony, if someone mentions something negative but does so in an ironic way, the AI will simply deem that post as negative even though it could be much more than that.

Having said this, there is research being conducted in the field of sarcasm detection with NLP although it is a very narrow field of research within NLP research. This field of research usually doesn’t focus on the whole sentiment spectrum but rather the focus is on sarcasm. Therefore, the task of this field is to detect if a given text is sarcastic or not (Berasategi, 2020).

# Methodology

To start, the two algorithms that were chosen out of the five were K-NN and Majority Vote. These algorithms will be explored in more detail later but there will be a brief discussion about the other three, Naïve Bayes, Decision Tree and Stacking.

## Naïve Bayes

Naïve Bayes is a classifier learns to discriminate between features by starting with the assumption that each feature depends only on the label but NOT on other features i.e., features are independent and equal to each other (Edge Hill University, 2021). The naïve bayes classifier does work well for sentiment analysis however as with decision tree it usually works best in a hybrid approach. Naïve bayes also faces the ‘zero-frequency problem’ where it assigns zero probability to a categorical variable whose category in the test data set wasn’t available in the training dataset.

## Decision Tree

Decision Trees work by building the tree on training data and for each leaf node, it finds the number of training instances that activate that leaf node, Set Confidence = no. of activations. The tree can then be applied to test data. Decisions trees can work well for sentiment analysis but usually better in a hybrid format, however for this task it was decided K-NN and Majority Vote would be best as they are easy to use and they have good performance on the test data.

## Stacking

In order to learn how to discriminate between features, stacking aims at finding an optimal combination of sub-classifiers and makes use of a meta-classifier (The classifier that takes in all the predicted values of your sub classifiers) to learn to combine the individual sub-classifiers (Edge Hill University, 2021).

However, when using stacking pre-defined features do not always yield good performance and some features may be irrelevant to the classification task. Stacking seemed like a very good option but instead the alternatives were chosen as better options for their simplicity and ease-of-use.

## K-Nearest Neighbour (K-NN)

The first algorithm that was used was K-Nearest Neighbour which is a supervised classification algorithm with the hypothesis that “instances that look similar to each other should share the same class label” (Edge Hill University, 2021). The way K-NN works is that for each test instance, it looks at the neighbour training instances and finds the most popular label among the neighbouring training instances and assigns the most popular label to the test instance.

The value of the K parameter can be adjusted to fine tune the accuracy of the algorithm. The K parameter defines the size of the “neighbourhood”, k=1 will examine only the closest training instances, k=2 will examine the closest two instances and so on.

As for strengths, K-NN is robust to noisy training data and is also very effective when the training data is large. The disadvantages being that you need to fine tune it and determine the best value of the K parameter for it to work optimally. Computation cost is also quite high because the distance of each query instance to all training samples needs to be computed. Indexing may reduce this computational cost though e.g., K-D Tree (Kardi Teknomo, 2022).

The K-NN algorithm has been used in countless applications over the years, one such example being a Sentiment Analysis of Twitter’s US Airlines Data. A twitter US Airline Sentiment dataset was compiled with 14,601 tweets about each major U.S. airline which are labelled as one of three categories, positive, negative or neutral based on the nature of the respective Twitter user’s feedback in regard to that particular airline.

The dataset was also further divided into training and test sets in a stratified fashion. The training sets contains 11.680 tweets whereas the test set contains 2,921 tweets (Butt, 2021).

The aim of the project was to develop and train a K-NN classifier on the training set and use it to predict sentiment classes of tweets that are recorded in the test data set. An example of the training dataset is shown below:

Table

Description automatically generated with medium confidence

Image Reference: <https://towardsdatascience.com/sentiment-analysis-of-twitters-us-airlines-data-using-knn-classification-91c7da987e13>

K-NN is very useful in this scenario as it is a simple and easy-to-use supervised machine learning algorithm that also makes highly accurate predictions. It is also known as a Lazy Learner (Instanced Based Learning) meaning it does not learn anything in the training period as it does not derive any discriminate functions from the training data it is used on.

## Majority Vote

The second algorithm tested was majority vote which is an Ensemble method. Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produce more accurate solutions than a single model would, learning to discriminate between features by predicting the label with more than half the vote from the individual sub-classifiers (Necati Demi, 2021).

Diagram

Description automatically generatedDiagram

Description automatically generatedAs for the parameter’s, voting type was adjusted and tested for hard voting and soft voting. In classification, hard voting ensembles involve summing the votes for class labels from other models and predicting the class with the most votes whereas soft voting ensembles sum up the predicted probabilities for class labels and predicting the class label with the largest sum probability (Jason Brownlee, 2020). Shown below are two diagrams outlining how Hard Voting (Left) and Soft Voting (Right) is carried out by the classifier.

Images Reference: <https://vitalflux.com/hard-vs-soft-voting-classifier-python-example/>

When it comes to the advantages of the majority Vote classifier, they can create a lower variance and lower bias and an ensemble model also creates a deeper understanding of the data. Ensembles in general are used when the upmost accuracy is required (Kapalko, 2019). Ensembles also lower the spread and dispersion of the predictions and model performance (Jason Brownlee, 2020).

As accuracy is of paramount importance in this task, majority vote is a great choice as one of its main advantages is that it is an ensemble method which in this task made use of the K-NN classifier and a voting classifier, giving far more accurate results that most other classifiers.

In a paper published by Nikunj C Oza at NASA, he explains how Majority Vote classifiers are used against other methods of classification when discriminating between different designs for neural networks (Oza, 2008). This real-world application further reinforced that using Majority Vote for this task was a good idea.

# Experiments

The experiments performed involved creating the five chosen algorithms, loading the training and test data into each and executing the algorithm on the data sets. After this was done the summary report was used to determine which two algorithms would be chosen as fit for purpose for the task. The best two performing algorithms were K-NN and Majority Vote which achieved F-Scores of 0.64 and 0.68 respectively.

For K-NN the K parameter was tested from values of 1-20 and it was discovered that a value of 11 for K yielded the best Weighted F-Score. As for Majority Vote, the algorithm was tested with hard voting and soft voting with hard voting coming out on top by 0.01.

The accuracy of both algorithms was not the best overall they were 2nd and 3rd out of the five when scoring accuracy with scores of 0.64 and 0.68 respectively. The highest accuracy algorithm was actually naïve bayes, but it was not used as although some of the results in the summary were impressive, its not a good algorithm for sentiment analysis because of a number of disadvantages this algorithm has.

These disadvantages being that fact that it assumes that all features are independent which is rarely the case in real life and the ‘zero frequency problem’ that was mentioned earlier in the report. Naïve bayes can work well as part of a hybrid model (Vadapalli, 2021).

Although both algorithms work well in the given task, Majority Vote seems to be the best choice for this task as it makes use of the K-NN classifier and the Voting Classifier making it a hybrid model. Ensemble methods such as Majority Vote also have great advantages especially when working with large datasets.

The performance in general is better with ensemble methods as they can make better predictions and achieve better performance more than any single contributing model on its own. Ensemble models also have a robustness to them in the way that they reduce spread or dispersion of the predictions.

Although ensembles have many advantages, they can also be quite costly to train and develop. The ROI (Return on Investment) of an ensemble approach to a task should be considered carefully. Return on investment is a ratio between net income and investment (Mansa, 2021).

The classifications report for each of the five algorithms can be seen below.

## Classification Reports

### Decision Tree

A screenshot of a computer

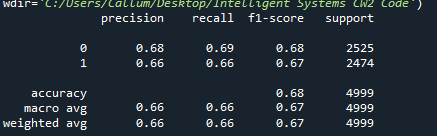
Description automatically generated with medium confidence

### Naïve Bayes

A screenshot of a computer

Description automatically generated with medium confidence

### Stacking



### K-NN

A picture containing graphical user interface

Description automatically generated

### Majority Vote

Calendar

Description automatically generated with medium confidence

# Conclusion

To surmise, the aims of the project was to develop and evaluate different AI algorithms in order to automatically analyse sentiment expressed though customer reviews of video games on the amazon webstore. There were five algorithms created and evaluated instead of the minimum requirement of two so that the investigation and evaluation would be thorough.

At the end of the project, five algorithms were coded and tested, all of which were briefly evaluated and explained when necessary. The final two algorithms that were chosen as the top two were then explored in more detail until the recommendation was finally made as to which algorithm best suited the presented task.

The aims of the project have been met well as there have been five algorithms tested, two tested and evaluated in detail, literature review and references to real world examples of how these types of algorithms are used and an abundance of extra reading and research was carried out for the project as can be seen from the number of references.

The code could have been more customised and fine tuned more in order to be more streamlined. The time it took to test and evaluate the algorithms could have been cut down a bit using other methods. For example, a for loop could have been used to run through a number of values for k and automatically set k as the value that gives the best F-Score.

At the time it did not occur to do this but looking back that is what would be done differently. There would also be more Hyper-Parameter testing for the other algorithms and evaluate whether they could achieve better F-Scores that the original two best performing algorithms with more fine tuning. Despite some of these developmental errors this project has been a success and these small errors will serve as valued learning for future projects.

# Bibliography

Berasategi, A., 2020. *Sarcasm detection with NLP.* [Online]   
Available at: https://towardsdatascience.com/sarcasm-detection-with-nlp-cbff1723f69a  
[Accessed 01 01 2022].

Butt, M. M., 2021. *Sentiment Analysis of Twitter’s US Airlines Data using KNN Classification.* [Online]   
Available at: https://towardsdatascience.com/sentiment-analysis-of-twitters-us-airlines-data-using-knn-classification-91c7da987e13  
[Accessed 02 01 2022].

Edge Hill University, 2021. *Blackboard - Week 6 Lecture - Bayes Classifier.* [Online]   
Available at: https://learningedge.edgehill.ac.uk/ultra/courses/\_275911\_1/cl/outline  
[Accessed 01 01 2022].

Edge Hill University, 2021. *Blackboard Week 8 Lecture - Random Forests and SVM.* [Online]   
Available at: https://learningedge.edgehill.ac.uk/ultra/courses/\_275911\_1/cl/outline  
[Accessed 21 12 2021].

Edge Hill University, 2021. *EHU Blackboard - Week 3 Lecture.* [Online]   
Available at: https://learningedge.edgehill.ac.uk/ultra/courses/\_275911\_1/cl/outline?legacyUrl=~2Fwebapps~2Fblackboard~2Fexecute~2Flauncher%3Ftype%3DCourse%26id%3D\_275911\_1%26url%3D  
[Accessed 03 01 2022].

HubSpot, 2022. *Content Management System.* [Online]   
Available at: https://www.hubspot.com/products/cms?hubs\_content=www.hubspot.com/&hubs\_content-cta=homepage-cms  
[Accessed 01 01 2022].

HubSpot, 2022. *Get Started With HubSpot Marketing Hub.* [Online]   
Available at: https://www.hubspot.com/products/marketing/get-started?hubs\_content=www.hubspot.com%2F&hubs\_content-cta=homepage-marketing  
[Accessed 01 01 01].

HubSpot, 2022. *Get Started With HubSpot Sales Hub.* [Online]   
Available at: https://www.hubspot.com/products/sales/get-started?hubs\_content=www.hubspot.com%2F&hubs\_content-cta=homepage-sales  
[Accessed 01 01 2022].

HubSpot, 2022. *Get Started With HubSpot Service Hub.* [Online]   
Available at: https://www.hubspot.com/products/service/get-started?hubs\_content=www.hubspot.com%2F&hubs\_content-cta=homepage-service  
[Accessed 01 01 2022].

HubSpot, 2022. *HubSpot - Home.* [Online]   
Available at: https://www.hubspot.com/  
[Accessed 01 01 2022].

HubSpot, 2022. *Operations Software.* [Online]   
Available at: https://www.hubspot.com/products/operations?hubs\_content=www.hubspot.com/&hubs\_content-cta=homepage-operations  
[Accessed 01 01 2022].

Jason Brownlee, P., 2020. *How to Develop Voting Ensembles With Python.* [Online]   
Available at: https://machinelearningmastery.com/voting-ensembles-with-python/  
[Accessed 03 01 2022].

Jason Brownlee, P., 2020. *Why Use Ensemble Learning?.* [Online]   
Available at: https://machinelearningmastery.com/why-use-ensemble-learning/  
[Accessed 21 12 2021].

Kapalko, R., 2019. *Predictive Model Ensembles: Pros and Cons.* [Online]   
Available at: https://blogs.perficient.com/2019/11/07/predictive-model-ensembles-pros-and-cons/  
[Accessed 03 01 2022].

Kardi Teknomo, P., 2022. *Strengths and Weaknesses of K Nearest Neighbor.* [Online]   
Available at: https://people.revoledu.com/kardi/tutorial/KNN/Strength%20and%20Weakness.htm  
[Accessed 04 01 2022].

Mansa, J., 2021. *Investopedia.* [Online]   
Available at: https://www.investopedia.com/terms/r/returnoninvestment.asp#:~:text=Return%20on%20investment%20(ROI)%20is,a%20number%20of%20different%20investments.&text=To%20calculate%20ROI%2C%20the%20benefit,the%20cost%20of%20the%20investment.  
[Accessed 04 01 2022].

Necati Demi, P., 2021. *Ensemble Methods: Elegant Techniques to Produce Improved Machine Learning Results.* [Online]   
Available at: https://www.toptal.com/machine-learning/ensemble-methods-machine-learning  
[Accessed 04 01 2022].

Oza, N. C., 2008. Classifier ensembles: Select real-world applications. *Information Fusion,* 1(1), p. 3.

Repustate, 2022. *Repustate - Home.* [Online]   
Available at: https://www.repustate.com/  
[Accessed 02 01 2022].

Rijwan Khan, P. S. A. K. A. T. A. M., 2020. SOCIAL MEDIA ANALYSIS WITH AI: SENTIMENT ANALYSIS TECHNIQUES FOR THE ANALYSIS OF TWITTER COVID-19 DATA. *Journal of Critical Reviews,* 0(0), pp. 1-2.

Vadapalli, P., 2021. *upGrad.* [Online]   
Available at: https://www.upgrad.com/blog/naive-bayes-explained/  
[Accessed 04 01 2022].