COMP2121 Data Mining
Coursework 3

Creating a Classification Model Using Online Sentiment to Determine Stock Price

Mining Minds:

Anna Jones (sc19akj)
Ayesha Rahman (sc21ar)
Conner Gulley (sc22cmg)
Geeyoon Lim (ml20g2l)
Nikolas Kontos (ed19nk2)
Tamar O'Rourke (sc21tor)

Background

In today's highly competitive and ever-evolving market, understanding customer sentiment and preferences is essential for businesses to remain competitive and successful. Online customer comments offer an unfiltered insight into the market's sentiment and a 2014 study by PowerReviews showed that 30% of 18-44 year-olds consult customer reviews every time they shop online- a percentage that is rising annually. By analysing these comments, we can determine the sentiment of consumers, which can be indicative of the market's perception of the company and its products/services. With advances in data analysis, it is now possible to extract and process data from the internet, on a large scale to find the overall sentiment and resulting trends towards a given product. Modern data analytics tools can be used to turn this large amount of indiscernible data (financial news, customer tweets, user reviews, etc.) into a classification model that can help investors make more informed investment decisions based on predicted market trends from said data.

Although sentiment analysis on social media has gained academic interest in recent times, there is still a limited understanding of how the market sentiment, represented by consumers' online reviews and tweets, correlates or indeed affects the stock price of publicly traded companies. Khedr et al. (1) conducted a study utilising a Naive Bayes classifier to analyse sentiment in financial news and historical stock market prices. They employed the K-Nearest Neighbours algorithm to predict future trends in the rise and fall of stock. This study was very successful, yielding a model that achieved an accuracy of 89.8% when predicting future stock fluctuations, which surpassed all previous studies.

Instead, our proposed model would employ data mining techniques to gather consumers' reviews and opinions directly from the review sections of retail websites, such as Amazon and Walmart, as well as social media posts. We will use this approach to search for comments related to newly released products. This data will enable us to perform sentiment analysis and determine whether the market's sentiment towards the product and its company is positive or negative. We will discretise comments on a daily basis and use our model to identify trends between days with positive/negative online reviews and corresponding increases/decreases in stock prices.

Evidence shows that a significant proportion of internet users rely on customer reviews to inform their purchasing decisions. This reliance demonstrates a vague relationship between widespread opinion and decision-making. The personal judgement of a product is not only publicised through online reviews but also through other internet platforms such as social media sites. Therefore, the widespread opinion of a product can be evaluated more accurately when this additional data is incorporated. The widespread opinion, or overall sentiment of a product can influence the market's desire to purchase it, and this demand is reflected in the stock price of a product. For this reason, correlations between the overall sentiment and stock price of a product exist and can be analysed using our model.

Our model is expected to provide significant insights into the product attributes that have the greatest impact on market sentiment, and, therefore, stock prices. This development is bound to have ramifications on the financial industry and data analysts, as it offers a novel tool for market analysis.

In summary, studies conducted both in the UK and internationally have shown that online sentiment has the potential to be valuable in predicting stock prices. In the past, researchers have explored the relationship between news articles and stock prices, while more recent studies have shifted towards the use of social media and online sentiment. We will utilise this approach in our model to create a classification system that could assist investors in making more informed investment decisions and potentially develop profitable trading strategies.

Contribution to Knowledge

Other studies have already examined the relationship between stock price and opinion—the previously described article by Khedr et al (2017) sought to predict stock market behaviour using financial news articles and historical behaviour (p.1). Another study, by Pagolu et al (2016) described gathering tweets over a set period and comparing it to stock behaviour over that set period (p.1346). These articles define “opinion” in different ways, with one relating to the opinion of experts in the field (and their written contemporary articles) while another gathers the opinions of Twitter users. The purpose of our proposal is to expand on these studies in two ways. Firstly, we intend to gather data during a product’s release window, instead of a preselected random period; and secondly, the data will be sourced not only from tweets but also from other social media platforms and customer reviews on e-commerce websites like Amazon and Walmart.

There are a number of disciplines this proposal can contribute to, in addition to data science (and computing more generally). Most notable is that of the business sciences, where having a stock market predictor based on public opinion would be useful in modifying well-known practices. Pathak (2021) notes that companies such as Amazon use data analytics to provide suggestions to users based on their preferences and habits (p.1). This proposal (and its analytic tools) could expand on that by studying whether certain advertising algorithms could lead to better-predicted stock market trends. For example, a study could be done to experiment with different advertising strategies for the same product and evaluate whether specific strategies are more influential than others to changes in the overall sentiment and stock price of the product. Once the study is completed, the results can be utilised to make more precise predictions about stock market trends based on the advertising strategies implemented by different companies. These predictions can subsequently facilitate more informed investment decisions.

There is also a psychological element to be explored, especially among this predictor’s usage with stockbrokers. Bakar and Chui Yi (2016) found that elements such as overconfidence and conservatism during a product launch period were major factors in stockbroker decision-making (p.319). Studies could easily be done to see if using the proposed model would lead to overconfidence or conservatism among stockbrokers, and cause them to buy more. Conversely, Naseem et al (2021) found that negative emotions and pessimism due to the COVID-19 era investment environment led to lacklustre investment in the market (p.1). Studies could also be done to the converse to discover whether a negative analysis result from our proposed model during a launch period would affect broker psychology and thus the market.

For each product we experiment with, our model will generate a larger dataset than those used in previous studies. This is because we will not limit the data retrieval to one platform but instead incorporate data from a wider variety of online platforms of different types. By doing so, a greater use of online sources is guaranteed, providing a more accurate overall sentiment to conclude more reliable correlations between data.

Overall, our proposal is distinct from the past—in specifying a time frame around product releases and widening the data set—and can be easily expanded on through implementation and testing under different strategies. Furthermore, many fields could take this proposal and find a use in it, as it were, by including it in their research.

Importance

In addition to academia, businesses have felt major effects from the advent of economic financial planning, analysis and customer feedback. According to a statistical analysis conducted by Ani Petrosyan (2023), the number of global internet users in 2022 reached 5.3 billion, demonstrating the widespread use of the internet and other related domains thus making it almost impossible to review each tweet or opinion on social media manually. Customers express varied opinions using complex and different languages, short texts, idioms, and different grammatical structures. Manually labelling such subjective feedback can give ambiguous results, leading to poor decision-making, which poses a challenge for companies and organisations looking to promote or release new products. Our proposal focuses on helping gather data during product release windows from all social media platforms and guide investors by allowing greater insight into customer feedback.

A study conducted by Jiekun Huang (2018) in the Journal of Financial Economics gathered a data set of more than 14.5 million customer product reviews on Amazon from 2004 to 2015. He found evidence related to stock pricing being affected by consumer opinions (p.164). Using various sources (such as Wikipedia Searches) and approaches he eventually arrived at a large dataset of user reviews with similar products from rival firms. Duplicates were deleted, and then by analysing all the reviews, and comparing them to the rival firms, he came to the conclusion that customer reviews impact a high value of cash flow in the financial market.

While the above is one example of opinion mining, it is necessary to look at another example of this technique to truly understand its importance. In a study done by Derakhshan and Beigy (2019), two datasets in Persian and English were collected. These included comment analysis from a market social network dataset and market data that shows daily prices per share in different languages. A Support Vector Machine (SVM) was used to predict stock change by extracting a set of features of each language and classifying them based on categories - increases or decreases the stock price. The study used the “LSA-POS" method incorporating parts-of-speech tags into topic modelling. His study targeted two different languages and their intonations in speech and writing proving that customer response can majorly change the flow of cash in the financial market irrespective of language (pp. 572-574).

A study done by Asur and Huberman (2010) measured the polarity ratio, which is the ratio between extremely positive and negative reviews on all platforms, as it can provide a more nuanced understanding of customer sentiment towards a particular product or company (p.498). This can be particularly useful for investors, as the polarity ratio can help them make better-informed decisions and potentially influence stock prices. Multiple studies have shown that customer reviews play a significant role in the direction of stock change. Therefore, by analysing customer responses and discussions based on datasets acquired through various sources such as tweets, social media, and e-commerce websites like Amazon and Walmart, we propose to develop a sentiment analysis method to predict stock changes. This approach will provide emerging industries and investors with a reliable method for predicting stock changes regardless of any other factors.

Overall, the use of a classification model in financial analysis has the potential to significantly enhance the efficiency and effectiveness of the stock market. One potential advantage of this is the ability to facilitate more informed investment decisions for investors, leading to more efficient allocation of capital and potentially contributing to economic growth. This could benefit society as a whole by creating job opportunities and improving living standards.

Research Hypothesis & Objectives

The objective of this proposal is to develop a model that can identify whether customers are reacting positively or negatively and find correlations in discretized stock price data. Understanding how important product qualities affect sentiment can help analysts and investors make better investment decisions.

The research hypothesis is that stock market movements for recently released items could be predicted using sentiment analysis of social media posts and consumer evaluations. We particularly postulate that it is possible to use online reviews to predict changes in stock prices. This research may have larger implications for understanding how people use online platforms to communicate their opinions and feelings in addition to its possible effects on investment choices.

The proposed project is innovative and timely since it looks at the possibility of forecasting stock market movements utilising sentiment analysis of customer reviews. While research has been done in this area, it has never been complete enough to form a well-rounded picture of the relationship between consumer sentiment and stock behaviour. Previous studies have mostly used historical behaviour and financial news items to examine the relationship between consumer mood and stock market trends. Khedr et al (2017) discovered that a more thorough approach is required to comprehend this link. This concept involves gathering data during a product's delivery period and leveraging a larger variety of data sources to gauge customer sentiment, building on previous studies.

The creation of a model for forecasting stock market changes based on consumer sentiment research is one of the project's main goals. The effort will also evaluate the suggested model's effectiveness and accuracy and contrast it with other strategies that are currently in use. The project's measurable objectives are as follows in order to meet these objectives:

1. Gather and evaluate user reviews for a recently released item sample.
2. Use the gathered data to create and verify a sentiment analysis model for forecasting stock market movements.
3. Analyse the connections between relevant product qualities and the sentiment shown in customer reviews to understand how these attributes affect market sentiment and stock prices.
4. Examine the possibility of using the sentiment analysis model to guide investment choices and evaluate the model's influence on stockbrokers' behaviour.

The project will provide for predicting stock market movements based on consumer sentiment analysis and customer review datasets for a sample of recently released items such as the study by Ho T and Huang Y (pp. 2-3). It is expected that the outcomes will include a deeper understanding of the relationship between consumer mood and stock market patterns, along with the possibility of using this correlation to guide investment decisions. Consequently, analysts and investors are better educated to make informed investment decisions, which can result in improved stock market forecasting accuracy and efficacy.

In summary, the purpose of the proposal is to construct a stock market trend prediction model based on customer sentiment research and to compare the accuracy of the model and efficiency to those of other approaches. This study could assist investors in making more knowledgeable investment decisions by identifying the critical product characteristics that affect market sentiment and stock prices.

Programme and Methodology

We will use the following sentiment analysis method to evaluate customer response over a given period, building a model to evaluate sentiment across user reviews on the Internet. This method will be repeated across a set of chosen products in iterations, for a more robust and general model.

In identifying stakeholders, it is important to note that markers of progress at each stage of the experiment are vital. To that end, our plan, as outlined in the work plan Diagram below, is to have weekly benchmarks set so that each half of the team can present results (whether it be data collected, or improvements to the classifier, among others) to the stakeholders (i.e. the body and funds and monitors the experiment itself). This should allow for both continuous feedback and a clear distinction between the team and the outside partners and stakeholders.

Data Collection

User responses to a given product will be collected from social media, such as Twitter, and e-commerce websites, such as Amazon. This can be achieved using tools such as SketchEngine, or using website-specific API. SketchEngine's WebBootCat is a viable option for fast, filtered web page retrieval. As it is for text extraction only, SketchEngine will leave HTML syntax and other artefacts, which must be removed by hand or by using/developing a tool. For social media data, many social media sites have an API, which may allow for downloading posts on a mass scale. However, due to social media's general-purpose use, it is expected that many posts including our product name may not be about the user's opinion on the product itself. Filtering out unhelpful tweets will be part of our process.

Our data collection will begin by downloading and manually reviewing and labelling 500 user posts (for our proof of concept as shown in the Appendix, we used 100 posts—we plan to expand on this within our full research project). Each individual post will also have an appropriate weighting for the popularity of the post (for example, views or likes regarding a tweet). This will be used when calculating the overall sentiment during our testing period, but will not be considered when classifying a single post's sentiment. If a platform cannot gauge a post's popularity, it will automatically be given a weight of 1.

Assigning a Post's Sentiment

Approach

There are multiple ways to assign the sentiment of a user's post, with several methods in the research mentioned above. While we may use existing tools to demonstrate the viability of the project, these tools will not be the main source for sentiment analysis. Instead, we will be building our own bespoke sentiment analysis model. This process will require manual classification and training. We will go through our collected user posts, and manually classify each post as one of three possibilities – "positive", "negative" or "neutral". This may be stored in a simple database format, such as a csv. Once we have a working model, we will begin automating part or all of this process in order to finetune the model's accuracy.

Posts we cannot classify will be discarded, and adjustments will be made to our data collection method (such as keyword filtering and profanity filtering). Examples of posts that will be discarded are as follows - "troll" posts, posts in languages other than English, "empty" posts (no text, only attachments), etc. We may be able to use a user's location, or a pre-existing language classifier, to determine a language from a post. It would be ideal to build a sentiment classifier to be compatible with other languages, using a representative corpus. This is outside the scope of our proposal, but is a viable next step, given this research is successful.

Problems

Negation Tagging

All posts will also be analysed for negation, by tagging words following negating statements, such as “not”, according to English syntax. The antonyms will be labelled by appending an appropriate tag, eg, “_NOT” to the end of each word. For example, “I did not like this product, it’s terrible” becomes “I did like_NOT this_NOT product_NOT, it’s terrible” - stopping negation at a comma, or other connective syntax. These search and replace tasks can easily be performed using regular expressions. Negation will be considered when using Word2Vec to determine a word’s sentiment.

Word and Emoji Removal

We will also remove frequently occurring words, which appear commonly throughout all categories of classification (grammatical features such as “a”, “the”, the name of our product or company, etc). This will prevent potentially erroneous classification for posts featuring these words. This will be achieved by using a simple program to count word frequency across all documents, knowing the sentiment - if a word appears frequently across all sentiments, then remove the word from all posts. Duplicate words will also be removed from a post.

We will also be removing emojis – these are very important for online communication, hence this is a weakness in our strategy. Considering the context-based nature of emojis, the problem of annotating them would be too complex for our project. It would be ideal to build a corpus of language used on social media, however, that is not within the scope of our project.

Using Word Embeddings to Represent Meaning of User Reviews

We can use an existing Word2Vec model to analyse the meaning of each post and output a vector representation of each token within the post. The number of dimensions depends on the model - typically, a large corpus will have 300 or below different dimensions. If a word does not exist in the model, we will remove that word from our tweet, preventing it from having an effect on a post’s overall sentiment. If all words are removed from a tweet, we can discard the post.

For our negated terms, we can attempt to find the n-tuple value of an antonym for the word – either we can use the average difference in vector space between a collection of well-known antonyms (“light”, “dark”), and apply it to our root word tuple representation to this different ([word value] - ([light]-[dark])) – however this method may not be effective. If the resultant n-tuple has any neighbouring words within a certain radius (this will be determined during testing), then we can assume that this is a valid antonym. Else, if no antonym exists, then we can assume it is not negatable and can use the sentiment of the original word. We expect this to happen for nouns. There are also existing antonym tools, which we may use to address this problem.

Now that we have our word embeddings, we can now determine the numerical value of a tweet’s sentiment. By calculating the average n-tuple of the words within the tweet, we now have a set of n numerical values, which can be used for sentiment classification in WEKA. These will be saved in a csv file, which will be converted into an arff file for use in WEKA.

Analysis via WEKA

We can now run our analysis via WEKA, as each post has a standard set of n numerical values (making up its vector), and its assigned class (“positive”, “negative”, “neutral”). We will create a .arff file that includes our n dimensions. The classified information will be split into two sets: a training set, and a test set.

We use a classifier in WEKA to build a model. We may test this with a series of classifiers, but an example of an appropriate classifier is Naive Bayes. This is appropriate, as the classified posts are represented numerically only. Our numerical representations will be tested against multiple different classifiers until an example above a certain threshold (90%) is found. For proof-of-concept purposes, J48 and Naive Bayes were used, but the intention for the full experiment is to use a variety of classifiers. In the event that no classifier is found, then either our method is insufficient, or we have not obtained enough manually classified information. After repeating this method using 1000 example user reviews, if a sufficient classifier is still not found, then we must alter the way in which we approach assigning word values.

Capturing the Overall Sentiment Across All Posts

Now that we have a classifier, we can now determine the classification of all suitable posts, using the mathematical equation below. This classification method can simply be automated, knowing the post’s average Word2Vec, and its trending weight. The sentiment will be a normalised decimal value between -1 and 1.

$$product\ sentiment = \sum_{i \in posts} (sentiment(i) \cdot trending(i))$$

The sentiment of a single post will be representative of the classification, output from our classifier:

- Positive sentiment results in a sentiment value of 1
- Negative results in a sentiment value of -1
- Neutral results in a sentiment value of 0

The trending value will be our “weight” for the given post. We may also normalise this behaviour by dividing our overall sentiment by the total weighting of the response.

Comparing to Stock Market Data

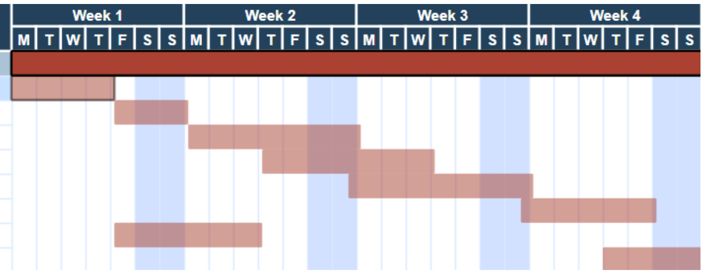
We will now compare the overall sentiment to our stock market data. We will find the average stock market price between a release window for our particular product, and prior to a release. The percentage change in stock price before and after our release window will then be stored along with the normalised sentiment.

We will then train a classifier, similar to our sentiment classifier, in order to recognise the expected growth rate of stocks following the normalised sentiment response on social media. This will require a repeat of our methodology across a wide range of stock values in the market.

We will train this model by repeating our methodology for many different companies/posts.

Workplan Diagram (Gantt Chart)

	Task Name	Duration	Start
1	Complete project execution	28 days	W1, D1
2	Project Planning & Task Assignment	4 days	W1, D1
3	Classifer Design and Discussion	3 days	W1, D1
4	Data Collection - Twitter & SketchEngine	7 days	W2, D1
5	Labelling of Data	7 days	W2, D4
6	Classifier Training and Reworking	7 days	W3, D1
7	Classifer Testing & Result Collection	5 days	W4, D1
8	Stock Market Data Collection	6 days	W1, D4
9	Final Results and Writeup	3 days	W4, D4



Gantt chart template by diagrams.net (see References)

Explanation of Chart Items

- *Complete project execution*: Time estimate to complete the project as a whole.
- *Project Planning & Task Assignment*: Planning period at start of project; also involves task assignment to each member/pair.
- *Classifier Design and Discussion*: Done in parallel with Project Planning and Task Assignment; planning of classifier design based on preconceived ideas and prior planning.
- *Data Collection - Twitter and SketchEngine*: Collection of data using SketchEngine and Twitter.
- *Labelling of Data*: Manual labelling of data for sentiment analysis purposes & placing reviews into Word2vec model for later use.
- *Classifier Training and Reworking*: Training of the classifier based on labelled data from *Labelling of Data* and reworks based upon results.
- *Classifier Testing and Result Collection*: Use of the now-reworked classifier on the test set and collection of results (i.e. accuracy, etc) from it.
- *Stock Market Data Collection*: Collection of stock market data from various sources.
- *Final Results and Writeup*: After the results have been collected from the classifier, time is given here to allow for these to be examined and conclusions drawn.

Workplan Task Possibilities

The work plan as laid out allows for much parallel work; once the initial parameters have been set out in *Project Planning and Task Assignment*, the team can split in two; one half focusing on classifier design and the other beginning to collect stock market data. At the end of Week 1, the classifier assignment half can move on to begin to collect data as specified by the *Data Collection - Twitter and SketchEngine* task. By the middle of Week 2, there should be enough data to begin manual labelling, which can be taken over by the former stock market data collection half. This leapfrogging pattern can continue through the middle of Week 4, with the classifier beginning to be trained, and then tested, as Week 2 ends, Week 3 runs, and Week 4 begins. In the last few days of the project (and Week 4), the team once again comes together to view the results and draw conclusions, outlined by the *Final Results and Writeup* task.

Appendix

In order to demonstrate the feasibility of our research, we created prototypes of the major parts of our proposal. These prototypes demonstrate the strengths of our method and work plan, as well as any accommodations needed to guarantee the success of this project.

Data Collection & Labelling

Collection & Labelling - Methods

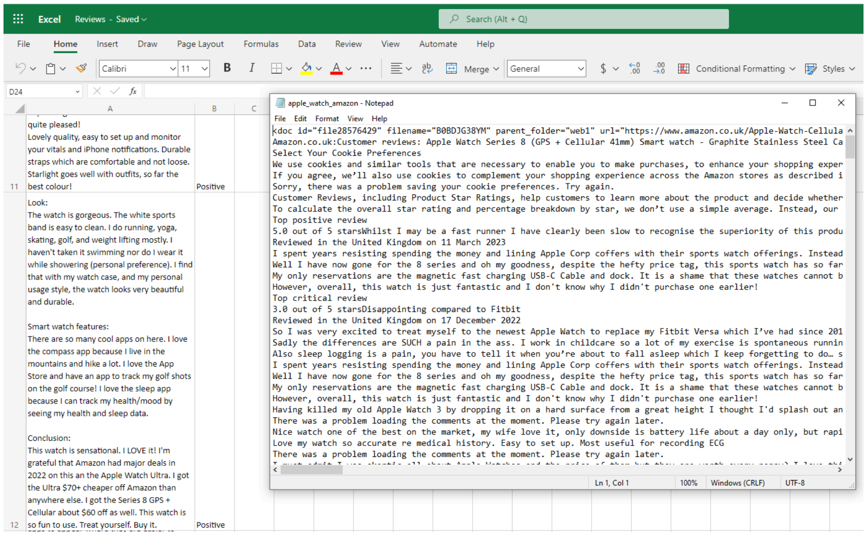
For the purposes of sentiment analysis, SketchEngine was used to collect raw webpages including user reviews from websites such as Amazon.co.uk and other e-commerce sites using keywords. These websites provided user reviews for the chosen product, which were loaded into a .txt file by SketchEngine. Those reviews were then manually labelled to be positive, negative, or neutral as specified by the Programme and Methodology section above. Tweets were also attempted to be gathered using the Twitter API, but as explained below in detail generally speaking the tweets themselves were irrelevant for sentiment analysis purposes.

Collection - SketchEngine Difficulties

There were several problems with using SketchEngine—most notably that often HTML elements were left in the corpus gathered. HTML tags and duplicate reviews led to a more time-intensive process in finding and labelling user reviews. In addition, labelling of the reviews themselves proved difficult, as having no filters selected for product reviews often left a majority of them being positive, leading to a skewed dataset with only a few negative reviews for training & classification purposes.

Collection - Twitter Difficulties

There are a limited number of options available when trying to collect large numbers of tweets and each has its drawbacks. Twitter's own API requires users to apply for access to its developer tools, but we were denied, and our reapplication has had no response. This led us to try alternatives eventually finding TweetDeck which allowed us to place filters on the search for Twitter so that only tweets including the phrase “apple watch” and not containing “www/com/win/prize/amazon/sale” left us with primary customers commenting on their apple watch experience. Due to Twitter’s UI, when using SketchEngine or other web scraping techniques, we were unable to capture the likes, views and retweets of each tweet. Had we gotten paid access to Twitter's API we could have kept this key data and used it to weigh out sentiment analysis. We then used ChatGPT to “remove any tweets not resembling customer reviews or opinions and format them so that it is just the user, date and their comment”. This yielded an almost perfect output, leaving only 1 out of the 28 adverts from the original 90 total tweets.



Labelling reviews from SketchEngine-generated .txt file

Twitter tweets are still the most viable form of user data, however—easiest to access (once approved) and short enough to not take up much storage space—and after this process with user reviews, they still seem fairly relevant. Collecting impact details of them as well as much easier—likes, retweets, etc. are more easily captured than having to scrape and discover how many people were “helped” by a review, for example. In conclusion, while manual labelling will be important at the start of this project, automating the classification of training data (and filtering irrelevant data) will be important early on.

Using one training dataset with WEKA

Input & Difficulties

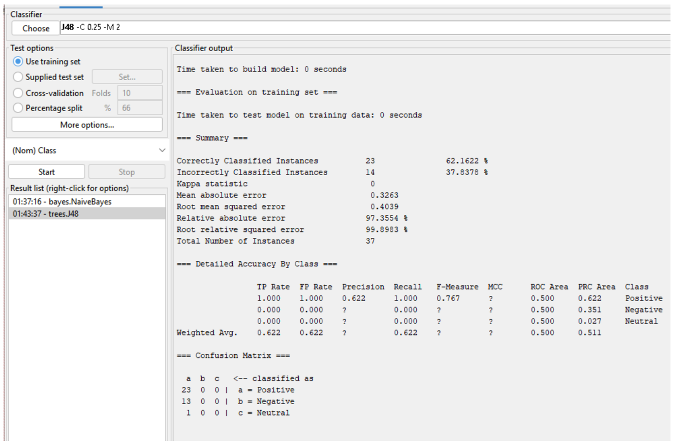
The dataset had two attributes: the text of the review and the sentiment class, which could be either Positive, Negative, or Neutral. We trained two machine learning classifiers, Naive Bayes and J48, to predict the sentiment class of customer reviews on Amazon.

A CfsSubsetEval attributes evaluator and the BestFirst Search strategy were used to train the J48 classifier. The most educational aspect was determined to be the Text Property. The default parameters were used to train the Naive Bayes classifier. On the training set, the accuracy of both classifiers was assessed.

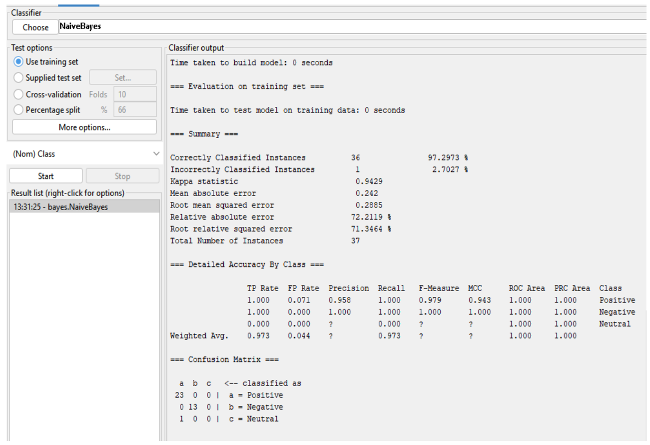
During our testing of WEKA, we experienced several drawbacks due to the word limit of our dataset. General WEKA did not allow exceeding a certain word limit per input in the test set. We also found that the accuracy of our results decreased with shorter tests, as WEKA requires a larger sample size to produce accurate results. Another factor affecting the results obtained was the lack of flexibility in the algorithms provided, as we were unable to customise or modify them to suit our specific needs. This made it difficult for us to tailor our approach to the problem we were trying to solve. Additionally, we found that WEKA was not well-suited for processing large datasets due to its limited scalability, which was another drawback that we had to work around.

Results

Given the training set, the J48 classifier reached an accuracy of 62.2%, properly classifying 23 cases as positive and incorrectly classifying 13 examples as negative. The Naive Bayes classifier has a 97.3% accuracy rate with just one occurrence misclassified. According to the Naive Bayes classifier detailed accuracy by class, the Positive class obtained a true positive rate of 1.0 and a false positive rate of 0.071, whereas the Negative class attained a TP rate of 1.0 and an FP rate of 0.0. For both classes, the recall and accuracy were likewise quite impressive. However, because there were no instances of the Neutral class in the training set, the TP rate, accuracy, and recall were all 0.



J48 Classifier



Naive Bayes Classifier

From the finding of the research, the J48 classifier dramatically underperformed the Naive Bayes classifier on the tiny dataset of customer reviews. This is probably considering Naive Bayes is a more straightforward and effective algorithm that works better with fewer datasets. J48, on the other hand, is a more complicated algorithm and could work more efficiently with bigger datasets. Nevertheless, it should be emphasised that the dataset was rather small, and the classifiers' performance might vary on bigger datasets. The dataset was also unbalanced, with the Positive class receiving the majority of the cases. The performance may have been impacted because of this, since they may have been biased towards the dominant class.

In summary, this study discovered that, on a small dataset of customer reviews, the Naive Bayes classifier outperformed the J48 classifier in terms of accuracy. The results indicated that J48 may surpass Naive Bayes on bigger datasets, whereas Naive Bayes is ideal for sentiment analysis tasks on short datasets. To confirm these results on bigger and more varied datasets, however, more study is required.

- Tools used to write and formulate documents- chat gpt (for the project name, pilot study), grammarly to check the grammar of documents, google scholar

Report Write Up

To conduct our research, we used Google Scholar, an online database that enables users to search for relevant materials, in which we located articles and studies relating to our proposed topic. It allowed us to filter our search results to obtain the most suitable information, saving us considerable time and enhancing the quality of our report. Additionally, we utilised Grammarly, a text analysis tool, to identify grammatical, spelling, and punctuation errors and provide suggestions for alternative sentence structures throughout the report. In summary, the incorporation of Google Scholar and Grammarly into our research and writing process contributed significantly to the production of an accurate and well-written report.

References

Asur, S. and Huberman, B.A. 2010. Predicting the Future with Social Media. *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. pp.492-499.

Bakar, S. and Chui Yi, A.N. 2016. The Impact of Psychological Factors on Investors' Decision Making in Malaysian Stock Market: A Case of Klang Valley and Pahang. *Procedia Economics and Finance*. **35**, pp.319-328.

Derakhshan, A. and Beigy, H. 2019. Sentiment analysis on stock social media for stock price movement prediction. *Engineering Applications of Artificial Intelligence*. **85**, pp.569-578.

Huang, S. 2018. The customer knows best: The investment value of consumer opinions. *Journal of Financial Economics*. **128**(1), pp.164-182.

Ho T. and Huang Y. 2021. Stock Price Movement Prediction Using Sentiment Analysis and CandleStick Chart Representation. [Online]. [Accessed 15 March 2023]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8659448/

JGraph. 2021. "diagrams.net, previously draw.io, is an online diagramming web site that delivers the source in this project." [Online]. [Accessed 30 April 2023]. Available from: https://github.com/jgraph/drawio

Khedr, A.E., Salma, S.E., and Yaseen, N. 2017. Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. *International Journal of Intelligent Systems and Applications(IJISA)*. **9**(7), pp.22-30.

Naseem S, Mohsin M, Hui W, Liyan G and Penglai K. 2021. The Investor Psychology and Stock Market Behavior During the Initial Era of COVID-19: A Study of China, Japan, and the United States. *Frontiers in Psychology*. **12**, pp.1-10.

Pagolu, V.S., Reddy, K.N., Panda G. and Majhi B. 2016. Sentiment analysis of Twitter data for predicting stock market movements. *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*. pp. 1345-1350.

Pathak, S. 2021. *How Amazon Uses Big Data?*. [Online]. [Accessed 30 March 2023]. Available from: https://www.analyticssteps.com/blogs/how-amazon-uses-big-data

Petrosyan, A. 2023. *Number of internet users worldwide from 2005 to 2022*. [Online]. [Accessed 29 April 2023]. Available from: https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/

Power Reviews. 2014. *The Power of Reviews*. [Report]. [Accessed on 20 March 2023] Available from : https://www.powerreviews.com/wp-content/uploads/2015/08/13185402/ThePowerofReviews-Report.pdf