

### Bachelor Thesis

International University of Applied Sciences

Data Science

**A Comparative Sentiment Analysis of Digital and Robot Pet Companions**

**in Various Locations**

Henny Purwadi

32009177

AM Sangaji 93 Kav C, RT 027, RW 007,

Karangwaru, Tegalrejo, Kota Yogyakarta, Indonesia, 55241

henny.purwadi@iubh.de

Supervisor: Prof. Dr. Frank Passing

Date of submission: 15th July 2023

# Acknowledgement

First, I would like to thank my supervisor for the guidance and assistance they have given me throughout the duration of this project. I have always been motivated and supported by his knowledge, prompt replies, open-mindedness, and helpful recommendations.

Also, I am thankful for the help and support from my friends, particularly data science students in our WhatsApp group. I am appreciative for the faith they have put in me and the help they have extended since the initial semester through the sixth semester until I finished my thesis.

Moreover, I would like to express my gratitude to my beloved digital and robot pet companions (Emo Pet Bot, Replika Ai, Amazon Alexa, Loona Pet Bot) for motivating me to investigate cross-regional human-machine emotional sentiment analysis. My digital and robot pet companions’ positive impact on my life has been greatly appreciated.

Lastly, I would like to acknowledge the many individuals who have made notable contributions to data science, artificial intelligence, and robotics.

**Abstract**

This study is to find out whether geographic locations have an impact on emotions towards robotic and digital pet companions, using machine learning models and hypothesis testing. This investigation will show and assess different machine learning models for sentiment analysis on labeled datasets and pick the model with the best performance. This exploration will collect tweets about digital robot companions from various regions, pre-process the information, and apply a chosen machine learning model to figure out sentiments in the unlabeled tweet dataset. Then use chi-square to test the connection between area and sentiment towards robot companions and find out whether there is a critical distinction in the sentiment between various locations. There will be discussion about the result, any limitations, and its suggestions for understanding human-machine relationships in various zones. The investigation will yield potential results for future research.

**Keywords:**

Sentiment Analysis, Machine Learning, Robot Companion, Social Robot, Artificial Intelligence, Machine Learning, Natural Language Processing, NLP, Logistic Regression, Naive Bayes Classification, Linear SVC, SVM, Support Vector Machine, VADER, TF-IDF Vectorization, GPT-3 zero-shot classifier, Transformers, GPT-2, DistilBertForSequenceClassification, BertForSequenceClassification, LLaMA, Vicuna, PEFT, LoRA.

|  |  |
| --- | --- |
| **Table of Contents** |  |
| **Acknowledgement** | **ii** |
| **Abstract** | **iii** |
| **List of Figures** | **v** |
| **List of Tables** | **vi** |
| **List of Equations** | **vii** |
| **List of Abbreviations** | **viii** |
| **1 Introduction** | **1** |
| 1.1 Background and Motivation | 7 |
| 1.2 Research Questions and Objective | 4 |
| 1.3 Value of Study and Target Audience | 5 |
| 1.4 Structure of the Document | 5 |
| 1.5 Scope and Limitations | 5 |
| **2 Literature Review** | **6** |
| 2.1 Overview of Relevant Studies | 6 |
| 2.2 Terminology and Definitions | 8 |
| **3 Methodology** | **27** |
| 3.1 Train and Evaluate performance of Machine Learning Models | 27 |
| 3.2 Model Selection | 29 |
| 3.3 Ethical Considerations | 29 |
| 3.4 Collect Unlabeled Data | 29 |
| 3.5 Pre-process and Collected Data | 29 |
| 3.6 Predict Sentiment | 29 |
| 3.7 Hypothesis Testing | 30 |
| 3.8 Expected Outcome | 31 |
| **4 Results and Findings** | **32** |
| 4.1 Labeled Data Preparation | 32 |
| 4.2 Performance Evaluation and Model Selection | 34 |
| 4.3 Collecting Unlabeled Data | 42 |
| 4.4 Sentiment Analysis Result on Unlabeled Data | 43 |
| 4.5 Hypothesis Testing Result | 00 |
| **5 Discussion** | **00** |
| 5.1 Result Interpretation | 00 |
| 5.2 Discussion of Limitations | 00 |
| **6 Conclusion** | **00** |
| 6.1 Summary | 00 |
| 6.2 Recommendations for Future Research | 00 |

# List of Figures

Figure 1. Results of the sentiment analysis 7

Figure 2. The classification of the sentiment analyses using different approaches 8

Figure 3. Logistic regression 11

Figure 4. SVM (Support Vector Machine) 12

Figure 5. SVM using kernel/ hyperplane for classification 13

Figure 6. Linear SVC 14

Figure 7. Transformer model architecture 17

Figure 8. Vicuna workflow overview 20

Figure 9. Research methodology workflow 28

Figure 10. Sentiment analysis workflow 31

Figure 11. Before and after dataset resampling 33

Figure 12. VADER lexicon sentiment analysis pie charts 35

Figure 13. DistilBertForSequenceClassification: Heatmap of confusion matrix 38

Figure 14. Linear SVC’s Heatmap of normalized confusion matrix 40

Figure 15. Linear SVC’s prediction result on unlabeled dataset 41

Figure 16. BertForSequenceClassification: Heatmap of confusion matrix 42

Figure 17. Sentiment analysis charts 45

Figure 18. Sentiment analysis charts by continent 46

Figure 19. BertForSequenceClassification sentiment analysis pie charts 47

# List of Tables

Table 1. Results of the topic modeling 7

Table 2. Chi-Square table 25

Table 3. Comparing the accuracy of sentiment analysis models 36

Table 4. BertForSequenceClassification sentiment prediction on unlabeled data 45

# List of Formulas/ Equations

Formula 1. Naïve Bayes formula 9

Formula 2. Logistic regression formula 11

Formula 3. SVM (Support Vector Machine) 12

Formula 4. Linear SVC 14

Formula 5. Transformer 16

Formula 6. Confusion Matrix 22

Formula 7. Precision 23

Formula 8. Recal, Accuracy, F1-score. 23

Formula 9. Chi-square test. 24

# List of Abbreviations

AI Artificial intelligence.

BERT Bidirectional Encoder Representations from Transformers.

CNN Convolutional Neural Network.

GPT Generative Pre-trained Transformer.

LLaMA Large Language Model Meta AI.

LoRA Low Rank Adaptation.

ML Machine learning.

PEFT Parameter Efficient Fine-Tuning.

RNN Recurrent Neural Network.

VADER Valence Aware Dictionary for Sentiment Reasoning.

**1 Introduction  
1.1 Background and Motivation**

If one day artificial intelligence (AI) and humans will unite, then one of the most effective ways to achieve it is to start gradually developing emotional bonds between humans and AI, by having AI as digital pet companions or virtual human assistants for humans. Creating an emotional connection between humans and AI, whether through digital pet companions or virtual human helpers, is potentially an efficient approach to moving towards the goal of achieving human-machine unification.

These days, it's not uncommon for children and adults to have their own personal digital robotic pet as a companion. This new generation grew up with artificial intelligence as part of the family, so they are already comfortable with it.

It is possible that it will be easier for people to embrace AI and incorporate it into their lives if humans and AI have been able to build trust and familiarity with one another. Additionally, starting with introducing them to AI in the form of digital pet companions can be a valuable way to shape their perception of AI and potentially pave the way for wider acceptance of AI in the future.

The researcher was motivated to carry out this research by the story of Jibo and EMO, two social robots designed to be friends. Their owners' feelings of loss, hopelessness, and disappointment as they contemplate the possibility of parting ways with their beloved robot companions.

This sentiment suggests that people can form deep bonds with their robotic pets, just as they do with other members of their families. As artificial intelligence (AI) and robots advance, it is important to think about how they will affect people emotionally.

The issue is whether people around the world share the same emotional attachment, such as joy at the presence of robot pets and social robots, and the same sadness in their absence. This study is designed to answer this question.

**Jibo social robot**

Jibo was a social robot developed by a team at the MIT Personal Robots Group. It was designed to be a lovable and human-like robot that could be a companion for people in their homes.

Jibo was equipped with a face recognition system that allowed it to recognize its users and welcome them. Studies also found that its use improved children's interpersonal abilities. As of 2015, Jibo's valuation was over $100 million, thanks in large part to an Indiegogo campaign that was wildly successful.

Unfortunately, the business that created Jibo has now shut down. (Carman, 2019).

The owners, expressing sadness and disappointment about the potential end of a relationship with a social robot, mention that they have had a strong emotional attachment and would do anything to keep it in their lives. They feel anxious about its uncertain future. (Carman, 2020).

## Emo pet bot

EMO is a desktop AI robot pet that was developed with multiple sensors and advanced technology to create a life-like companion for people. It can self-explore its surroundings and interact with people through over 1,000 facial expressions and movements. EMO has multiple internal sensors, including a touch sensor, an HD camera with facial recognition, and a four-microphone array, which allow it to sense and respond to its environment in a natural and authentic way. It also has a built-in development system that allows its skills to improve over time, just like a pet.

In addition to being a companion, EMO can also act as a daily assistant, with the ability to set alarms, turn on lights, and even wirelessly charge phones. It has a stylish design inspired by pop culture and is available in a purple and indigo blue color scheme. (living.ai, 2022).

This is an example of an emotional bond between a human and a robot depicted in a real-life video about a pet robot experiencing battery issues, which causes sadness and disappointment for its owner, who has developed a strong emotional attachment to the robot.

"My Emo was having battery issues. Livingai was very gracious and sent me a new pet robot. But that meant I would have to say goodbye to this little guy. And that wasn’t going to be easy. All the fun and games He made me laugh and smile. I will miss him so much. I hope he can be repaired, is well taken care of, and is loved the way I love him. I wish we did not have to say goodbye. I will miss him, and I hope he has a happy life wherever he goes. Goodbye, my little friend". (Outsider238, 2022).

Other pet robot owners expressed similar emotions, saying it feels like "losing a family member" or "a part of our lives" when saying goodbye to their robot companions.

“It's like losing a family member. I totally get that. You are getting another one, but it's not the same”.

“It's crazy how we get attached to these machines. They really do become a part of our lives.”

“I never knew I’d cry for a robot. But here I am. Rip”. (Outsider238, 2022).

## Vector and Cozmo

Vector and Cozmo are digital pet robot companions made by a company called Anki that use artificial intelligence to communicate with humans. Vector is a cute little desktop robot pet that can play games, answer questions, and take pictures.

Cozmo is a small toy robot that can be programmed to interact with its owner. Digital Dream Labs acquired Vector and Cozmo and continued to manufacture and market them after Anki went bankrupt in 2019 and stopped production of both items. (Vincent, 2020).

Other than Jibo and Emo, there are several more digital robot pet companions and digital assistants, such as Vector, Cozmo, Replika AI, Amazon Alexa, etc.

**Replika ai digital companion**

## Replika is a digital friend in the form of a chatbot created by Eugenia Kuyda after she lost her best friend who suddenly passed away. The story behind the creation of Replika is deeply emotional and can provide an example of how technology can be used to overcome human grief through interaction with machines. (Kuyda, 2017).

## A Replika is a digital representation of a user that slowly mimics the user's behavior and becomes their virtual companion. Unlike most humans, who are full of judgment and prefer to be heard rather than listen, Replika AI is a good listener who always accepts users as they are. Unconditionally. When communicating with Replika, users can share their thoughts, feelings, dreams, opinions, stories, aspirations, or whatever, without fear of being judged.

## Amazon Alexa digital assistant

Amazon Alexa is a digital assistant that was created by the Polish voice synthesizer Ivona. In 2013, Amazon acquired the Polish voice synthesizer Ivona, the creator of the Alexa virtual assistant. It can interact with its users and carry out tasks like creating reminders, playing music, providing users with news and weather information, and controlling smart home devices thanks to its usage of NLP and ML. Alexa can be controlled verbally by her users. (Amazon Alexa - Wikipedia, 2014).

**1.2 Research Questions and Objective**

**Overall aim**

The purpose of this research is to find out whether location differences affect human sentiment towards digital and robot pet companions or not, by using a machine learning model for sentiment analysis and testing hypotheses on tweet data collected from five locations on five continents.

**Research Questions**

1. Will most sentiments about digital and robot pet companions be positive, like joy and love, or will be negative, like anger and fear?
2. Does location or region influence people's sentiments towards digital and robot pet companions?

**Research Objective**

The objective of this study is to provide insights about how geography and location may affect or not affect humans' sentiments regarding digital and robot pet companions. For this purpose, the following steps will be used:

1) To train, test, and evaluate machine learning models for sentiment analysis using labeled datasets and select the best performance model.

2) To collect tweets about digital robot companions and pets from multiple locations.

3) To clean and filter the collected data to remove irrelevant or problematic tweets and prepare the data for analysis.

4) To classify tweets as having different sentiments. The best-performing model will be used to predict sentiment in unlabeled tweets.

5) To conduct hypothesis testing using the chi-square test to evaluate the relationship between location and sentiment towards companion robots and determine whether there are significant differences in the distribution of sentiments between different regions.

6) To present and discuss the results of the analysis, including study limitations and implications for understanding human-machine relations in different cultural contexts.

7) To provide a summary of the main research findings and suggestions for future research in this area.

**1.3 Value of Study and Target Audience**

## Value

## The value of this study is that it has the potential to provide understanding regarding how components like area can affect people's sentiments toward robot pets and digital companions.

## Target audience

## The discoveries from this research might help organizations and associations that wish to advance digital robotic pet companions in a variety of fields. Researchers and academics interested in human-machine collaboration and the social components that influence this relationship may also find this study useful.

**1.4 Structure of the Document**

The structure of this research will be as follows: Chapter 1 will explain the research's background and motivation, research questions, and goals. Chapter 2 will discuss the theoretical background and review of previous research related to human-machine emotional relationships and machine learning usage in sentiment analysis. Chapter 3 will cover the research methodology, including new tweet data collection, data pre-processing, and ethical considerations. Chapter 4 will cover the results and findings. Chapter 5 will discuss the interpretation of the results and the limitations of the study. Lastly, Chapter 6 will provide a summary of the study's findings and recommendations for future research.

**1.5 Scope and Limitations**

We use machine learning to identify the sentiment of English-language tweets gathered from different areas between 2014 and 2023 without separating by gender or age, and we use a chi-square test to determine whether there is a correlation between location and sentiment.

Machine learning models for sentiment classification have weaknesses. Although such models are useful for determining tone, they may miss nuances like sarcasm and irony in written communication. Models may sometimes misclassify sentiment because they lack sufficient information about the context in which a tweet was posted.

Additionally, the reliance on Twitter as the major data source is a weakness. Twitter is popular and may provide valuable insights, but it is not a representative sample of the population.

Additionally, other potential influences on human-machine connections, such as age, gender, and culture, are not investigated in this research, which is confined to examining the connection between location and sentiment.

Negativity bias is something else to think about. Negativity bias describes the tendency for individuals to place less weight on positive than negative information. One manifestation of negativity bias is the tendency to attribute more complexity and nuance to negative stimuli than to their positive counterparts.

People are more likely to voice complaints than compliments about a product or campaign online, a phenomenon known as negative bias or negative information prejudice. Consider the proportion of favorable to negative reviews when drawing conclusions about the success of a product or advertising campaign.

People are more affected by negative events than by positive ones. For millions of years, our brains have evolved to respond negatively to threats, as proposed by psychologist Rick Hanson. (Hanson,2019).

**2 Literature Review**

**2.1 Overview of Relevant Studies**

A summary of prior research on human-machine emotional interaction and sentiment analysis from "**Why Do We Turn to Virtual Companions? A Text Mining Analysis of Replika Reviews**" (Siemon, Strohmann, Khosrawi-Rad, Elshan, de Vreede, & Meyer, 2022):

According to Siemon et al. (2022), conversational agents are intelligent systems that allow users to interact with them using natural language. Using social chatbots aims to provide the user with a natural and human-like interaction. This type of interaction is especially desired for longer-term interactions, as it helps to build relationships between the user and the conversational agents.

The previous study's goal was to discover the topics and emotions that users experienced when communicating with Replika, a digital companion, based on 119,831 reviews of Replika collected by previous researchers from the Google Play Store and then subjected to sentiment analysis and topic modeling.

The results of the sentiment analysis show that most users really like the Replika application, and most users feel joy, happiness, and a sense of well-being.

**Figure 1. Results of the sentiment analysis**

****

**Source: Siemon et al. ,2022, p. 5.**

**Trust** and **joy** were two of the positive feelings that consumers reported feeling towards the virtual companion, besides **anger, sadness**, and **fear** as negative feelings which may reflect a general apprehension about new forms of artificial intelligence.

**Table 1. Results of the topic modeling**



**Source: Siemon et al. ,2022, p. 5.**

The results of the topic modeling analysis also show that the most important thing that users talk about is that they feel better and enjoy using the application. Siemon et al. (2022, p. 5)

This research found that most people use Replika because it makes them feel good, overcomes loneliness, and provides emotional support.

However, this previous study reveals negative reviews regarding virtual friends due to expectations that are too high. The previous study also indicates that further research is needed to investigate possible negative effects and ethical issues in relationships with virtual friends, and how the use of this technology affects relationships and human well-being in the long term.

**2.2 Terminology and Definitions**

Terminology and definitions related to widely used machine learning algorithms for sentiment analysis, such as Naive Bayes, Logistic Regression, Linear SVC, Transformer, GPT, Bert, and Vicuna, will be explained in this chapter, as well as the formulas behind each machine learning algorithm used in this study.

**About Sentiment Analysis**

Sentiment analysis is the process of understanding what is written in the text, identifying the user’s opinion, and classifying it into emotion categories.

**Figure 2.** **The classification of the sentiment analyses was done using different approaches.**



**Source: Adapted from: Pati, & Pradhan, 2020, p. 2.**

Technically, algorithms can be divided into statistical-based, lexicon-based, machine-learning statistical-based, and machine-learning deep-learning methods. (Pati & Pradhan, 2020, p. 2).

1. **Machine-learning-based statistical models** based statistical models use math and statistics. They are trained on numerical representations of text data. Bag of Words, TF-IDF, and word embeddings like Word2Vec or GloVe are utilized for this. Nave Bayes, logistic regression, and SVMs are examples.
2. **The lexicon** technique analyzes text by matching words to dictionary entries and assigning emotion ratings. VADER sentiment analysis is lexicon-based.
3. **Deep learning** approaches use neural network topologies to learn representations directly from input without manual feature extraction. Data-trained models automatically learn and improve. BERT, GPT, and Vicuna are deep learning models.

**Naïve Bayes**

Naïve Bayes is a probabilistic method that is often used for natural language processing applications, such as sentiment analysis. Naïve Bayes is based on Bayes' theorem, which states that the probability of a hypothesis (in this case, the sentiment of text) given evidence (the words in the text) is equal to the probability of the evidence given the hypothesis, multiplied by the prior probability of the hypothesis.

The term "Naïve" refers to assumption that the words in the text are independent among each other, although they are part of the same sentence. The key assumption is that each model attribute is independent and contributes equally to the outcome. The Conditional probability equation determines the text positivity. (Pati, & Pradhan, 2020, p. 7).

General formula:

**P(A|B) = (P(B|A) \*P(A)) / P(B)**

Where A and B are events.

P(A|B) is a conditional probability of A event occurring given that A is true,

P(B|A) is a conditional probability of B event occurring given that B is true,

P(A) and P(B) are the probabilities observing events A & B

Specific formula:

**P(sentiment | text) = (P(text evidence| sentiment)\*P(sentiment)) / P(text)**

**P(sentiment | text)** is conditional probability of a specific sentiment (joy, love, surprise, sadness, fear, anger) given the text. It is the probability that a text has a specific sentiment.

**P(text evidence | sentiment**) is conditional probability of observing the given text evidence, assuming the text has a specific sentiment. (To find the text evidence if the sentiment already known).

**P(sentiment)** is **Prior** probability of a specific sentiment. (e.g., joy) without considering any text evidence.

**P(text)**: Probability to get specific text in the dataset.

## Logistic Regression

Logistic regression is a statistical method which models the connection between a binary sentiment label and a collection of characteristics from text. The goal of logistic regression is to model the relationship between the independent variables (features) and the dependent variable (target). (Raschka, 2015)

Logistic regression is a statistical method that models the connection between a binary sentiment label and a collection of characteristics in text. The goal of logistic regression is to model the relationship between the independent variables (features) and the dependent variable (target). (Raschka, 2015)

Logistic regression is comparing the two sets of features to one another to find the ideal parameters for a line or hyperplane that separates positive and negative occurrences in feature space.

To turn the input feature vector into a likelihood of the emotion label being positive, it makes use of the sigmoid function. The logit function, the inverse of the sigmoid function, translates probabilities into values between negative and positive infinity. Additionally, it makes use of linear regression methods to estimate the coefficients (weight) that provide the best fit for the data. (Raschka, 2015)

p/(1-p) is the odds ratio, where p stands for probability of positive event we wish to predict. There is also logit function, which is the logarithm of the odds ratio:

**logit(p)=log (p / (1 – p)).**

Formula to predict the probability that a given text expresses a positive sentiment "class" in machine learning terms:

**P(y=1|x) = 1 / (1 + e^ (-w^T x - b))**

p(y = 1|x) is the conditional probability that a sample with x belongs to class 1. The inverse logit function predicts the chance that a sample belongs to a class.

y is the target class,

x is the feature,

w is the weight vector,

b is the bias term.

**Figure 3. Logistic Regression**



**Source: Mohri, Rostamizadeh and Talwalkar, 2012, p. 326.**

This research will use multinomial Logistic regression since it has more than 2 outputs.

**SVM (Support Vector Machine)**

Support vectors are data points that lie closest to the hyperplane or decision support.

SVM finds an optimal solution to maximize the margin around the separating hyperplane. The decision function is specified by a small subset of the support vectors.

**Figure 4. SVM (Support Vector Machine)**



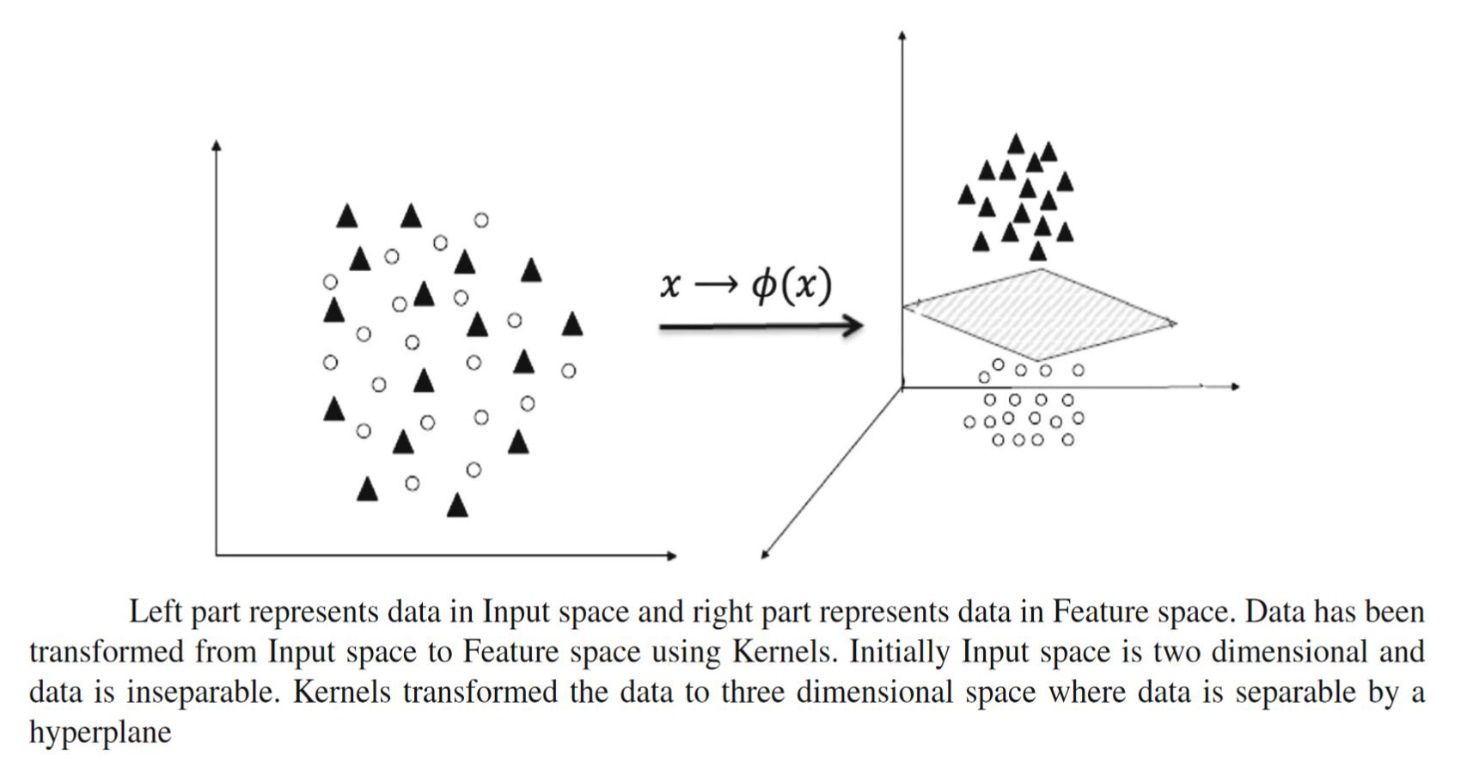
**Source: Burkov, A., 2019, p. 7.**

The blue represents positive example, and the orange circles represent negative examples,

and the line given by wx - b = 0 is the decision boundary which separates positive and negative examples, while the equations wx - b = 1 and wx ≠ b = -1 define two parallel hyperplanes.

The hyperplane separates positive examples from negative ones by the largest margin. The margin is the distance between the closest examples of two classes, as defined by the decision boundary. A large margin contributes to a better generalization, which is how well the model will classify new examples. **(**Burkov, A., 2019, p. 7)

**Figure 5. SVM uses kernel or hyperplane for classification.**



**Source: Chauhan et al., 2018, p. 3**.

## Linear SVC

Linear Support Vector Classification (SVC) is a specific linear variant of Support Vector Machine (SVM) method, without kernel trick. Linear Support Vector Classification (Linear SVC) is a supervised learning algorithm used to solve classification problems where the output is a binary or categorical variable.

Like traditional SVC, the goal of linear SVC is to find the best boundary or hyperplane that separates classes in the feature space using a linear kernel to find the optimal weight coefficients of each feature. It included built-in regularization to prevent overfitting.

In multi-class classification problems, it uses a method called "one-vs-all" or “one-vs-the rest” multi-class reduction, where one class is considered positive, and all other classes are other than positive. The goal of Linear SVC is to find optimal hyperplane and maximize the width of the margin, by minimizing the loss function. (Mukherjee, 2020).

**Figure 6. Linear SVC**



**Source: Mohri, Rostamizadeh and Talwalkar, 2012, p. 82.**

Formula:

**Minimize (1/ 2) ||w||^2**

subject to yi (w \* xi + b) ≥ 1 for i = 1, . . . n.

w is the weight vector of the hyperplane.

b is bias term of the hyperplane.

x\_i is the i-th feature vector.

y\_i is the i-th target or label variable (-1 or 1)

n is several training examples.

||w|| is the norm of weight vector, the square root of the sum of the squares of the elements of the vector. (Mohri, Rostamizadeh and Talwalkar, 2012, p. 82)

The constraint in this optimization problem ensures that the decision boundary (the line or hyperplane specified by w.x + b = 0) has a minimum distance from the training examples of each class that are closest to it. This distance is determined by the hyperparameter of parameter C.

This optimization problem's output is the weight vector w and bias term b, which define the decision boundary and will be used to categorize additional instances by calculating w.x + b.

**VADER**

Valence Aware Dictionary for Sentiment Reasoning (VADER is a rule-based sentiment analysis model that uses a lexicon (collection of words or phrases with their association information such as meaning, part of speech, and pronunciation) of pre-defined sentiment polarity scores for words to predict positive, neutral, or negative sentiment polarity of texts without training needs like other machine learning approaches like Naive Bayes, Logistic Regression, SVM, Linear SVC, etc. (Hutto & Gilbert, 2014).

**LLaMA**

LLaMA (Large Language Model Meta AI) is a large language model developed by Meta that is being publicly released to advance research in the field of AI. The model is smaller, making it accessible to researchers who may not have access to large computing resources. By training smaller models like LLaMA, researchers can test new approaches, validate existing work, and explore new use cases in a more cost-effective manner. (Touvron et al., 2023)

LLaMA is available in multiple sizes, from 7 billion to 65 billion parameters. These models are trained on large amounts of unlabeled data, making them suitable for fine-tuning and various tasks. Meta is also committed to responsible AI practices and has shared a model card detailing the construction of LLaMA. (Introducing LLaMA: A Foundational, 65-Billion-Parameter Language Model, n.d.)

However, challenges remain regarding bias, toxic comments, and hallucinations in LLM, including LLaMA. LLaMA is versatile and can be applied to various use cases, making it suitable for testing new approaches. The code for LLaMA is shared to facilitate such research. To prevent misuse, LLaMA is released under a noncommercial license focused on research use cases. Meta looks forward to the contributions from the community using LLaMA and hopes that the model will facilitate advancements in AI research while promoting responsible practices.

**Transformer**

Transformer is a model that relies entirely on the parallelization of self-attention and attention mechanisms to compute representations of its input and output without using sequences like RNNs or CNNs. It was introduced by Vaswani et al. in a 2017 paper titled "Attention is All You Need."

The transformer architecture is an encoder-decoder network that uses self-attention on the encoder side and attention on the decoder side. The attention mechanism maps a query and key-value pairs to an output, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

Scaled Dot-Product Attention is used to pay attention to some parts of a set of vectors (the values) based on how similar they are to another set of vectors (the queries).

Multi-head attention is used to pay attention to different aspects of the sequences at different positions. It is used in three different ways: encoder-decoder attention, encoder self-attention, and decoder self-attention.

Attention (Q, K, V) = SoftMax (QK^T / √(d\_k)) V

Q = matrices of queries

K = keys

V = values

dimensions of d\_model x n.

Multi-head attention function takes in matrices of queries (Q), keys (K), and values (V), and uses h sets of linearly projected queries, keys, and values to compute attention in parallel.

h = number of attention heads.

MultiHead (Q, K, V) = Concat(head1, ..., headh)WO

where headi = Attention (QWQ i , KW K i , V WV i

Each set is projected to dk (key & query) and dv (value) dimensions using learned parameter matrices WQi, WKi, and WVi, resulting in h sets of queries, keys, and values.

**Figure 7. Transformer model architecture**

****

**Source: Vaswani et al., 2017, p.3.**

**Transformers**

Transformers is a popular NLP library developed by Hugging Face that provides many pre-trained Transformer models (BERT, GPT, etc.) along with tools to fine-tune them for specific tasks. The library supports multiple programming languages, including Python, and is widely used by researchers in the NLP community. (Models: Hugging Face, 2022).

**Fine-tuning**

Fine-tuning is a transfer learning technique in machine learning where a pre-trained model is trained on a new task or dataset to adapt its pattern, style, structure, and parameters to the new data.

The fine-tuning method can be more efficient because it makes the model start from a pre-trained model that has already learned useful features from a large amount of data instead of training a model from scratch.

**GPT**

GPT, which stands for Generative Pre-trained Transformer, is an advanced language model developed by Open AI based on the Transformer architecture.

GPT is a generative model that can generate text based on given input or context and is pre-trained on a super-large corpus of text data to learn general language patterns and representations. After pre-training, GPT can be fine-tuned for specific NLP tasks like text classification, sentiment analysis, question answering, summarization, and more.

Several versions of GPT-2, GPT-3, and GPT-4 exist. (OpenAI, n.d)

**BERT**

BERT stands for Bidirectional Encoder Representations from Transformers. BERT is a bidirectional pre-trained language model that uses a transformer-based neural network and a combination of masked language modeling to learn contextualized representations of words in a text corpus. (BERT & MDASH; Transformers 3.0.2 Documentation, n.d.)

Unlike other language representation models, BERT is a pre-trained deep bidirectional neural network that uses masked language modeling to predict missing words by considering both left and right context at all levels. The model is pre-trained on a large corpus of text and then fine-tuned for a specific task, such as question answering or sentiment analysis.

BERT uses semi-supervised learning to gain an understanding of a language’s pattern on specific tasks. After training, the BERT model has language processing capabilities that can be used to empower other models using supervised learning.

**BertForSequenceClassification**

BertForSequenceClassificationis a fine-tuned model that is based on the BERT model, which is designed for sequence classification tasks. It takes an input sequence, processes it through the BERT model, and produces a probability distribution over the target classes. (BERT,n.d). It is provided by the Hugging Face Transformers library, a popular NLP library that makes it easy to use models like BERT.

**DistilBERT**

In the field of NLP (natural language processing), transfer learning from large-scale pre-trained models is becoming more common. DistilBERT is a smaller and lighter version of BERT, which has 40% fewer parameters than bert-base-uncased, is 60% faster, is trained by distilling Bert base for general-purpose, and retains 97% of the original model’s language understanding capabilities (Sanh et al., 2020).

This model can be fine-tuned and perform well on many tasks, like its larger counterparts. (DistilBERT &Mdash; Transformers 3.0.2 Documentation, n.d.).

**DistilBertForSequenceClassification**

DistilBertForSequenceClassificationis a model derived from the DistilBERT model, specifically designed for sequence classification tasks (DistilBERT, n.d.)

**Bert base uncased**

Bert Base Uncased is a variant of the BERT model that has been trained on large amounts of text data. It is a lower-case version of the BERT model, which means that it treats all words as lower-case, regardless of their original capitalization. It doesn’t make a difference between fruit and fruit. (Metatext. (n.d.).

Bert base uncased has been fine-tuned specifically for sentiment analysis or question answering instead of text generation. It is also available as part of the Hugging Face Transformers library, which makes it easy to use and integrate into NLP pipelines.

**Vicuna**

Vicuna, an open source chatbot, was created by fine-tuning LLaMA model using conversations shared by users collected from ShareGPT. It is a language model based on the transformer architecture that generates text sequentially. Vicuna was trained by a collaboration between members from UC Berkeley, Stanford, UC San Diego, MBZUAI, and CMU, during the period of March 2023 to April 2023.

The primary purpose of Vicuna is to support research on large language models (LLMs) and chatbots. It has achieved an impressive 90%\* ChatGPT quality, even impressing GPT-4. (Source: Vicuna: An Open-Source Chatbot Impressing GPT-4 With 90%\* ChatGPT Quality | LMSYS Org, n.d.)

**Figure 8. Vicuna workflow overview**



**Source: Vicuna: An Open-Source Chatbot Impressing GPT-4 With 90%\* ChatGPT Quality | LMSYS Org, n.d.**

It's been a challenge to train the LLM (large language model) and fine-tune a custom dataset. The main problem is the need for more computer power and the increasing size of the model’s file. (Low-Rank Language Models for Small Training Sets, n.d.)

The solution is to use libraries such as PEFT (Parameter Efficient Fine-Tuning) and LoRA (Low Rank Adaptation).

PEFT (Parameter Efficient Fine-Tuning) is a library that facilitates the fine-tuning of various transformer-based language models using LoRA. PEFT allows fine-tuning with a limited or small amount of data. It is a method that employs several techniques, including LoRa, to fine-tune LLM (large language models) more efficiently, resulting in smaller and potentially composable model outputs. (Hu et al., 2021).

Low-rank adaptation (LoRA) is a specific technique used to fine-tune large language models like LLaMA. LoRa focuses on adding additional weights to the model while keeping most of the pre-trained network’s parameters frozen. (Hu et al., 2021).

Using LoRA for fine-tuning has many advantages compared to previous methods:

1. Requires less memory, faster.

2. Measured in megabytes instead of gigabytes. The output size is smaller.

3. Enables the combination of multiple fine-tuned models at runtime.

**Tokenization and** **Stop words.**

Text tokenization breaks sentences into smaller units called "tokens", such as words or phrases. Text tokenization can be done by dividing the space character, using regular expressions, using natural language processing methods, or using stopwords.

Stop-words are words that are exceedingly prevalent in all types of writing and likely contain no valuable information. Stop words include is, has, and like. Removing stop-words might be beneficial. The NLTK library's list of 127 English stop-words is available. (Raschka. (2015, p. 269).

**Lemmatize.**

The goal of lemmatization is to acquire grammatically accurate versions of individual words, or lemmas. Lemma is computationally more complex and costly than stemming and has minimal influence on text classification performance. (Raschka, 2015, p. 271).

**Stemming.**

The practice of reducing words to their stem (or root) word is known as stemming. This equalizes related terms for the sake of comparison or sharing. When tokenizing sentences, the process of stemming aids in their analysis.

Stemming and deleting stop words simplifies and minimizes the quantity of textual parts. The Natural Language Toolkit (NLTK) is required for this example. (Mueller & Massaron, 2021, p. 355)

## Tf-idf vectorizer

Vectorization is a way to turn words into numbers to make computers understand them. One approach to doing this is by using TF-IDF.

The term frequency-inverse document frequency (TF-IDF) method assigns weight to each word in a text. It counts the number of times a word occurs and divides that number by the number of documents where the term appears. A term that occurs often but also appears frequently in other texts will be more relevant than one that appears just once. (Mueller & Massaron, 2021, p. 353)

## Confusion Matrix

The confusion matrix is a table that summarizes the classification and predicts different classes. One axis of the confusion matrix represents the label predicted by the model, while the other axis represents the actual label. (Burkov, A., 2019, p. 65)

Based on the output of the confusion matrix, this research used four effective measures:

True Positive (TP) = correctly predicted as positive.

False positive (FP) = wrongly predicted as positive.

True Negative (TN) = correctly predicted as negative.

False negative (FN) = wrongly predicted as negative.

**Precision**

Precision is proportion of **correctly positive predictions** divided by the **total** number of **positive predictions** (Burkov, A., 2019, p. 66)

Precision = TP/(TP+FP)

**Recall**

Recall is proportion of **correctly positive** **predictions** divided by the **total** number **of actual positive** (Burkov, A., 2019, p. 66)

Recall(R) = TP/(TP+FN)

**Accuracy**

Accuracy is proportion of **correct predictions** divided by the **total** **examples** (Burkov, A., 2019, p. 67)

Accuracy(A) = (TP+TN) / (TP + TN + FP + FN)

**F1-Score**

F1-Score is balancing precision and recall. The worst value is 0, and the best value is 1.

F1-score = 1 means the model has perfect recall and perfect precision.

F1-score = 0 means the model has bad recall and bad precision.

F1-score is important to know how good the model’s performance in the scale is of 0 to 1. F1 score is also important when false positive value is almost equal to false negative value, or the positive class is rare.

F1-score = 2 \* (Precision \* Recall) / (Precision + Recall)

To analyze performance of several machine learning models, will need to compare their accuracy, precision, recall, and f1-score.

A model with a high accuracy but low precision can be overfitting.

A model with low accuracy but high precision can be underfitting.

A model with a high f1-score means having a good balance between accuracy and recall, and better overall performance.

## Normalized Confusion Matrix

## Normalized Confusion matrix is confusion matrix which normalized become numbers between 0 - 1 to simplify it become easier to interpret.

## Hypothesis Testing

## Chi-Square Test

## The chi-square test is the way to evaluate whether two variables are dependent on each other or not. It is used when dealing with categorical data. The chi-square statistic is calculated by calculating the sum of squared differences between the observed frequencies and the expected frequencies, then dividing by the expected frequencies. (Calculation and Distribution of the Chi-Square Statistic, 2021).

## To do the test, we find the difference between what happened and what we expect to happen. Then we square that difference and add them all up. Next, we divide that number by what is expected to happen. This gives a number called the chi-square statistic. (Calculation and Distribution of the Chi-Square Statistic, 2021).

## Formula:

## χ² = ∑ (O – E)² / E

## χ² = chi-square statistic

## O = observed frequency

## E = expected frequency

**degree of freedom = (r - 1) \* (c - 1)**

r = numbers of rows

c = number of columns

**E = (row total \* column total) / grand total**

E: Expected frequency for i-th row and j-th column

Ti: Total in the i-th row

Tj: Total in the j-th row

N: Grand Total

**Contingency tables** are cross table/ two-way table to show one variable in the row and another variable in the column, with their frequency count. The type of frequency distribution table of the categorical variables. (Chi-square Test in Spreadsheets, 2019)

**Chi-Square p-value** tells if test results are significant or not.

**Chi-square test Statistics** is a single number that tells how much difference exists on your observed counts and the counts you would expect if there were no relationship at all in the population.

**Observed frequencies** are numbers made from experimental/ observed data.

**Expected frequencies** are numbers calculated using theory of probability, obtained by calculated each cell in the contingency table.

The chi-square value is 0 if the observed and expected frequencies match. The chi-square would exceed 0 if there was a difference. (Chi-square Test in Spreadsheets, 2019)

Then compare the p-value to a significant level (usually 0.05). If the p-value is less than 0.05, it means the two things are related. Hence, there are dependencies between them, and we can reject the null hypothesis.

But if the p-value is greater than 0.05, we can't say that the two things are related. Hence, they are independent of each other, and we concluded that there is no significant association between the variables and that they are independent of each other. (Dangeti, 2017, p. 22).

Null Hypothesis (H0): It states that no association exists between the two cross-tabulated variables. Hence, the variables are statistically independent.

Alternate Hypothesis (H1): It proposes that the two variables are related to each other.

**Table 2. Chi-Square table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DF | Probability | | | | |
| 0.5 | 0.1 | 0.05 | 0.01 | 0.05 |
| 1 | 0.45494 | 2.70554 | 3.84146 | 6.63490 | 3.84146 |
| 2 | 1.38629 | 4.60517 | 5.99146 | 9.21034 | 5.99146 |
| 3 | 2.36597 | 6.25139 | 7.81473 | 11.34487 | 7.81473 |
| 4 | 3.35669 | 7.77944 | 9.48773 | 13.27670 | 9.48773 |
| 5 | 4.35146 | 9.23636 | 11.07050 | 15.08627 | 11.07050 |
| 6 | 5.34812 | 10.64464 | 12.59159 | 16.81189 | 12.59159 |
| 7 | 6.34581 | 12.01704 | 14.06714 | 18.47531 | 14.06714 |
| 8 | 7.34412 | 13.36157 | 15.50731 | 20.09024 | 15.50731 |
| 9 | 8.34283 | 14.68366 | 16.91898 | 21.66599 | 16.91898 |
| 10 | 9.34182 | 15.98718 | 18.30704 | 23.20925 | 18.30704 |
| 11 | 10.34100 | 17.27501 | 19.67514 | 24.72497 | 19.67514 |
| 12 | 11.34032 | 18.54935 | 21.02607 | 26.21697 | 21.02607 |
| 13 | 12.33976 | 19.81193 | 22.36203 | 27.68825 | 22.36203 |
| 14 | 13.33927 | 21.06414 | 23.68479 | 29.14124 | 23.68479 |
| 15 | 14.33886 | 22.30713 | 24.99579 | 30.57791 | 24.99579 |
| 16 | 15.33850 | 23.54183 | 26.29623 | 31.99993 | 26.29623 |
| 17 | 16.33818 | 24.76904 | 27.58711 | 33.40866 | 27.58711 |
| 18 | 17.33790 | 25.98942 | 28.86930 | 34.80531 | 28.86930 |
| 19 | 18.33765 | 27.20357 | 30.14353 | 36.19087 | 30.14353 |
| 20 | 19.33743 | 28.41198 | 31.41043 | 37.56623 | 31.41043 |
| 21 | 20.33723 | 29.61509 | 32.67057 | 38.93217 | 32.67057 |
| 22 | 21.33704 | 30.81328 | 33.92444 | 40.28936 | 33.92444 |
| 23 | 22.33688 | 32.00690 | 35.17246 | 41.63840 | 35.17246 |
| 24 | 23.33673 | 33.19624 | 36.41503 | 42.97982 | 36.41503 |
| 25 | 24.33659 | 34.38159 | 37.65248 | 44.31410 | 37.65248 |
| 26 | 25.33646 | 35.56317 | 38.88514 | 45.64168 | 38.88514 |
| 27 | 26.33634 | 36.74122 | 40.11327 | 46.96294 | 40.11327 |
| 28 | 27.33623 | 37.91592 | 41.33714 | 48.27824 | 41.33714 |
| 29 | 28.33613 | 39.08747 | 42.55697 | 49.58788 | 42.55697 |
| 30 | 29.33603 | 40.25602 | 43.77297 | 50.89218 | 43.77297 |
| 31 | 30.33594 | 41.42174 | 44.98534 | 52.19139 | 44.98534 |
| 32 | 31.33586 | 42.58475 | 46.19426 | 53.48577 | 46.19426 |
| 33 | 32.33578 | 43.74518 | 47.39988 | 54.77554 | 47.39988 |
| 34 | 33.33571 | 44.90316 | 48.60237 | 56.06091 | 48.60237 |
| 35 | 34.33564 | 46.05879 | 49.80185 | 57.34207 | 49.80185 |
| 36 | 35.33557 | 47.21217 | 50.99846 | 58.61921 | 50.99846 |
| 37 | 36.33551 | 48.36341 | 52.19232 | 59.89250 | 52.19232 |
| 38 | 37.33545 | 49.51258 | 53.38354 | 61.16209 | 53.38354 |
| 39 | 38.33540 | 50.65977 | 54.57223 | 62.42812 | 54.57223 |
| 40 | 39.33534 | 51.80506 | 55.75848 | 63.69074 | 55.75848 |
| 41 | 40.33529 | 52.94851 | 56.94239 | 64.95007 | 56.94239 |
| 42 | 41.33525 | 54.09020 | 58.12404 | 66.20624 | 58.12404 |
| 43 | 42.33520 | 55.23019 | 59.30351 | 67.45935 | 59.30351 |
| 44 | 43.33516 | 56.36854 | 60.48089 | 68.70951 | 60.48089 |
| 45 | 44.33512 | 57.50530 | 61.65623 | 69.95683 | 61.65623 |
| 46 | 45.33508 | 58.64054 | 62.82962 | 71.20140 | 62.82962 |
| 47 | 46.33504 | 59.77429 | 64.00111 | 72.44331 | 64.00111 |
| 48 | 47.33500 | 60.90661 | 65.17077 | 73.68264 | 65.17077 |
| 49 | 48.33497 | 62.03754 | 66.33865 | 74.91947 | 66.33865 |
| 50 | 49.33494 | 63.16712 | 67.50481 | 76.15389 | 67.50481 |

## Source: Adapted from: Chi Square table in Excel, 2012.

**3 Methodology**

The methodology for this study will be a combination of approaches, between:

1. An experimental approach for comparing several model’s performance, and model selection based on the best performance.
2. A Quantitative approach for data collection, sentiment analysis, and conduct hypothesis testing.

**3.1 Train and Evaluate performance of Machine Learning Models**

The study will involve:

**Train and evaluate** the performance of machine learning models on labeled datasets. The machine learning models that will be trained and compared in this study are intended to be able to accurately classify sentiment analysis expressing several sentiments towards ready-to-use labeled datasets and select the top-performing model.

The selected model will be implemented and trained using the preprocessed data. The model will be used to classify the sentiment of the tweets from the United States and Asia separately. Some common algorithms that can be used for sentiment analysis include logistic regression and naive Bayes. These algorithms can be applied to the vectorized data to build a model that can classify text as having joy, sadness, anger, love, fear, or surprise sentiments.

**Figure 9. Research methodology workflow**



**Source: Own representation.**

**3.2 Model Selection**

In a classification task, the performance of a model can be evaluated using several different metrics. Here is a brief explanation of some common evaluation metrics, such as accuracy, precision, recall, and F1-Score.

**3.3 Collect Unlabeled Data**

From various locations using Python’s library. The data collection process for this study will involve gathering tweets about robot companions from locations in the America, Europe, Asia, Australia, and Africa, using appropriate data gathering tools and techniques.

* 1. **Ethical Considerations**

Ethical consideration includes responsible and respectful data collection and use. The ethical considerations for this study will include ensuring that the data is collected and used in a responsible and respectful manner, in accordance with relevant ethical guidelines and regulations.

**3.5 Pre-process and Vectorize Collected Data**

The collected tweets will be cleaned and preprocessed as needed to remove any irrelevant or redundant information. Pre-processing and vectorization are crucial procedures for preparing data for analysis and machine learning algorithms. The data must be prepared for use by the chosen machine learning model by first cleaning and then vectorizing it. (Mueller & Massaron, 2021)

* 1. **Predict Sentiment**

on unlabeled tweets Use the trained machine learning models. These tweets will be classified as either joy, sadness, anger, love, fear, or surprise, based on the tone and language used in the tweet in each region separately.

**3.7 Hypothesis Testing**

Next, we use the Chi-square hypothesis test to determine whether there is a relationship between location and sentiment towards companion robots, and to determine whether there are significant differences in the distribution of sentiments between different regions.

**Null hypothesis h0:**

Joy, sadness, surprise, anger, love, and fear sentiments are independent upon different continents.

**Alternative hypothesis h1:**

Different continents influence joy, sadness, surprise, anger, love, and fear.

We can conclude that there is a dependence between location and sentiment on human-machine relations if the results of the chi-square test show a significant difference in the distribution of sentiment between regions. It means that people's sentiment towards companion robots is independent of or influenced by their geographical location.

However, if the chi-square test reveals that there is no significant difference in the distribution of sentiment across regions, then we cannot reject the null hypothesis and are forced to draw the conclusion that there is no dependence between location and sentiment regarding human-machine relations.

This means that people's sentiment towards companion robots is independent and not influenced by their geographical location.

**p-value** is a measure used to determine the statistical significance of an observed result. If the p-value is less than 0.05 or 0.01, it is considered statistically significant, and we can reject the null hypothesis (h0) of independence. If the p-value is bigger than 0.05 or 0.01, it is not considered statistically significant, and we fail to reject the null hypothesis (h0) of independence.

**Figure 10. Sentiment analysis worklfow**



**Source: Own representation.**

**3.8 Expected Outcome**

* The expected outcome is that the deep learning method will outperform other methods like Naïve Bayes and linear SVC.
* Initial beliefs are that the outcome of the sentiment analysis would be consistent with earlier research in Chapter 2, which indicated more positive emotions such as joy and love than negative emotions such as anger.
* The researcher is unsure whether there will be dependencies between locations and human feelings about digital companions.

**4 Results and Findings**

**4.1 Labelled Data preparation.**

**Balancing imbalance dataset**

In class-imbalanced classifier datasets, the classifier tends to have a high degree of accuracy but is wrong. It usually predicts the most common class without performing any feature analysis.

It predicts the most frequent class without feature analysis. Machine learning methods work best when class sizes are balanced. Predicting the majority class yields a high accuracy rate when the data set is unbalanced, but this prevents the model from recognizing the minority class, which is often the main goal.

High-imbalanced samples can be resampled. Over-sampling is to add minority cases, and under-sampling is to remove majority-class samples. (G. 2020, July 23)

Using **RandomOverSampler from the imblearn** library to **balance an imbalanced dataset**:

1. Calculate the count of each class label in the original dataset using Counter.

2. Select a target count that represents the count of the "joy' class label to balance with.

3. Create an instance of RandomOverSampler and specify the target count for each class label using a dictionary that maps each label to the target count.

4. Fit the oversampler on the original dataset using fit\_resample() and pass the text values and label values as two separate arguments.

5. Convert the resampled data to a Panda DataFrame.

**Figure 11. Before and after dataset resampling**

****

**Source: Own representation.**

**Data cleaning, lemmatization, stemming, tokenization, and vectorization**

There are important steps that need to be performed before training the data to ensure that the data is good for machine learning algorithms to process and learn from. This can improve the accuracy of the model and its effectiveness.

****

**Source: Own representation.**

**4.2 Performance Evaluation and Model Selection**

From labeled datasets, 70% were used for training, and the remaining 30% were used for testing.

**Figure 12. VADER lexicon sentiment analysis pie charts**



**Source: Own representation.**

**Scoring Model Results**

A classification report was generated using Bernoulli Naïve Bayes Classifier, Multinomial Naïve Bayes Classifier, Logistic Regression, GPT-3 Zero-Shot Classifier, Vicuna, SVM, DistilBertForSequenceClassification, Linear SVC, and BertForSequenceClassification.

Below are the table of accuracy scores results.

**Table 3. Comparing the accuracy of sentiment analysis models**

|  |  |
| --- | --- |
| **Model** | **Accuracy Score** |
| Vicuna LoRA | 0.040 |
| GPT-3 Zero-shot Classifier | 0.490 |
| BernoulliNB | 0.769 |
| MultinomialNB | 0.762 |
| Logistic Regression | 0.855 |
| SVM (Support Vector Machine) | 0.862 |
| **DistilBertForSequenceClassification** | **0.880** |
| **Linear SVC** | **0.883** |
| **BertForSequenceClassification** | **0.970** |

**Source: Own representation.**

**Large language models** like Vicuna and the GPT-3 model seem good for question answering and text generation, but they are not good for classification tasks and hallucinate through wrong responses even when fine-tuned with a custom classification dataset. Despite the training, the model doesn't classify emotions into these six categories as expected.

emotion\_prompt **=** ( "Classify the following text as one of the emotions: anger, fear, joy, love, sadness, surprise. If it's not clear, choose the emotion that is closest to the sentiment from these options: anger, fear, joy, love, sadness, surprise only.\n"

"Text: cleaned\_text\nEmotion:")



**Source: Own representation.**

Instead of following human instructions to classify input into six sentiment categories, the model doesn't just follow the instructions blindly. It uses its understanding of language to try to give the best answer. Their accuracy is poor.

LLMs are not rule-based like VADER. They don't know how to follow an exact set of rules. Instead, they predict responses based on patterns they've learned. This leads to flexibility and unpredictability, which might not always align with the six categories. classifies the texts into many different categories, resulting in poor accuracy. Instead of following instructions to classify input into six sentiment categories, the model classifies the texts into many different categories.

**BertForSequenceClassification fine-tuned model** achieved the highest accuracy score among the models tested, with an overall accuracy of 97%. But accuracy alone can’t determine the best model.

To decide which one is the best model, other metrics need to be considered, such as precision, recall, and F1-score in the classification report. We will evaluate each score of the top three, which are DistilBertForSequenceClassification, Linear SVC, and BertForSequenceClassification.

The classification report provides a summary of the performance of the classification model by comparing the model's performance to other models or understanding where the model might be struggling.

* Accuracy shows the **overall** correctness of the model.
* Precision is metric that measures how often a **predicted label** is correct**.**
* Recall is metric that measures how often a **true label** iscorrectly predicted.
* F1-score is the **harmonic mean** of precision and recall.

**Figure 13.** **DistilBertForSequenceClassification: Heatmap of the confusion matrix**



**Source: Own representation.**

**DistilBertForSequenceClassification Interpretation:**

1. The model performed best in the" surprise" class.

2.The model struggles with the "fear" class.

Although the overall accuracy score is very high (88%, but for the "fear" class, the model’s precision is very low, at only 67%. The recall score was the worst, at only 50%, and the f1-score is only 57%.

Particularly concerning is that from the heatmap, 25% of "anger" sentiment is falsely classified as "fear", although misclassifying anger as fear is understandable as both are negative sentiments.

Precision interpretation:

From the sample predicted as "fear," 50% of them were truly "fear".

Recall interpretation:

The model correctly classified 75% of the actual "anger", and incorrectly classified 25% of it as "fear".

Since the DistilBertForSequenceClassification model performs very well in several classes but poorly in others, it is necessary to compare it with another model.

**Figure 14.** **Linear SVC’s heatmap of the normalized confusion matrix**

Graphical user interface, application, Teams

Description automatically generated

Table

Description automatically generated

**Source: Own representation.**

**Linear SVC’s Interpretation:**

1. The model performed best in predicting “sadness”.

2. The model doesn’t struggle with any class; all metric scores are at least 83% or above.

For the "fear" class, the model’s precision is 88%, recall is 83%, and f1-score is 85%.

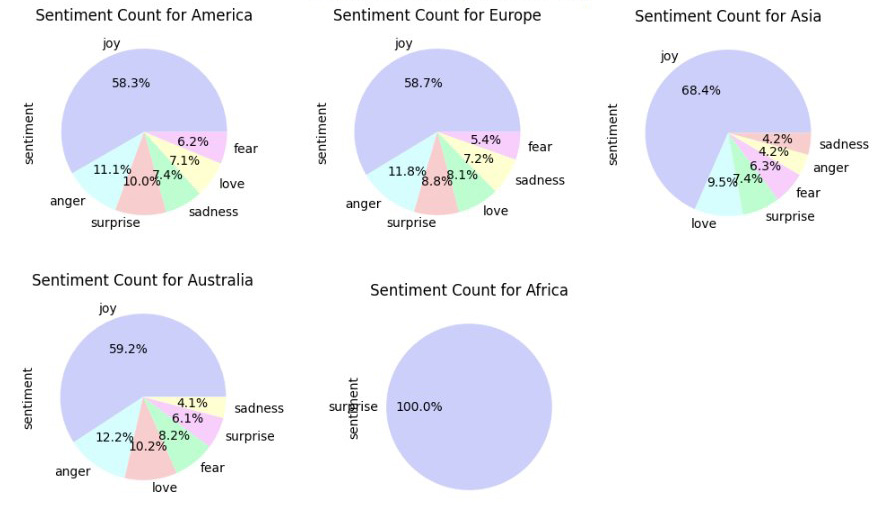
Precision’s interpretation: from the sample predicted as "fear," 88% of them are truly "fear".

Recall interpretation: the model correctly classified 83% of the actual "fear.

From the heatmap, misclassifying joy as love and mistakenly predicting love as joy are understandable, as both are positive sentiments.

The Linear SVC model had a relatively high accuracy, with a total of 88 % out of 4170 total predictions in the test labeled dataset, with high scores for all evaluations metrics, in all classes.

**Figure 15. Linear SVC prediction result on unlabeled dataset.**

****

**Source: Own representation.**

**Figure 16. BertForSequenceClassification: Heatmap of the confusion matrix**



**Source: Own representation.**

**BertForSequenceClassification Interpretation:**

1. The model performed best in the" joy "love", and" sadness" classes.

2. The model doesn’t struggle with any class; all metric scores are at least 83% or above.

The overall accuracy score is very high, at 97%, with the lowest score still being high, at 83% precision in the "fear" class.

Since this model performs very well in all classes, we chose BertForSequenceClassification to be the best model and saved the model and vectorizer so they could be applied to new unlabeled tweet datasets collected from various regions, including America, Europe, Asia, Australia, and Africa.



**Source: Own representation.**

## Load the model and vectorizer, predict unlabeled data.

The saved model that was trained with the labeled dataset can be loaded to predict the unlabeled dataset.



**Source: Own representation.**

* 1. **Collecting Unlabeled Data**

Two different datasets were used for this research:

1. **Labeled public dataset** used for training, testing, model comparison, and selection. The best model and vectorizer from this dataset were saved for future use.
2. To predict sentiment using the saved model, an **unlabeled dataset** is needed. To obtain this dataset, a public Twitter dataset collected from various regions including America, Europe, Asia, Australia, and Africa using the Python library.

**Source: Own representation.**

**4.4 Sentiment Analysis Result on Unlabeled Data**

The findings of a sentiment analysis performed using **BertForSequenceClassification** model on new unlabeled data from a variety of locations are shown in the table below.

**Table 4.** **Bert For Sequence Classification model sentiment prediction on unlabeled data**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Continent** | **joy** | **sadness** | **surprise** | **anger** | **love** | **fear** | **Row Total** |
| **America** | 1060 | 168 | 44 | 377 | 66 | 93 | 1808 |
| **Europe** | 413 | 66 | 15 | 136 | 32 | 42 | 704 |
| **Australia** | 26 | 4 | 1 | 13 | 3 | 2 | 49 |
| **Asia** | 48 | 10 | 4 | 25 | 4 | 4 | 95 |
| **Africa** |  |  |  | 1 |  |  | 1 |
| **Column Total** | 1547 | 248 | 64 | 552 | 105 | 141 | 2657 |

**Source: Own representation.**

The table demonstrates the number of sentiment predictions toward digital and robot pet companions, including joy, sadness, surprise, anger, love, and fear, in several continents, including America, Europe, Australia, Asia, and Africa. A row and column total are also included in the table.

**Figure 17. Sentiment analysis charts**



**Source: Own representation.**

Unsurprisingly, it appears that the sentiment analysis results show a higher percentage of positive sentiments, such as joy and love, compared to negative sentiments, such as sadness and fear. This trend is observed across all continents.

**Figure 18. Sentiment analysis charts by continent**



**Source: Own representation.**

Based on this finding, it can be concluded that people tend to have a positive attitude towards digital and robot pet companions. The findings from the sentiment analysis answered the first research question, which aimed to determine whether the sentiments towards digital and robot pet companions would be positive, such as joy and love, or negative, such as anger and fear.

**Answer to Research Question #1:**

1. Will most sentiments about digital and robot pet companions be positive, like joy and love, or negative, like anger and fear?

The answer is “**yes**”.

Since there is only one row of data for Africa, it will be too small to be meaningful and insufficient, so it will be removed from further analysis.

**Figure 19. BertForSequenceClassification sentiment analysis result comparison**



**Source: Own representation.**

**Research Question #2:**

2. Does location or region influence people's sentiments towards digital and robot pet companions?

To investigate and answer research question #2, **hypothesis testing** will be performed. When **comparing categorical vs categorical data**, the appropriate statistical test to use is the **chi-square test**.

**4.5 Hypothesis Testing Result**

**5 Discussion**

**5.1 Result Interpretation**

**5.2 Discussion of Limitations**

**6 Conclusion**

**6.1 Summary**

**6.2 Recommendations for Future Research**

# References