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: 21st July 2023

Acknowledgement

First, 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.

Also, 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.

Moreover, 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.

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 emotion 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 give 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

List of Figures

List of Tables

List of Formulas/ Equations

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 anyone have a dream that one day artificial intelligence (AI) and humans will unite, then one of the most effective ways to achieve this goal 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 the 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 them.

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 usage 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 life. 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 1000 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 is 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 & smile. I will miss him so much. I hope he can be repaired…and 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 company called Anki, which 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 which 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 & love, or will be negative, like anger & fear?
2. Does location/ 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, evaluate machine learning model 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 different sentiments. The selected best performing model will be used to predict sentiment in Unlabeled tweets.

5) To conduct a hypothesis testing using 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 give understanding regarding how component like area can affect people sentiments of robot pet 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 & 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 tweets data collection, data pre-processing, and ethical considerations. Chapter 4 will cover the results and findings. Chapter 5 will discuss interpretation of results and 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 has 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 good 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 prone 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

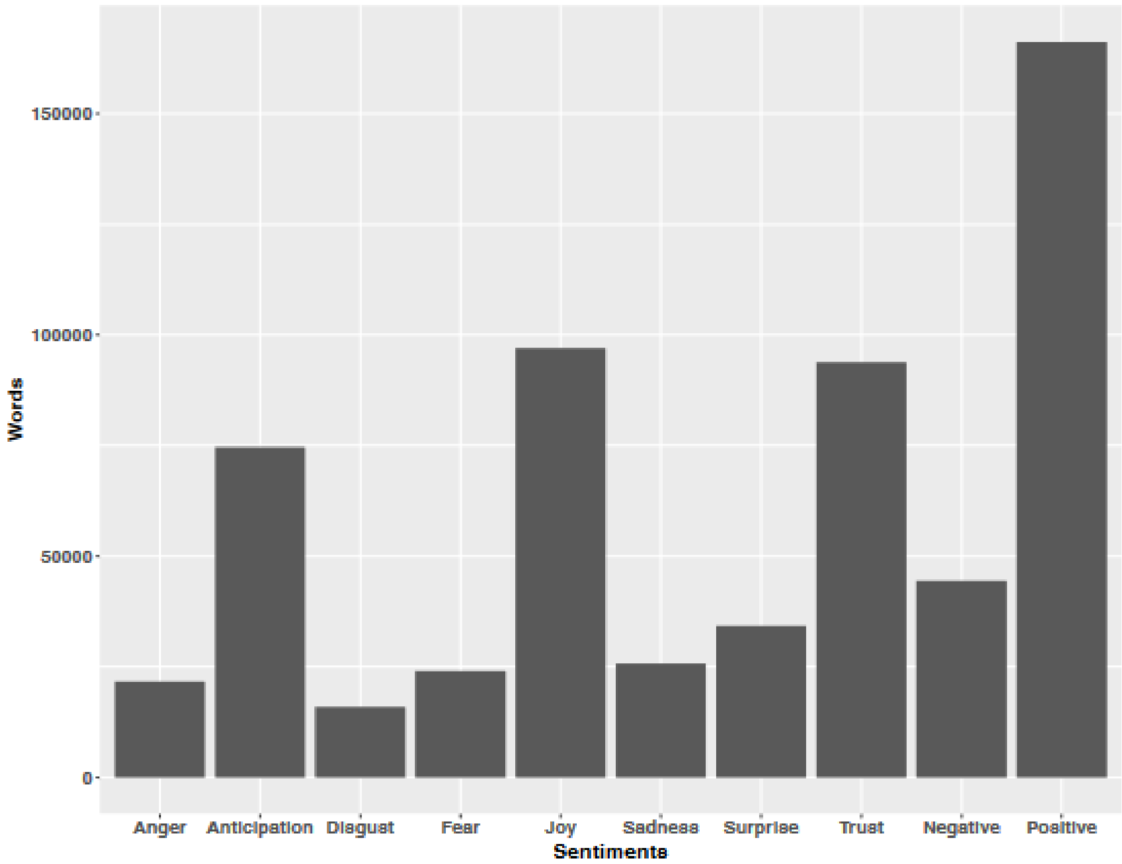
A summary of prior research on human-machine emotional interaction and sentiment analysis from "**Why Do We Turn to Virtual Companions? A Text Mining Analysis of Replika Reviews**" (Siemon, Strohmann, Khosrawi-Rad, Elshan, de Vreede, & Meyer, 2022):

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

Topic name	**Topic 1 – Excitement about AI**	**Topic 2 - Companionship**	**Topic 3 – Well-Being and Support**	**Topic 4 – Enjoyment**
Relevant words	AI Good Fun Pretty Cool Interesting Conversation	Like Real Talking Person Someone Friend Nice	Asked Great Feel Good Helpful Better Talk	Amazing App Love Good Awesome Fun AI

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.

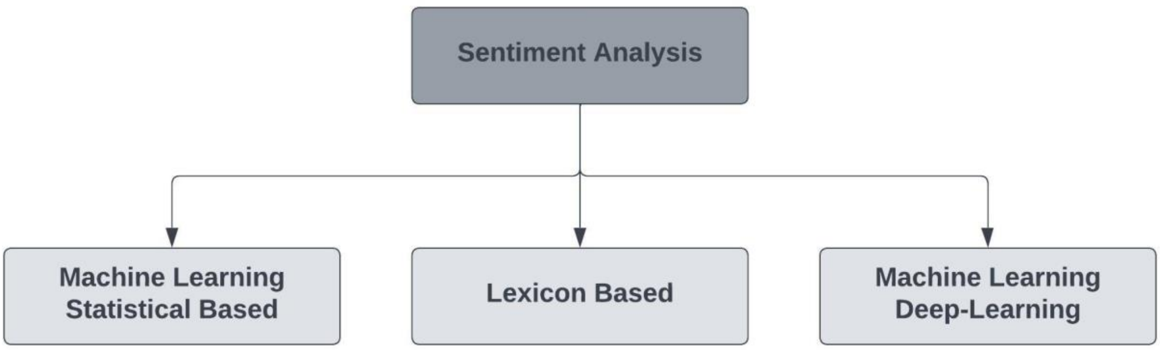
2.2 Terminology and Definitions

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 formulas behind each machine Learning algorithms used in this study.

About Sentiment Analysis

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

Figure 2. The classification of the sentiment analyses 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 dMachine learning eep-learning method. (Pati, & Pradhan, 2020, p. 2).

1. **Machine Learning-based statistical models** rely on mathematical or statistical analysis. They are trained on text data transformed into numerical representation that the algorithm can process. Techniques such as Bag of Words, TF-IDF, and word embeddings (like Word2Vec or GloVe) are commonly used for this purpose. Examples of machine learning models include Naïve Bayes, Logistic Regression, and Support Vector Machines (SVMs).
2. **The lexicon method** uses a pre-compiled dictionary or set of rules, to analyze text and extract sentiments from the text, matching the words in text with entries in the dictionary and assigning sentiment scores. An example of a lexicon-based method is the VADER sentiment analysis tool.
3. **Machine learning deep learning-based methods** can learn directly from data, due to their complex architecture. They use neural network architectures and learn representations directly from the data, and often without the need for manual feature extraction. The models are trained using tons of data and algorithms such as backpropagation. This process allows the model to automatically learn and improve from experience. Examples of deep learning models include BERT, GPT, and Vicuna.

Naïve Bayes

Naïve Bayes is a probabilistic method that is often used for natural language processing applications, such as sentiment analysis. It 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 features (words in the text) are independent of one another, 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 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, inverse of the sigmoid function, translates probabilities to 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.

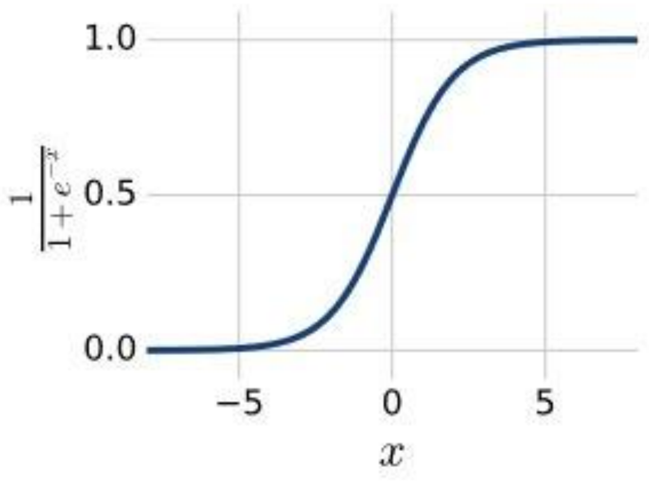
y is the target class,

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.

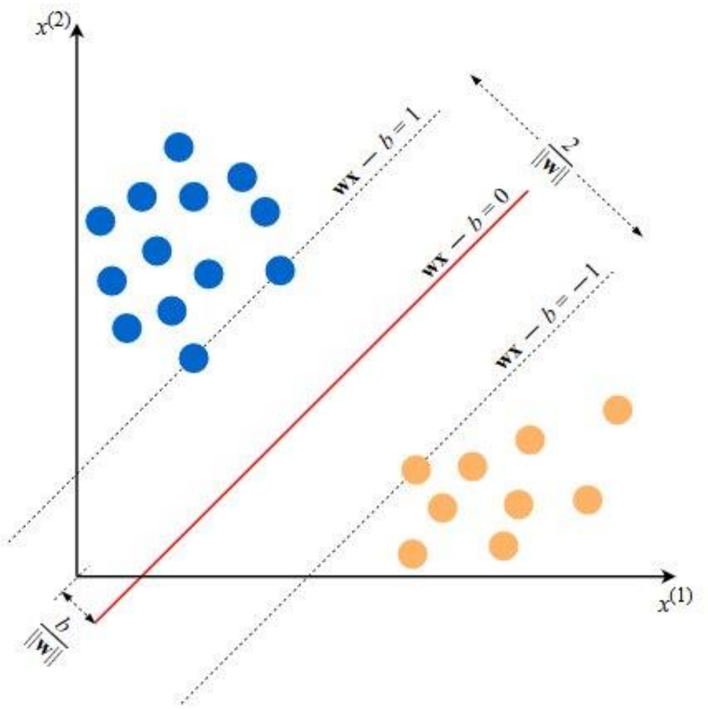
This research will use multinomial Logistic regression since it has more than 2 outputs.

SVM (Support Vector Machine)

Support Vectors are data points that lie closest to the hyperplane / decision support.

SVM finds an optimal solution to maximize the margin around separating hyperplane. The decision function 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 with 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 good the model will classify new examples. **(**Burkov, A., 2019, p. 7)

Linear SVM

Linear SVM is a type of SVM using linear kernel/ hyperplane for binary classification.

Figure 5. Linear SVM hyperplanes

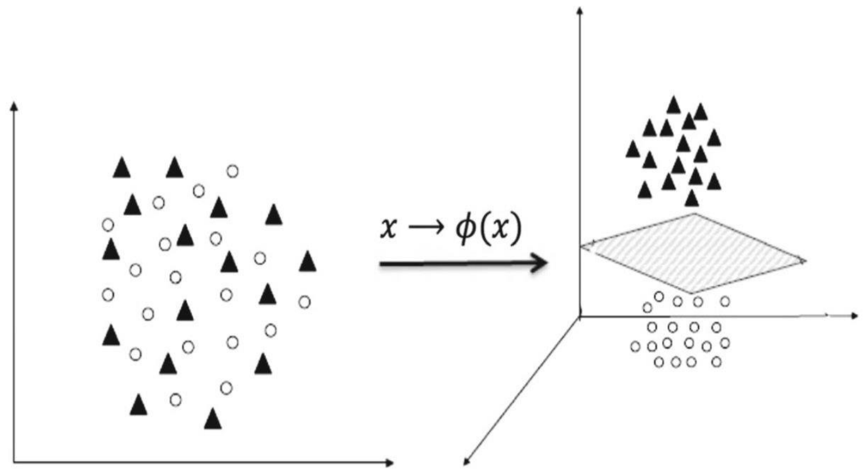


Fig. 2 Left part represents data in Input space and right part represents data in Feature space. Data has been transformed from Input space to Feature space using Kernels. Initially Input space is two dimensional and data is inseparable. Kernels transformed the data to three dimensional space where data is separable by a hyperplane

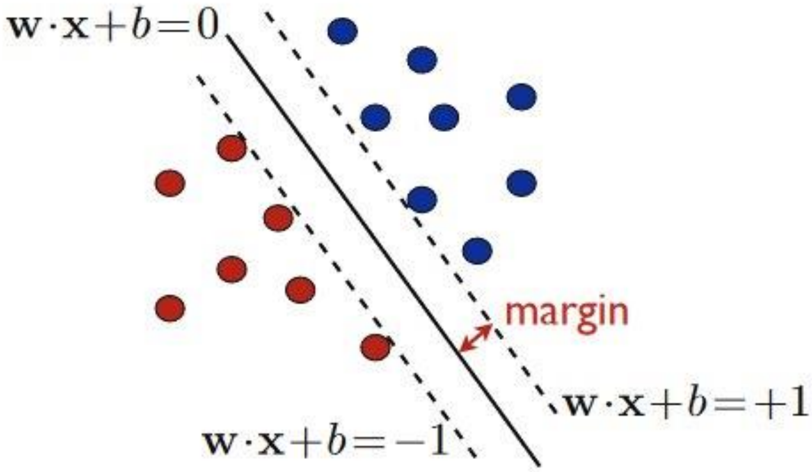
Source: Chauhan et al., 2018, p. 3.

Linear SVC

Linear Support Vector Classification (SVC) is a specific linear variant of Support Vector Machine (SVM) method. 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 which uses a lexicon (collection of words or phrases with their association information such as meaning, part of speech, 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 Naïve 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, which 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 it 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 model that relies entirely on parallelization of self-attention and attention mechanism to compute representations of its input and output without using sequence like RNNs or CNNs, introduced by Vaswani et al. in a paper titled "**Attention is All You Need**" in 2017.

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.

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T / \sqrt{d_k})\, V$$

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.

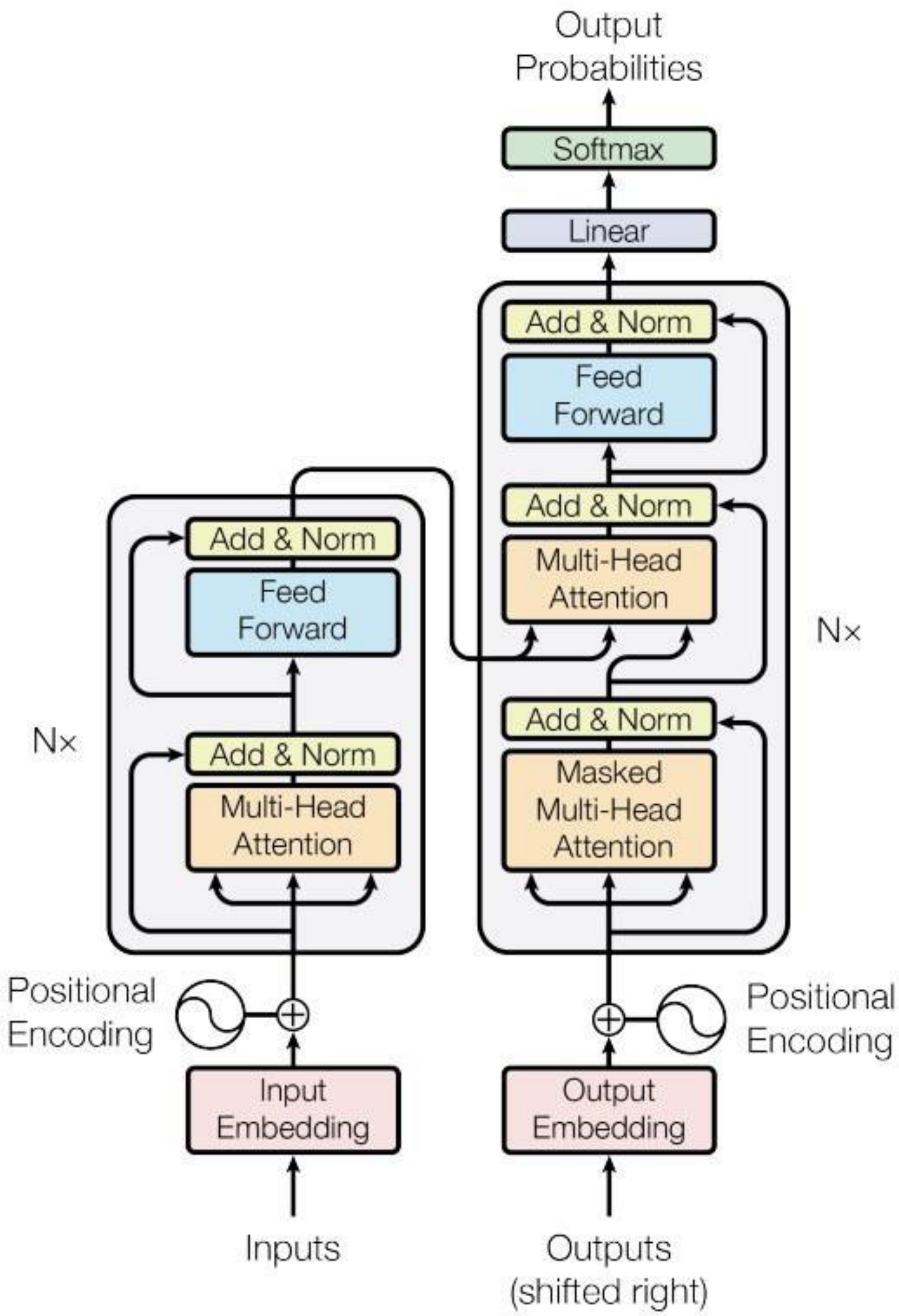
h = number of attention heads.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, ..., head_h)W^O$$

$$\text{where } head_i = \text{Attention}(QW^Q_i, KW^K_i, VW^V_i$$

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 which provides many pre-trained Transformer models (BERT, GPT, etc.) beside 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.

Fine-tuning method can be more efficient, because it makes the model start from pre-trained model which 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 which can generate text based on given input or context, 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. (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 using 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 a 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 on a specific task, such as question answering or sentiment analysis.

BERT uses Semi-Supervised Learning to gain understanding of 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

BertForSequenceClassification is a fine-tuned model that is based on the BERT model, which 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 model is becoming more common. DistilBERT is a smaller/ lighter version of BERT which has 40% less parameters than bert-base-uncased, 60% faster, trained by distilling Bert base, for general-purpose, and retains 97% of 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

DistilBertForSequenceClassification is a model derived from 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.).

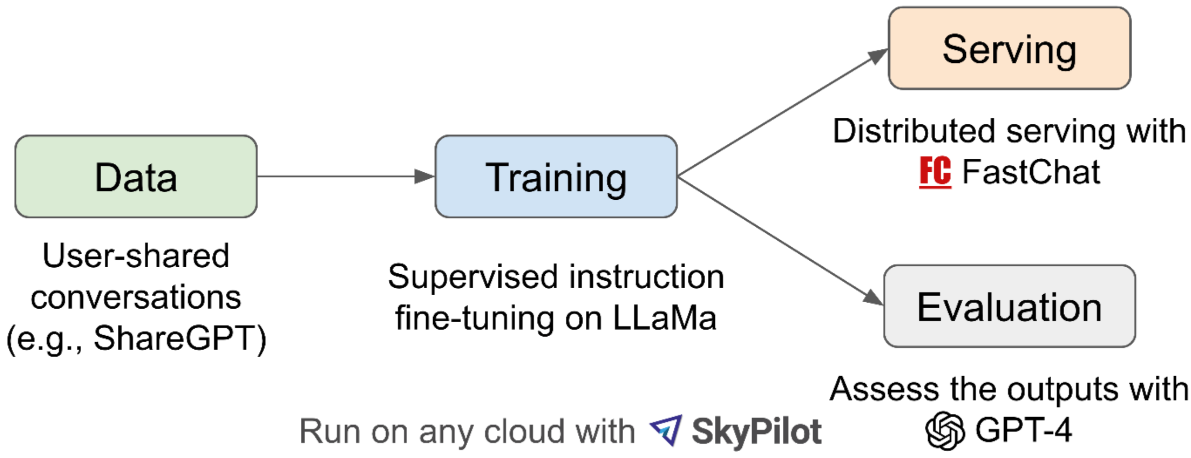
Bert base uncased has been fine-tuned specifically for sentiment analysis or question answering, instead of text-generation. It is also available as part of the Hugging Face Transformers library, which makes it easy to use and integrate into NLP pipelines.

Vicuna

Vicuna is an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT. It is an auto-regressive language model, based on the transformer architecture. Vicuna was trained between March 2023 and April 2023 by members from UC Berkeley, Stanford UC San Diego, MBZUAI, and CMU.

The main use of Vicuna is research on LLM (large language models) and chatbots. (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 LLM/Large Language Model and fine tuning a custom dataset. The main problem is because of the need for more compute power, and the increasing of model's file size. (Low Rank Language Models for Small Training Sets, n.d.)

The solution is with using library 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 enables fine-tuning with a small amount of data. It is a method that employs several techniques, including LoRa, to fine-tune LLM (large language models) more efficiently. This approach results in smaller and potentially composable model outputs. (Hu et al., 2021).

Low-rank adaptation (LoRA) is a technique used to fine-tune large language models like LLaMA. LoRa focuses on adding extra weights to the model while freezing most of the pre-trained network's parameters. (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, or using natural language processing methods, or using stopwords.

Stop-words are words that are exceedingly prevalent in all types of writings 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 do 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 to predict 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 confusion matrix output, 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

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:

$$\chi^2 = \sum (O - E)^2 / E$$

χ^2 = 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 observed and expected frequencies matched. Chi-square would exceed 0 if there is 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 each other, and we can reject the null hypothesis.

But, if the p-value is more than 0.05, we can't say that the two things are related. Hence, they are independent of each other, and 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.

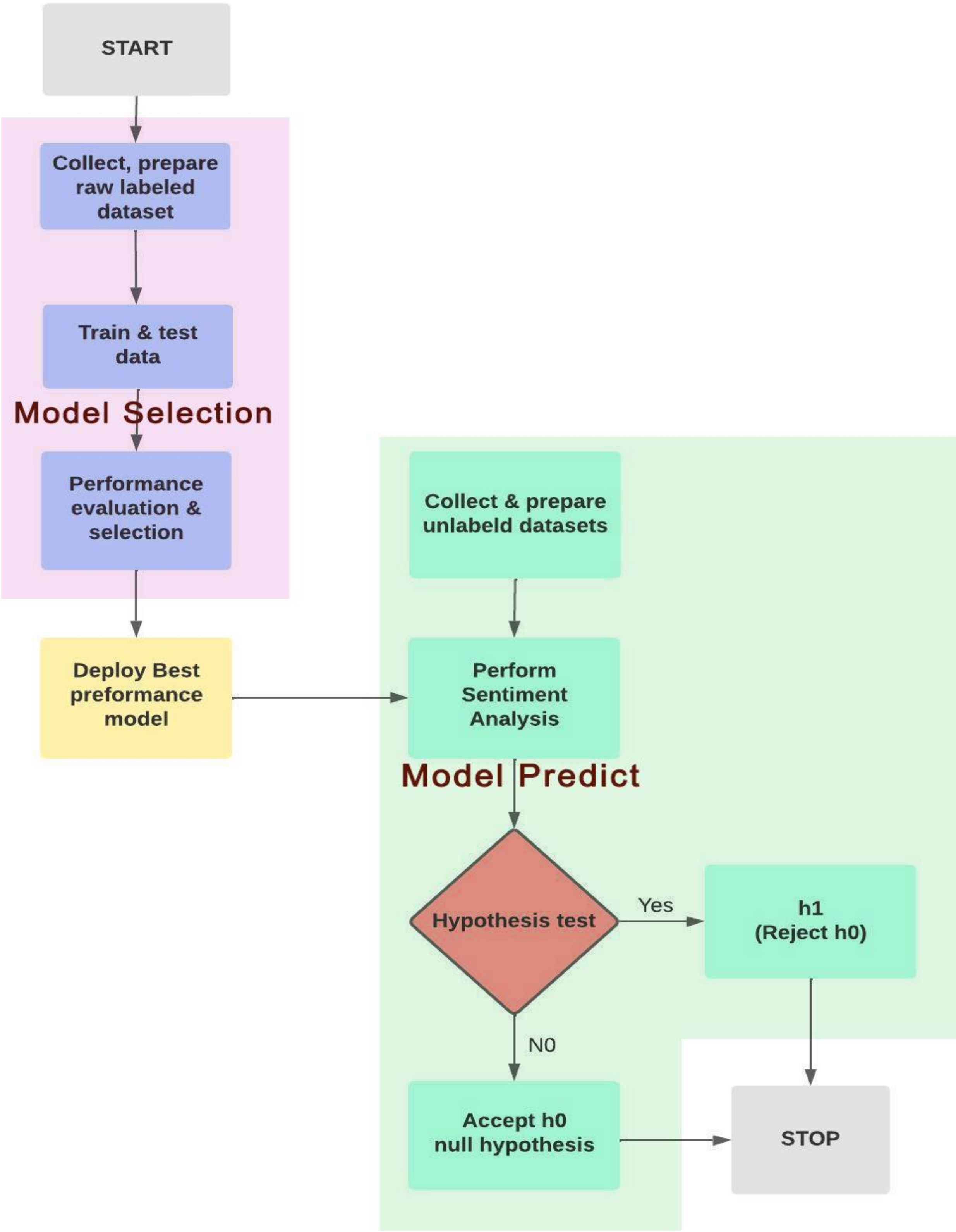
3.1 Train and Evaluate performance of Machine Learning Models

The study will involve:

Train and Evaluate performance of Machine Learning Models on labeled datasets. The machine learning models 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 **joy, sad, anger, love, fear, 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.

3.4 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 usage by the chosen machine learning model by first cleaning and then vectorizing them. (Mueller & Massaron, 2021)

3.6 Predict Sentiment

on Unlabeled tweets Use the trained machine learning models. These tweets will be classified as either joy, sadness, anger, love, fear, 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, Fear sentiments are INDEPENDENT upon different continents.

Alternative hypothesis h1:

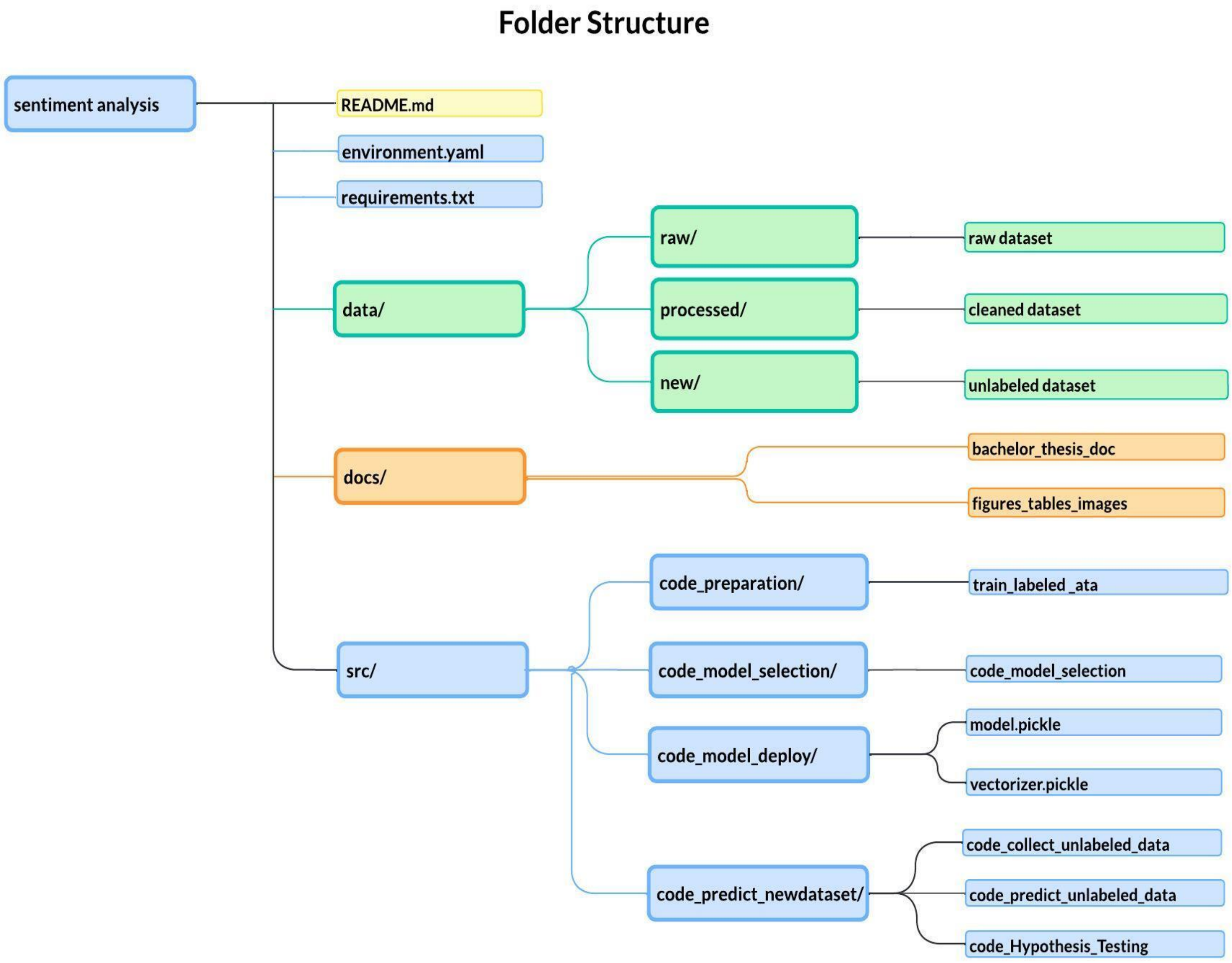
Joy, Sadness, Surprise, Anger, Love, Fear sentiments are DEPENDENT upon different continents.

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 depending on/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, not influenced by their geographical location.

Figure 10. Sentiment analysis worklfow



Source: Own representation.

3.8 Expected Outcome

- The expected outcome is that logistic regression method will outperform other methods like Bayes and Linear SVC.
- Initial beliefs are that the sentiment analysis outcome would be consistent with earlier research in chapter two, 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 will have a high degree of accuracy, but 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 balanced. Predicting the majority class yields a high accuracy rate when the data set 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 adds minority cases, and under-sampling is to removes 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 ()

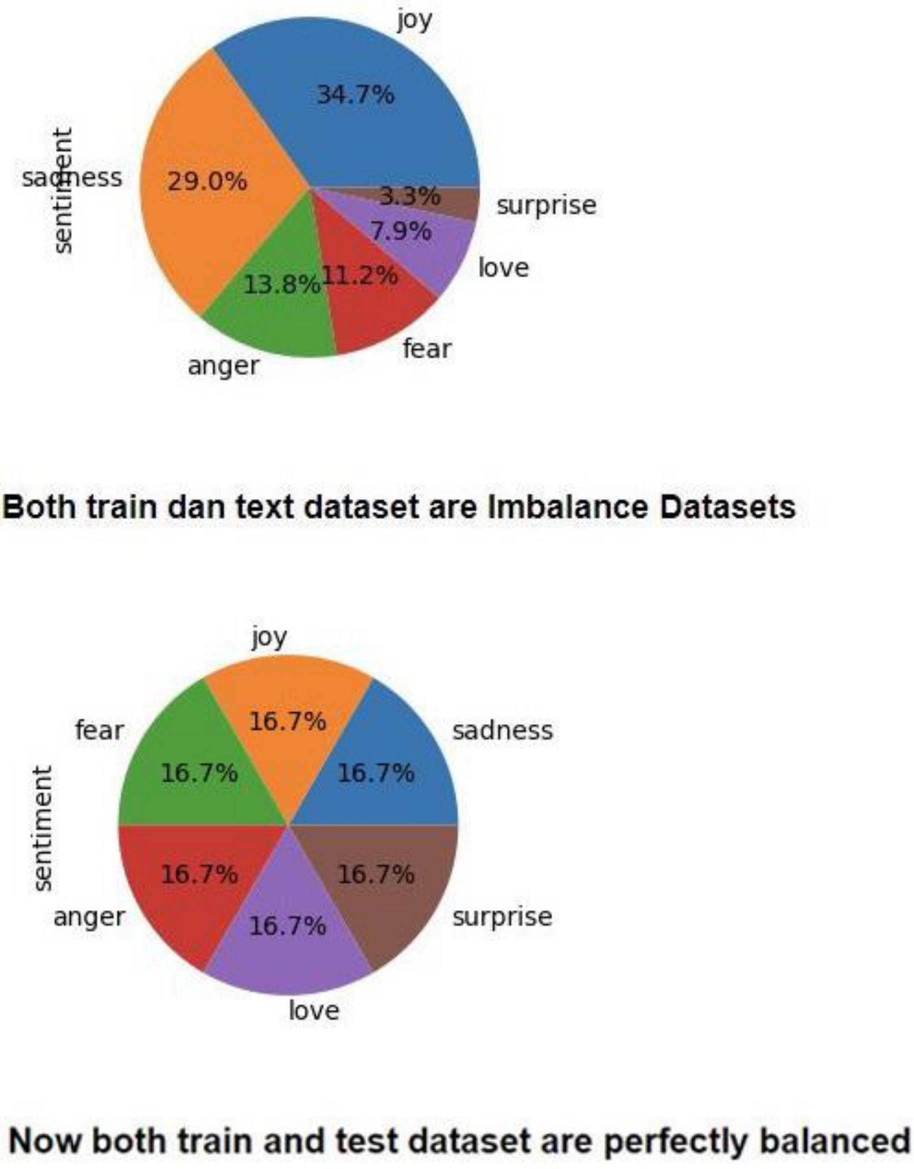
2.Select a target count that represents the count of 'joy' class label to balance with.

3.Create an instance of RandomOverSampler & 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 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 and model's effectiveness.

Cleaning

```python
def clean_text(text):
    import re
    from string import punctuation
    text=re.sub(r'(http|ftp|https):\/\/([\w\-_]+(?:(?:\.[\w\-_]+)+))([\w\-\.,@?^=%&:/~\+#]*[\w\-\@?^=%&/~\+#])?',
                ' ', text)
    text=re.sub(r'['+punctuation+']',' ',text)
    text=re.sub(r'#(\w+)',' ',text)
    text=re.sub(r'@(\w+)',' ',text)
    text = text.lower() # Convert  to Lowercase

    token=RegexpTokenizer(r'\w+')
    tokens = token.tokenize(text)

    lemmatizer = WordNetLemmatizer()
    stems = [lemmatizer.lemmatize(t) for t in tokens]
    stemmer = PorterStemmer()
    stems = [stemmer.stem(t) for t in stems]

    return ' '.join(stems)

def tokenize(text):
    token=RegexpTokenizer(r'\w+')
    tokens = token.tokenize(text)
    return tokens
```

Vectorizer tf-idf

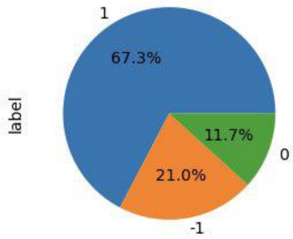
```python
# Define the vectorizer and fit it to the training data
cv = TfidfVectorizer(lowercase=True, preprocessor=clean_text, stop_words='english',
                             ngram_range=(1,3), tokenizer=tokenize)

x_train = cv.fit_transform(train_df['text'].values.astype('U'))
y_train = train_df['sentiment']

# Vectorize the test data using the same vectorizer
x_test = cv.transform(test_df['text'].values.astype('U'))
y_test = test_df['sentiment']
```

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.

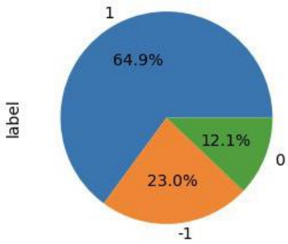
Figure 12. VADER lexicon sentiment analysis pie charts



America

```
1 sentiment_count_J = dfP_Am1['label'].value_counts()
2 sentiment_count_J
```

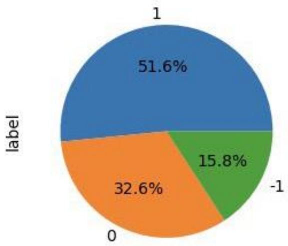
```
positive    1217
negative     379
neutral      212
Name: label, dtype: int64
```



Europe

```
1 sentiment_count_J = dfP_Eu1['label'].value_counts()
2 sentiment_count_J
```

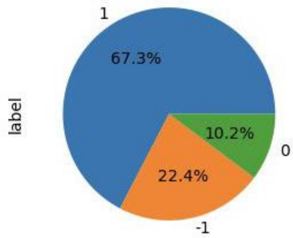
```
positive    457
negative    162
neutral      85
Name: label, dtype: int64
```



Asia

```
1 sentiment_count_J = dfP_As1['label'].value_counts()
2 sentiment_count_J
```

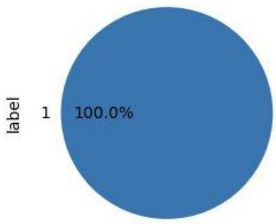
```
positive    49
neutral     31
negative    15
Name: label, dtype: int64
```



Australia

```
1 sentiment_count_J = dfP_Au1['label'].value_counts()
2 sentiment_count_J
```

```
positive    33
negative    11
neutral      5
Name: label, dtype: int64
```



Africa

```
1 sentiment_count_J = dfP_Af1['label'].value_counts()
2 sentiment_count_J
```

```
positive    1
Name: label, dtype: int64
```

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 GPT-3 model seem good for question answering and text generation, but they are not good for classification task, and hallucinated through wrong response even they're 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:")
```

```
[72] Counter({' sadness': 1121,
              ' love': 341,
              ' joy': 1352,
              ' fear': 710,
              ' anger': 371,
              ' none of the above': 15,
              ' surprise': 222,
              ' exhaustion': 1,
              ' uncertainty': 4,
              ' sympathy/love': 7,
              ' jealousy (not one of': 3,
              ' curiosity': 12,
              ' pride': 1,
              ' determination': 1,
              ' sympathy': 8,
              ' neutral': 1})
```

Source: Own representation.

Instead of following human instructions to classify input into 6 sentiment categories, the model doesn't just follow the instruction blindly. It uses its understanding of language to try and 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 6 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 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 Classification report. We will evaluate each score of the top three, which are **DistilBertForSequenceClassification, Linear SVC,** and **BertForSequenceClassification.**

Classification report provides a summary of the performance of the classification model on comparing the model's performance to other models or to understand where the model might be struggling.

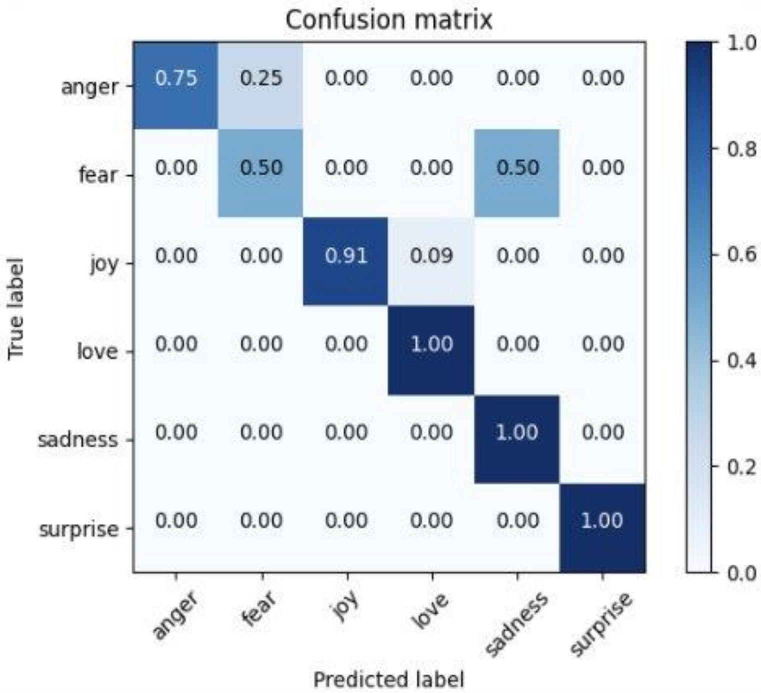
- Accuracy is about how often the model is right.
- Precision is how often the model **predicts** specific emotions, and it's **that** emotion.
- Recall is how often the model correctly predicts a specific emotion.
- F1-score is how good the model overall, mixed of precision and recall.

Figure 13. DistilBertForSequenceClassification: Heatmap of confusion matrix

```
              precision    recall  f1-score   support

       anger       1.00      0.75      0.86         4
        fear       0.67      0.50      0.57         4
         joy       1.00      0.91      0.95        11
        love       0.67      1.00      0.80         2
     sadness       0.83      1.00      0.91        10
    surprise       1.00      1.00      1.00         1

    accuracy                           0.88        32
   macro avg       0.86      0.86      0.85        32
weighted avg       0.89      0.88      0.87        32
```

```
# Get the confusion matrix
cm = confusion_matrix(df_metrics['Actual_class'].values, df_metrics['Predicted_class'].values)
classes = label2int.keys()
plot_confusion_matrix(cm, classes, normalize=True, title='Confusion matrix')
```

```
Normalized confusion matrix
[[0.75       0.25       0.         0.         0.         0.        ]
 [0.         0.5        0.         0.         0.5        0.        ]
 [0.         0.         0.90909091 0.09090909 0.         0.        ]
 [0.         0.         0.         1.         0.         0.        ]
 [0.         0.         0.         0.         1.         0.        ]
 [0.         0.         0.         0.         0.         1.        ]]
```



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, which is only 67%. The Recall score was the worst, only 50%, and f1-score is only 57%.

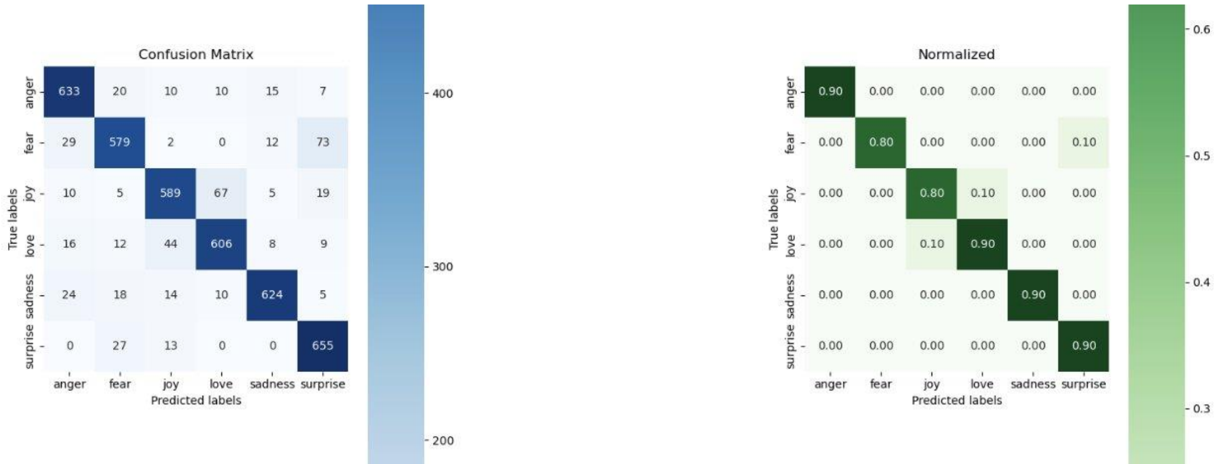
Particularly concerning is that from the heatmap, 25% of “anger” sentiment is falsely classified as “fear”, although misclassifying anger as fear is understandable, as both are negative sentiments.

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 them as “fear”.

Since DistilBertForSequenceClassification model performs very good in several classes, but poorly in several classes, then it is necessary to compare it with another model.

Figure 14. Linear SVC's heatmap of normalized confusion matrix




```
Cross Validation Scores:  [0.91495711 0.9129678  0.91918438 0.9129678 ]
Average Cross Validation Score:  0.9150192714161384
Accuracy with L1 regularization and cross validation:  0.8839328537170263
```

```
performance_evaluation()
```

```
              precision    recall  f1-score   support

       anger       0.89      0.91      0.90       695
        fear       0.88      0.83      0.85       695
         joy       0.88      0.85      0.86       695
        love       0.87      0.87      0.87       695
     sadness       0.94      0.90      0.92       695
    surprise       0.85      0.94      0.90       695

    accuracy                           0.88      4170
   macro avg       0.88      0.88      0.88      4170
weighted avg       0.88      0.88      0.88      4170

[[633  20  10  10  15   7]
 [ 29 579   2   0  12  73]
 [ 10   5 589  67   5  19]
 [ 16  12  44 606   8   9]
 [ 24  18  14  10 624   5]
 [  0  27  13   0   0 655]]

[[0.9 0.  0.  0.  0.  0. ]
 [0.  0.8 0.  0.  0.  0.1]
 [0.  0.  0.8 0.1 0.  0. ]
 [0.  0.  0.1 0.9 0.  0. ]
 [0.  0.  0.  0.  0.9 0. ]
 [0.  0.  0.  0.  0.  0.9]]
```

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.

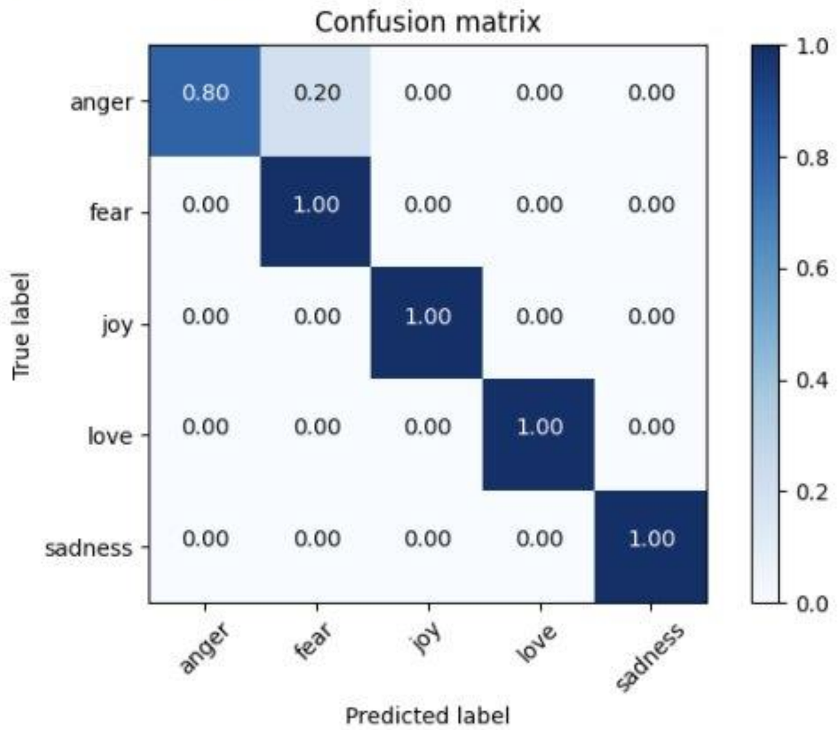
The Linear SVC model had a relatively high accuracy, with a total of 88 % out of 4170 total predictions in the test labeled dataset, with high scores for all evaluations metrics, in all classes.

Figure 15. BertForSequenceClassification: Heatmap of Confusion matrix

```
              precision    recall  f1-score   support

       anger       1.00      0.80      0.89         5
        fear       0.83      1.00      0.91         5
         joy       1.00      1.00      1.00        11
        love       1.00      1.00      1.00         1
     sadness       1.00      1.00      1.00        10

    accuracy                           0.97        32
   macro avg       0.97      0.96      0.96        32
weighted avg       0.97      0.97      0.97        32
```

```
In [ ]:  cm = confusion_matrix(df_metrics['Actual_class'].values, df_metrics['Predicted_class'].values)
         classes = label2int.keys()
         plot_confusion_matrix(cm, classes, normalize=True, title='Confusion matrix')
```

```
Normalized confusion matrix
[[0.8 0.2 0.  0.  0. ]
 [0.  1.  0.  0.  0. ]
 [0.  0.  1.  0.  0. ]
 [0.  0.  0.  1.  0. ]
 [0.  0.  0.  0.  1. ]]
```



Source: Own representation.

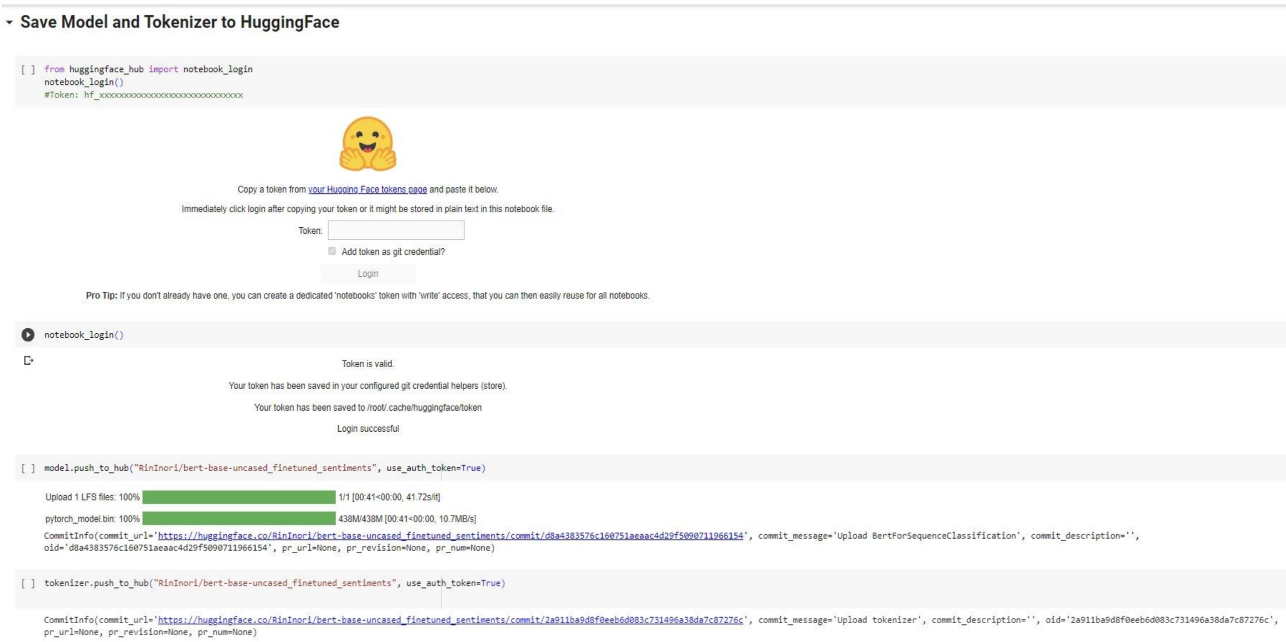
BertForSequenceClassification Interpretation:

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, 97%, with the lowest score still high, which is 83% precision in "fear" class.

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



Source: Own representation.

Load Model and Vectorizer, predict Unlabeled data.

The saved model which trained with labeled dataset can be loaded to predict new the unlabeled dataset.

Load Model From Hugging face

https://huggingface.co/RinInori

```
from transformers import AutoModelForSequenceClassification, AutoTokenizer
import torch
import pandas as pd

def predict_sentiments(model_name, tokenizer_name, input_file):

    model = AutoModelForSequenceClassification.from_pretrained(model_name)
    tokenizer = AutoTokenizer.from_pretrained(tokenizer_name)

    df = pd.read_csv(input_file)

    # Tokenize the input text
    test_inputs = tokenizer(list(df['text']), padding=True, truncation=True, max_length=128, return_tensors='pt')

    # Make predictions
    with torch.no_grad():
        model.eval()
        outputs = model(test_inputs['input_ids'], token_type_ids=None, attention_mask=test_inputs['attention_mask'])
        logits = outputs[0].detach().cpu().numpy()
        predictions = logits.argmax(axis=-1)

    # Map the predicted labels back to their original names
    int2label = {0: 'anger', 1: 'fear', 2: 'joy', 3: 'love', 4: 'sadness', 5: 'surprise'}
    predicted_labels = [int2label[p] for p in predictions]

    # Add the predicted labels to the test dataframe
    df['label'] = predicted_labels

    # Save the predictions to a file
    output_file = input_file.replace(".csv", "_predicted2.csv")
    df.to_csv(output_file, index=False)

model_name = "RinInori/bert-base-uncased_finetuned_sentiments"
tokenizer_name = "RinInori/bert-base-uncased_finetuned_sentiments"
```

Source: Own representation.

4.3 Unlabeled Data Collection

```
tweet_data = open('dc_America.csv', 'a', newline='', encoding='utf8')
csv.writer(tweet_data).writerow(['text'])
max_tweets = 5000

queries = ['jibopetbot','jiborobot','jibopetrobot','jibosocialrobot','savejibo', 'jibobot',
          'vectorpetbot','vectorrobot','vectorpetrobot','savevector','replika','repikaai','amazonalexa',
           'cozmopetbot','cozmorobot', 'cozmopetrobot','ankipetbot','ankirobot', 'ankipetrobot','saveanki']

for query in queries:
    for n,tweet in enumerate(sntwitter.TwitterSearchScraper(
        f"{query} since:2014-01-01 until:2023-06-30 near:Boston +\
        within:1000km lang:en -filter:links -filter:replies").get_items()):
        if n > max_tweets:
            break
        csv.writer(tweet_data).writerow([tweet.content])

tweet_data.close()
print("Done")
```

Source: Own representation.

4.4 Sentiment Analysis Result on Unlabeled Data

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

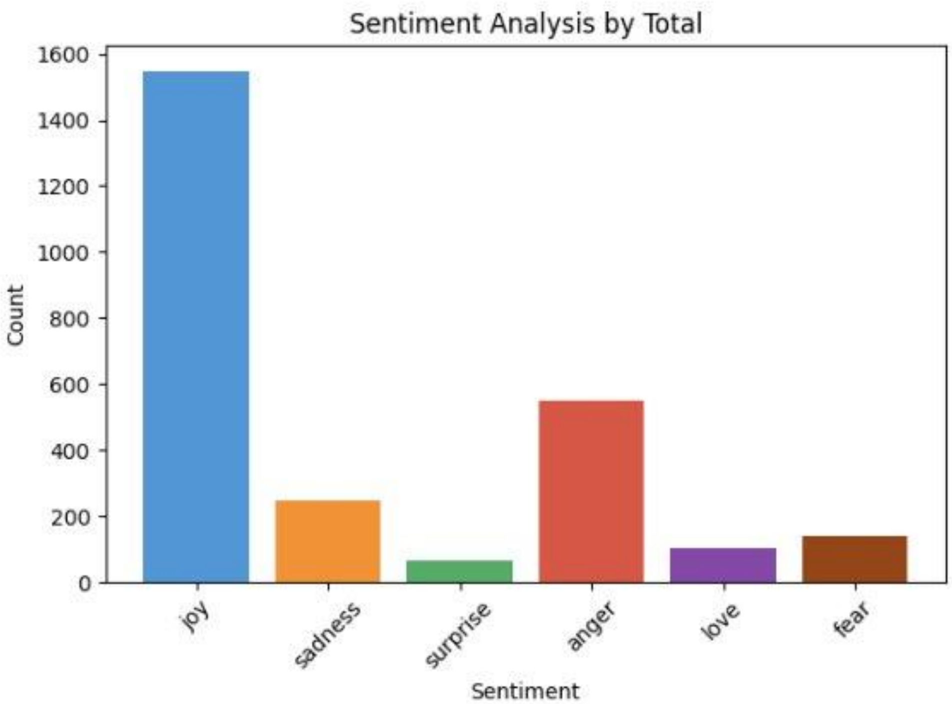
Table 4. Bert For Sequence Classification model sentiment prediction on unlabeled data

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 prediction 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.

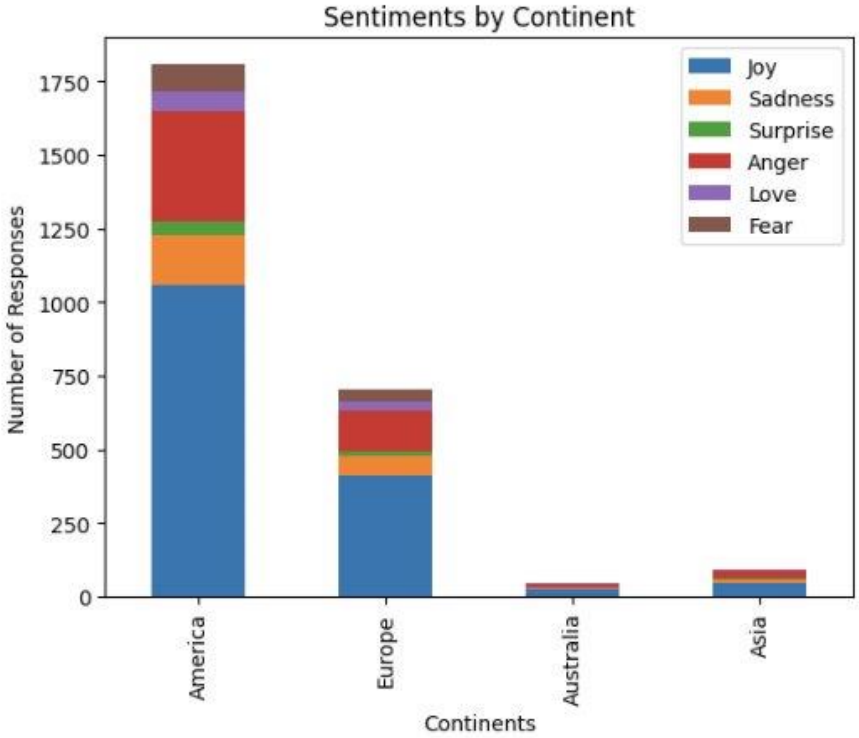
Figure 16. Sentiment analysis charts



Source: Own representation.

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

Figure 17. 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.

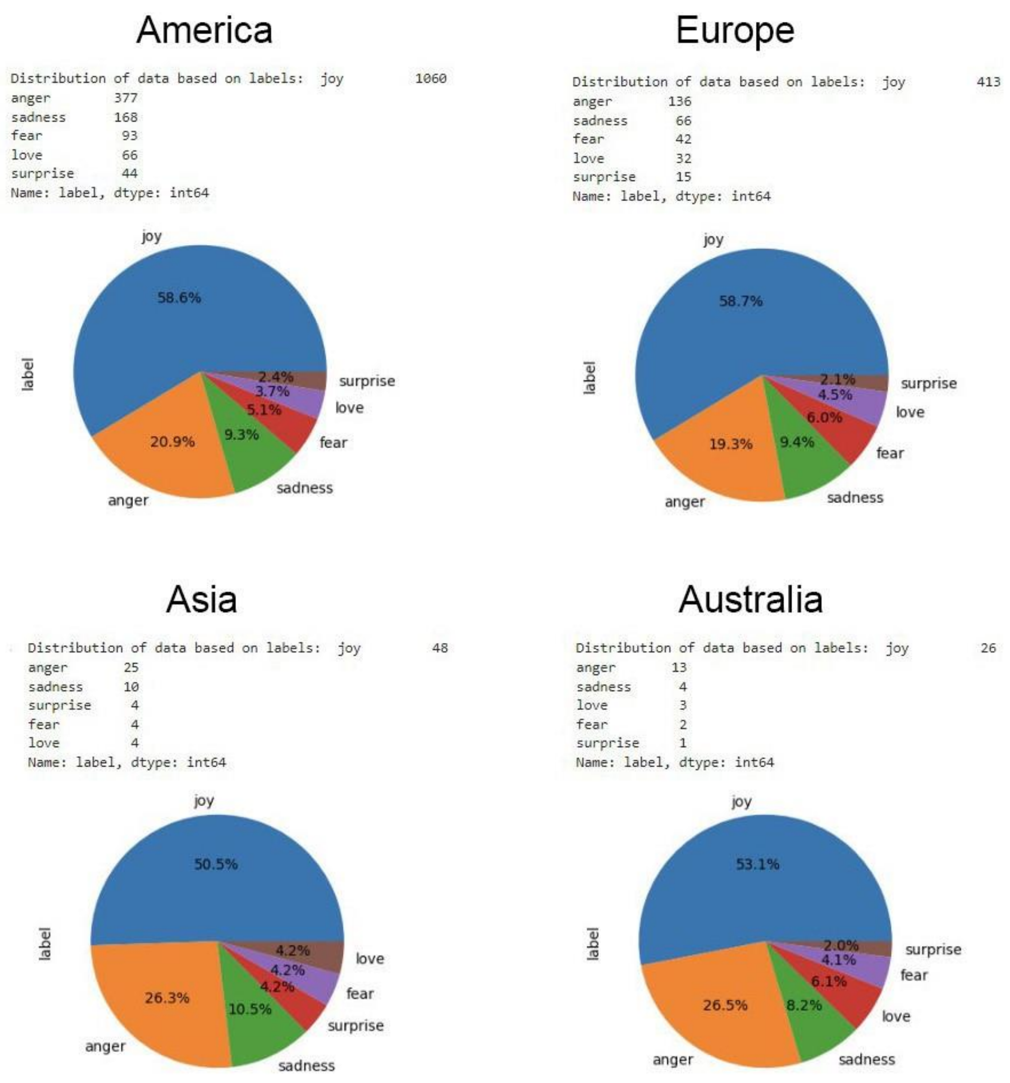
Answer for Research Question #1:

1. Will most sentiments about digital and robot pet companions be positive, like joy & love, or will be negative, like anger & 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, it will be removed from further analysis.

Figure 18. BertForSequenceClassification sentiment analysis pie charts



Source: Own representation.

Research Question #2:

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

To investigate and answer research question #2, hypothesis testing will be performed.

4.5 Hypothesis Testing Result

America and Europe

Table 5. Observed frequencies America and Europe

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
Column Total	1473	234	59	**513**	98	135	**2512**

Source: Own representation.

degree of freedom = (row - 1) * *(column - 1)*

df = (2 - 1) * *(6 - 1) = 5*

O = observed frequency

E = Expected Frequency = (Row total * Column total) / Grand total

For example: America anger = (1808*513) / 2512 = **369**

Table 6. Expected frequencies America and Europe

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
America	1060	168	42	**369**	71	97	1808
Europe	413	66	17	144	27	38	704
Column Total	1473	234	59	513	98	135	2512

Source: Own representation.

χ^2 = chi-square statistic = $\sum (O - E)^2 / E$

For example: America anger = ((377 - 369)^2) / 369 = **0.16**

Table 7. Chi-Square statistic America and Europe

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
America	0,000	0,001	0,095	**0,173**	0,292	0,179	0,690
Europe	0,000	0,003	0,143	0,420	0,749	0,459	1,773
Column Total	0,000	0,004	0,198	0,584	1,040	0,637	**2,463**

Source: Own representation.

From the calculation, **Chi-Square Statistic** is **2,463**

Table 8. 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**

Source: Chi Square table in Excel, 2012.

Based on the Chi-Square table, the **Critical value** for chi-squared statistic with degree of freedom of 5 and a p-value of 0.05 is **11.07.**

Null hypothesis h0:

Joy, Sadness, Surprise, Anger, Love, Fear sentiments are INDEPENDENT upon different continents.

Alternative hypothesis h1:

Joy, Sadness, Surprise, Anger, Love, Fear sentiments are INDEPENDENT upon different continents.

As the Chi-Square Statistic 2.463 lesser than Critical value 11,07, indicating that the observed frequencies are not significantly different from the expected frequencies.

This means that it is **not possible to reject the null hypothesis** of **independence** between the two variables.

Calculation using Python show the same result:

America and Europe

```
[5] tab_data = [[1060, 168, 44, 377, 66, 93],    [413, 66, 15, 136, 32, 42],]
    chi2_contingency(tab_data)

    Chi2ContingencyResult(statistic=2.4630207573107077, pvalue=0.782053628310583, dof=5, expected_freq=array([[1060.18471338,
    168.42038217,   42.46496815,  369.22929936,
             70.53503185,   97.1656051 ],
           [ 412.81528662,   65.57961783,   16.53503185,  143.77070064,
             27.46496815,   37.8343949 ]]))
```

Table 9. Hypothesis Testing result between America and Europe

Continent Pair	Chi-Squared Statistic	Critical Value	P-Value (significance level (0.05))	Degrees of Freedom	Independent (h0)?
America and Europe	2,463	11.07	0.782	5	Yes

Source: Own representation.

Chi-squared statistic's result, 2.463, is a value used to determine whether the observed frequencies of data significantly different from what expected.

A p-value benchmark used for determining statistical significance is 0.05. It is like a margin of error, and if the p-value is less than or equal to 0.05, it means that there is a real difference between the expected and observed results.

The p-value outcome is 0.782, which means that if the null hypothesis (independence) is true, there is an 80.8% probability of obtaining the observed results by chance.

Since the p-value is greater than 0.05, it is not possible to reject the null hypothesis of independence between America and Europe and sentiment analysis of digital and robot pet companions.

The degree of freedom, 5, shows how many ways the frequencies can change on their own. The result is an array that shows how often each cell in the contingency table is likely to happen.

The critical value for the chi-squared statistic with a degree of freedom of 5 and a p-value of 0.05 is 11.07. The observed Chi-Squared statistic, 2.463, does not exceed this critical value, indicating that the observed frequencies are not significantly different from the expected frequencies. This means that the null hypothesis that the two variables are independent cannot be rejected.

America and Asia

Table 10. Observed frequencies for America and Asia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
America	1060	168	44	377	66	93	1808
Asia	48	10	4	25	4	4	95
Column Total	1108	178	48	402	70	97	1903

Source: Own representation.

Table 11. Expected frequencies for America and Asia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
America	1053	169	46	382	67	92	1808
Asia	55	9	2	20	3	5	95
Column Total	1108	178	48	402	70	97	1903

Source: Own representation.

Table 12. Chi-Square Statistic America and Asia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
America	0.051	0.007	0.056	0.064	0.004	0.008	0.190
Asia	0.967	0.140	1.073	1.212	0.073	0.147	3.611
Column Total	1.018	0.147	1.130	1.276	0.077	0.154	**3.801**

Source: Own representation.

From the calculation, **Chi-Square Statistic** is **3.801**

America and Asia

```
tab_data = [[1060, 168, 44, 377, 66, 93],[48, 10, 4, 25, 4, 4]]
chi2_contingency(tab_data)
```

```
Chi2ContingencyResult(statistic=3.8012122038716587, pvalue=0.5783767032863041, dof=5,
expected_freq=array([[1052.68733579,  169.11403048,   45.6037835 ,  381.93168681,
          66.5055176 ,   92.15764582],
       [  55.31266421,    8.88596952,    2.3962165 ,   20.06831319,
           3.4944824 ,    4.84235418]]))
```

Table 13. Hypothesis testing result between America and Asia

Continent Pair	Chi-Squared Statistic	Critical Value	P-Value (significance level (0.05))	Degrees of Freedom	Independent (h0)?
America and Asia	3.801	11.07	0.578	5	Yes

Source: Own representation.

Chi-squared statistic's result, 3.801, is a value used to determine whether the observed frequencies of data significantly different from what is expected.

A p-value benchmark used for determining statistical significance is 0.05 is like a margin of error, and if the p-value is less than or equal to 0.05, it means that there is a real difference between the expected and observed results.

The p-value outcome is 0.578, which means that if the null hypothesis (independence) is true, there is a 57.8% probability of obtaining the observed results by chance. Since the p-value is greater than 0.05, it is not possible to reject the null hypothesis of independence between America and Asia and sentiment analysis of digital and robot pet companions.

The degree of freedom, 5, shows how many ways the frequencies can change on their own.

The critical value for the chi-squared statistic with a degree of freedom of 5 and a p-value of 0.05 is 11.07. The observed Chi-Squared statistic, 3.801, does not exceed this critical value, indicating that the observed frequencies are not significantly different from the expected frequencies. This means that the null hypothesis that the two variables are independent cannot be rejected.

America and Australia

Table 14. Observed frequencies for America and Australia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
America	1060	168	44	377	66	93	1808
Australia	26	4	1	13	3	2	49
Column Total	1086	172	45	390	69	95	1857

Source: Own representation.

Table 15. Expected frequencies for America and Australia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
America	1057	167	44	380	67	92	1808
Australia	29	5	1	10	2	3	49
Column Total	1086	172	45	390	69	95	1857

Source: Own representation.

Table 16. Chi-Square statistic America and Australia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
America	0.007	0.002	0.001	0.019	0.021	0.003	0.052
Australia	0.246	0.064	0.030	0.713	0.764	0.102	1.919
Column Total	0.253	0.066	0.030	0.733	0.785	0.105	1.971

Source: Own representation.

From the calculation, Chi-Square Statistic is **1.971**

America and Australia

```
[12] tab_data = [[1060, 168, 44, 377, 66, 93],[26, 4, 1, 13, 3, 2]]
     chi2_contingency(tab_data)

Chi2ContingencyResult(statistic=1.9712046958304237, pvalue=0.8531149043640054, dof=5, expected_freq=array([[1057.34410339,  167.46149704,   43.81260097,  379.7092084 ,
          67.17932149,   92.49326871],
       [  28.65589661,    4.53850296,    1.18739903,   10.2907916 ,
           1.82067851,    2.50673129]]))
```

Table 17. Hypothesis testing result between America and Australia

Continent Pair	Chi-Squared Statistic	Critical Value	P-Value (significance level (0.05))	Degrees of Freedom	Independent (h0)?
America and Australia	1.971	11.07	0.853	5	Yes

Source: Own representation

Chi-squared statistic's result, 1.971 is a value used to determine whether the observed frequencies of data significantly different from what expected.

A p-value benchmark used for determining statistical significance is 0.05. It is like a margin of error, and if the p-value is less than or equal to 0.05, it means that there is a real difference between the expected and observed results.

The p-value outcome is 0.853, which means that if the null hypothesis (independence) is true, there is 85.3% probability of obtaining the observed results by chance. Since the p-value is greater than 0.05, it is not possible to reject the null hypothesis of independence between America and Australia and sentiment analysis of digital and robot pet companions.

The degree of freedom, 5, shows how many ways the frequencies can change on their own. The result shows how often each cell in the contingency table is likely to happen.

The critical value for the chi-squared statistic with a degree of freedom of 5 and a p-value of 0.05 is 11.07. The observed Chi-Squared statistic, 1.971, does not exceed this critical value, indicating that the observed frequencies are not significantly different from the expected frequencies. This means that the null hypothesis that the two variables are independent cannot be rejected.

Europe and Asia

Table 18. Observed frequencies for Europe and Asia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
Europe	413	66	15	136	32	42	704
Asia	48	10	4	25	4	4	95
Column Total	461	76	19	161	36	46	799

Source: Own representation.

Table 19. Expected frequencies for Europe and Asia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
Europe	406	67	17	142	32	41	704
Asia	55	9	2	19	4	5	95
Column Total	461	76	19	161	36	46	799

Source: Own representation.

Table 20. Chi-Square statistic Europe and Asia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
Europe	0.114	0.014	0.181	0.242	0.002	0.053	0.607
Asia	0.847	0.103	1.342	1.792	0.018	0.395	4.496
Column Total	0.961	0.117	1.523	2.034	0.021	0.448	**5.103**

Source: Own representation.

From the calculation, Chi-Square Statistic is **5.103**

Europe and Asia

```
[8] tab_data = [[413, 66, 15, 136, 32, 42],[48, 10, 4, 25, 4, 4]]
    chi2_contingency(tab_data)

    Chi2ContingencyResult(statistic=5.103145801091831, pvalue=0.40342237407794435, dof=5, expected_freq=array([[406.18773467,
    66.96370463,  16.74092616, 141.85732165,
             31.71964956,  40.53066333],
           [ 54.81226533,   9.03629537,   2.25907384,  19.14267835,
              4.28035044,   5.46933667]]))
```

Table 21. Hypothesis testing result between Europe and Asia

Continent Pair	Chi-Squared Statistic	Critical Value	P-Value (significance level (0.05))	Degrees of Freedom	Independent (h0)?
Europe and Asia	5.103	11.07	0.403	5	Yes

Source: Own representation.

Chi-squared statistic's result, 5.103 is a value used to determine whether the observed frequencies of data significantly different from what expected.
A p-value benchmark used for determining statistical significance is 0.05. It is like a margin of error, and if the p-value is less than or equal to 0.05, it means that there is a real difference between the expected and observed results.

The p-value outcome is 0.403, which means that if the null hypothesis (independence) is true, there is a 19.3% probability of obtaining the observed results by chance. Since the p-value is greater than 0.05, it is not possible to reject the null hypothesis of independence between Europe and Asia and sentiment analysis of digital and robot pet companions.

The degree of freedom, 5, shows how many ways the frequencies can change on their own. The result is an array that shows how often each cell in the contingency table is likely to happen.

The critical value for the chi-squared statistic with a degree of freedom of five and a p-value of 0.05 is 11.07. The observed Chi-Squared statistic, 5.103, does not exceed this critical value, indicating that the observed frequencies are not significantly different from the expected frequencies. This means that the null hypothesis that the two variables are independent cannot rejected.

Europe and Australia

Table 22. Observed frequencies for Europe and Australia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
Europe	413	66	15	136	32	42	704
Australia	26	4	1	13	3	2	49
Column Total	439	70	16	149	35	44	753

Source: Own representation.

Table 23. Expected frequencies for Europe and Australia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
Europe	410	65	15	139	33	41	704
Australia	29	5	1	10	2	3	49
Column Total	439	70	16	149	35	44	753

Source: Own representation.

Table 24. Chi-Square statistic Europe and Australia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
Europe	0.016	0.005	0.000	0.078	0.016	0.018	0.133
Australia	0.231	0.068	0.002	1.126	0.229	0.260	1.915
Column Total	0.247	0.072	0.002	1.204	0.245	0.278	**2.049**

Source: Own representation.

From the calculation, Chi-Square Statistic is **2.049**

Europe and Australia

```
tab_data = [[413, 66, 15, 136, 32, 42],[26, 4, 1, 13, 3, 2]]
chi2_contingency(tab_data)
```

```
Chi2ContingencyResult(statistic=2.04863351474268, pvalue=0.8423755734754516, dof=5, expected_freq=array([[410.43293493,
65.44488712,  14.95883134, 139.30411687,
         32.72244356,  41.13678619],
       [ 28.56706507,   4.55511288,   1.04116866,   9.69588313,
          2.27755644,   2.86321381]]))
```

Table 25. Hypothesis testing result between Europe and Australia

Continent Pair	Chi-Squared Statistic	Critical Value	P-Value (significance level (0.05))	Degrees of Freedom	Independent (h0)?
Europe and Australia	2.049	11.07	0.842	5	Yes

Source: Own representation.

Chi-squared statistic's result, 2.049 is a value used to determine whether the observed frequencies of data significantly different from what expected.

A p-value benchmark used for determining statistical significance is 0.05. It is like a margin of error, and if the p-value is less than or equal to 0.05, it means that there is a real difference between the expected and observed results.

The p-value outcome is 0.842, which means that if the null hypothesis (independence) is true, there is an 84% probability of obtaining the observed results by chance. Since the p-value is greater than 0.05, it is not possible to reject the null hypothesis of independence between Europe and Australia about sentiment analysis of digital and robot pet companions.

The degree of freedom 5 shows how many ways the frequencies can change on their own. The result is an array that shows how often each cell in the contingency table is likely to happen.

The critical value for the chi-squared statistic with a degree of freedom of 5 and a p-value of 0.05 is 11.07. The observed Chi-Squared statistic, 2.049, does not exceed this critical value, indicating that the observed frequencies are not significantly different from the expected frequencies. This means that the null hypothesis that the two variables are independent cannot be rejected.

Asia and Australia

Table 26. Observed frequencies for Asia and Australia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
Asia	48	10	4	25	4	4	95
Australia	26	4	1	13	3	2	49
Column Total	74	14	5	38	7	6	144

Source: Own representation.

Table 27. Expected frequencies for Asia and Australia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
Asia	49	9	3	25	5	4	95
Australia	25	5	2	13	2	2	49
Column Total	74	14	5	38	7	6	144

Source: Own representation.

Table 28. Chi-Square statistic Asia and Australia

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
Asia	0.014	0.063	0.149	0.000	0.083	0.000	0.309
Australia	0.027	0.122	0.289	0.000	0.160	0.001	0.600
Column Total	0.040	0.186	0.438	0.001	0.243	0.001	0.909

Source: Own representation.

From the calculation, Chi-Square Statistic is **0.909**

Asia and Australia

```
[10] tab_data = [[48, 10, 4, 25, 4, 4],[26, 4, 1, 13, 3, 2]]
     chi2_contingency(tab_data)

     Chi2ContingencyResult(statistic=0.9093127570282862, pvalue=0.9695437657008573, dof=5, expected_freq=array([[48.81944444,
     9.23611111,  3.29861111, 25.06944444,  4.61805556,
             3.95833333],
           [25.18055556,  4.76388889,  1.70138889, 12.93055556,  2.38194444,
             2.04166667]]))
```

Table 29. Hypothesis testing result between Asia and Australia

Continent Pair	Chi-Squared Statistic	Critical Value	P-Value (significance level (0.05))	Degrees of Freedom	Independent (h0)?
Asia and Australia	0.909	11.07	0.969	5	Yes

Source: Own representation.

Chi-squared statistic's result, 0.909 is a value used to determine whether the observed frequencies of data significantly different from what expected.

A p-value benchmark used for determining statistical significance is 0.05. It is like a margin of error, and if the p-value is less than or equal to 0.05, it means that there is a real difference between the expected and observed results.

The p-value outcome is 0.969, which means that if the null hypothesis (independence) is true, there is a 96.9% probability of obtaining the observed results by chance. Since the p-value is greater than 0.05, it is not possible to reject the null hypothesis of independence between Asia and Australia and sentiment analysis of digital and robot pet companions.

The degree of freedom, 5, shows how many ways the frequencies can change on their own. The result is an array that shows how often each cell in the contingency table is likely to happen.

The critical value for the chi-squared statistic with a degree of freedom of 5 and a p-value of 0.05 is 11.07. The observed Chi-Squared statistic, 0.909, does not exceed this critical value, indicating that the observed frequencies are not significantly different from the expected frequencies. This means that the null hypothesis that the two variables are independent cannot be rejected.

All Pairing Hypothesis Test Result

Table 30. Hypothesis pair testing result

Continent Pair	Chi-Squared Statistic	Critical Value	P-Value (significance level (0.05))	Degrees of Freedom	Independent (h0)?
America and Europe	2,463	11.07	0.782	5	Yes
America and Asia	3.801	11.07	0.578	5	Yes
America and Australia	1.971	11.07	0.853	5	Yes
Europe and Asia	5.103	11.07	0.403	5	Yes
Europe and Australia	2.049	11.07	0.842	5	Yes
Asia and Australia	0.909	11.07	0.969	5	Yes

Source: Own representation.

Table 31. Independency result

Sentiment	**Europa**	**Asia**	**Australia**
America	Independent	Independent	Independent
Europe	N/A	Independent	Independent
Asia	Independent	N/A	Independent
Australia	Independent	Independent	N/A

Source: Own representation.

All Continents Together

Since sample from Africa is only 1, it’s too small to perform Hypothesis testing.

Table 32. Observed frequencies.

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
Column Total	1547	248	64	551	105	141	2656

Source: Own representation.

degree of freedom = (row - 1) * *(column - 1)*

df = (4 - 1) * *(*6 - 1*)* = **15**

Table 33. Expected Frequencies

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
America	1053	169	44	375	71	96	1808
Europe	410	66	17	146	28	37	704
Australia	29	5	1	10	2	3	49
Asia	55	9	2	20	4	5	95
Column Total	1547	248	64	551	105	141	2656

Source: Own representation.

Table 34. Chi-Square statistic

Continent	Joy	Sadness	Surprise	Anger	Love	Fear	Row Total
America	0.045	0.004	0.004	0.010	0.420	0.093	0.576
Europe	0.021	0.001	0.227	0.691	0.624	0.573	2.138
Australia	0.226	0.072	0.028	0.790	0.583	0.139	1.839
Asia	0.972	0.144	1.279	1.421	0.016	0.216	4.047
Column Total	1.265	0.221	1.538	2.913	1.643	1.020	**8.600**

Source: Own representation.

From the calculation. Chi-Square Statistic is **8.6**

Table 35. Chi-Square table

	Probability				
df	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**

Source: Adapted from: Chi Square Table Dalam Excel. (2012. July 1).

Based on Chi-Square table. the **Critical value** for chi-squared statistic with degree of freedom of 15 and a p-value of 0.05 is **24.9957.**

Null hypothesis h0:

Joy. Sadness. Surprise. Anger. Love. Fear sentiments are INDEPENDENT upon different continents.

Alternative hypothesis h1:

Joy. Sadness. Surprise. Anger. Love. Fear sentiments are DEPENDENT upon different continents.

As the **Chi-Square Statistic 8.6** is **lesser than Critical value 24.99579**. it means that the observed frequencies are not significantly different from the expected frequencies. This means that it is **not possible to reject the null hypothesis of independence** between variables.

```
import numpy as np
import pandas as pd
import scipy.stats as stats
from scipy.stats import chisquare
from scipy.stats import chi2_contingency

tab_data =[
    [1060, 168, 44, 377, 66, 93],
    [413, 66, 15, 136, 32, 42],
    [26, 4, 1, 13, 3, 2],
    [48, 10, 4, 25, 4, 4],
]

chi2_contingency(tab_data)
```

```
Chi2ContingencyResult(statistic=8.599593697366872, pvalue=0.8975061938439917, dof=15, expected_freq=array([[1053.07831325,  168.81927711,   43.56626506,  375.07831325,
          71.47590361,   95.98192771],
       [ 410.04819277,   65.73493976,   16.96385542,  146.04819277,
          27.8313253 ,   37.37349398],
       [  28.54028614,    4.5753012 ,    1.18072289,   10.16528614,
           1.93712349,    2.60128012],
       [  55.33320783,    8.87048193,    2.28915663,   19.70820783,
           3.75564759,    5.04329819]]))
```

Source: Own representation.

The first value 8.6. is the chi-squared statistic. This value is important to assess whether the observed frequencies are significantly different from the expected frequencies or not.

The second value. 0.897 is the p-value.

The p-value is 0.897 which means that there is a 89.7% chance of obtaining the observed result if the null hypothesis is true.

The p-value is greater than 0.05. which means it is not possible to reject the null hypothesis of independence based on the observed data.

Therefore. there is a high probability of getting the observed frequencies if the variables (continent and sentiment) are independent.

Therefore. we CAN NOT reject the null hypothesis since there are no significant dependencies between the continent and sentiment.

The third value 15 is the degrees of freedom.
Calculated by (row-1) * (column-1) = (6-1)*(4-1) = 5*3 = 15.

The critical value for the chi-squared statistic with a degree of freedom 15 and p-value 0.05 is **24.99579**. The chi-squared statistic is 10.964562027950564. It does not exceed this critical value. showing that the observed frequencies are not significantly differed from the expected frequencies. Therefore. it is not possible to reject the null hypothesis of independence among two variables.

5 Discussion

5.1 Result Interpretation

To determine which continents are good for AI robot digital companion business marketing based on the percentage of sentiment analysis. we could calculate the proportion of sentiments for each continent. We might measure each continent's sentiment % to decide which continents are good for AI robot digital companion marketing promotion.

To calculate each continent's positive (joy. love). neutral (surprise. sadness). and negative (anger. fear) feelings, need to divide the number of each sentiment category by the total number for each continent and multiply by 100.

America:
Positive (joy + love): (1060 + 66) / 1808 * 100 = 62.6%
Neutral (sadness+ surprise): (168 + 44) / 1808 * 100 = 11.7%
Negative (anger + fear): (377 + 93) / 1808 * 100 = 25.7%

Europe:
Positive: (413 + 32) / 704 * 100 = 63.2%
Neutral: (66 + 15) / 704 * 100 = 11.5%
Negative: (136 + 42) / 704 * 100 = 25.2%

Asia:
Positive: (48 + 4) / 95 * 100 = 54.7%
Neutral: (10 + 4) / 95 * 100 = 14.7%
Negative: (25 + 4) / 95 * 100 = 30.5%

Australia:
Positive: (26 + 3) / 49 * 100 = 59.1%
Neutral: (4 + 1) / 49 * 100 = 10.2%
Negative: (13 + 2) / 49 * 100 = 30.6%

Based on the calculated sentiment proportions, all continents have relatively high positive sentiment percentages.

Europe has the highest positive sentiment proportion (63.2%), followed by America (62.6%), Australia (59.1%), and Asia (54.7%).

All continents have strong positive sentiments. AI robot digital companion company marketing has a high possibility to be successful in all these locations.

Also, according to the Chi-Square test results, the null hypothesis of independence between variables cannot be rejected. This means that there is not enough evidence to conclude that sentiments depend on different continents. Therefore. it may not be a significant factor to determine which continents are suitable for AI robot digital companion business marketing.

Since all continents have similar positive sentiment percentages. then AI robot digital companion company marketing has a high possibility to be successful everywhere in any locations in this planet.

5.2 Discussion of Limitations

This study has several limitations. such as:

- Collected tweets don’t differentiate between age, gender, race, or cultural background.
- This study only looked at tweets written in English. This might not reflect tweets where English is not the language spoken by most people.
- The application of machine learning models for conducting sentiment analysis also comes with several potential drawbacks and biases. given that these algorithms are not always accurate and make frequent errors. For example, a model that uses machine learning could have difficulty accurately detecting tweets that contain irony or sarcasm. as well as tweets that show a negative attitude through negation.
- The sample size of tweets collected from Africa is too small. insufficient to accurately evaluate the relationship between location and sentiment in this region.

Therefore, more study is needed to validate the nature of the link between the variables and gain a deeper understanding of the underlying reasons for their relationships.

6 Conclusion

6.1 Summary

These findings contribute to our understanding of human-machine relationships in different cultural contexts and have implications for the design and use of social robots and other artificial intelligence technologies. Based on the results of this study. it can be concluded that location does not significantly influence people's sentiments towards digital and robot pet companions.

This suggests that businesses marketing AI robot digital companions can focus their marketing efforts on any location without considering the potential impact of location on sentiment.

6.2 Recommendations for Future Research

Further research is needed so that we can understand more deeply about other factors that might influence the human-machine relationship regarding robot companions. including other variables such as gender. age. and background cultural dimensions such as power distance. uncertainty avoidance. individualism - collectivism. masculinity-femininity. and short vs. long-term orientation as in Hofstede's Cultural Dimensions Theory. (The 6 Dimensions Model of National Culture by Geert Hofstede. n.d.).

In addition. further studies are needed regarding people's concerns about security and privacy when communicating with digital friends. Addressing these concerns will be important to ensure users feel comfortable and secure in their interactions with their digital companions.

Research in this area could explore further user perceptions of data handling practices. privacy measure effectiveness. regulatory and law potential impact frameworks on user acceptance of AI technology.

Future research can contribute to the development of more culturally sensitive and user-friendly AI technologies. Considering these aspects can ultimately result in better human-machine interactions and enhance the overall user experience.

References

Amazon Alexa - Wikipedia. (2014. November 6). Amazon Alexa - Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Amazon_Alexa

BERT — transformers 3.0.2 documentation. (n.d.). BERT &Mdash; Transformers 3.0.2 Documentation. https://huggingface.co/transformers/v3.0.2/model_doc/bert.html

BERT. (n.d.). BERT. https://huggingface.co/docs/transformers/v4.27.2/en/model_doc/bert

Burkov. A. (2019. January 1). The Hundred-Page Machine Learning Book.

Burkov. A. (2020. September 8). Machine Learning Engineering.

Carman. (2019. June 19). They Welcomed a Robot Into Their Family. Now They're Mourning Its Death - the Verge. Retrieved from https://www.theverge.com/2019/6/19/18682780/jibo-death-server-update-social-robot-mourning

Carman. (2020. July 23). Jibo. the social robot that was supposed to die. is getting a second life. Jibo. the Social Robot That Was Supposed to Die. Is Getting a Second Life - the Verge. Retrieved from https://www.theverge.com/2020/7/23/21325644/jibo-social-robot-ntt-disruptionfunding

Chauhan. V. K.. Dahiya. K.. & Sharma. A. (2018. January 16). Problem formulations and solvers in linear SVM: a review - Artificial Intelligence Review. SpringerLink. Retrieved from https://link.springer.com/article/10.1007/s10462-018-9614-6

Chi-Square Statistic: How to Calculate It / Distribution. (2021). Statistics How To. Retrieved from https://www.statisticshowto.com/probability-and-statistics/chi-square/

Chi-square Test in Spreadsheets. 2019. Retrieved from https://www.datacamp.com/tutorial/chi-square-test-in-spreadsheets

Dangeti. P. (2017. July 21). Statistics for Machine Learning.

Digital Dream Labs. (n.d.). Digital Dream Labs. Retrieved from https://www.digitaldreamlabs.com/

DistilBERT. (n.d.). DistilBERT.

https://huggingface.co/docs/transformers/v4.27.2/en/model_doc/distilbert

DistilBERT — transformers 3.0.2 documentation. (n.d.). DistilBERT &Mdash; Transformers 3.0.2 Documentation. https://huggingface.co/transformers/v3.0.2/model_doc/distilbert.html#distilbertmodel

EMO - LivingAi. (n.d.). LivingAi. Retrieved from https://living.ai/emo/

Frost. J. (2022. January 26). Chi-Square Table. Statistics by Jim. Retrieved from https://statisticsbyjim.com/hypothesis-testing/chi-square-table/

G. (2020. July 23). 10 Techniques to Solve Imbalanced Classes in Machine Learning (Updated 2023). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/

Hanson. Rick. (2019). Taking in the Good vs. The Negativity Bias

Hutto. C.. & Gilbert. E. (2014. May 16). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text | Proceedings of the International AAAI Conference on Web and Social Media. https://doi.org/10.1609/icwsm.v8i1.14550

Hu. E. J.. Shen. Y.. Wallis. P.. Allen-Zhu. Z.. Li. Y.. Wang. S.. Wang. L.. & Chen. W. (2021. June 17). LoRA: Low-Rank Adaptation of Large Language Models. arXiv.org. https://arxiv.org/abs/2106.09685v2

Hutchinson. B.. Ostendorf. M.. & Fazel. M. (2011. September). Low Rank Language Models for Small Training Sets. IEEE Signal Processing Letters. 18(9). 489–492.

Introducing LLaMA: A foundational. 65-billion-parameter language model. (n.d.). Introducing LLaMA: A Foundational. 65-billion-parameter Language Model. https://ai.facebook.com/blog/large-language-model-llama-meta-ai/

Joshua. S. (2022. April 25). How to combine multiple CSV files using Python for your analysis. Medium. Retrieved from https://medium.com/@stella96joshua/how-to-combine-multiple-csv-files-using-python-for-your-analysis-a88017c6ff9e

Kuyda. (2017. July 21). The story of Replika. the AI app that becomes you. YouTube. Retrieved from https://www.youtube.com/watch?v=yQGqMVuAk04

Low Rank Language Models for Small Training Sets. (n.d.). Low Rank Language Models for Small Training Sets | IEEE Journals & Magazine | IEEE Xplore. https://doi.org/10.1109/lsp.2011.2160850

Metatext. (n.d.). Metatext. https://metatext.io/models/bert-base-uncased

MUKHERJEE. (2022). The Maths behind Linear SVC Classifier. Retrieved from https://www.kaggle.com/code/soham1024/the-maths-behind-linear-svc-classifier

Mohri. M.. Rostamizadeh. A.. Talwalkar. A.. & Bach. F. (2012. September 7). Foundations of Machine Learning.

Models - Hugging Face. (2022. November 16). Models - Hugging Face. https://huggingface.co/models

Nandi. S. (2021. July 1). Twitter Sentiment Analysis Using Machine Learning Approaches. Medium. Retrieved from https://nandisoham2017.medium.com/twitter-sentiment-analysis-using-machine-learning-approaches-14fba1b8e357

Mueller. J. P.. & Massaron. L. (2021. February 9). Machine Learning for Dummies. For Dummies.

OpenAI API. (n.d.). OpenAI API. https://platform.openai.com

O. (2023. March 28). GitHub - openai/openai-cookbook: Examples and guides for using the OpenAI API. GitHub. https://github.com/openai/openai-cookbook

Pati. & Pradhan. (2020. December 12). Comparison Between Machine Learning Algorithms Used for Sentiment Analysis. *IAEME Publication.* Retrieved from https://iaeme.com/Home/article_id/IJARET_11_12_026

Raschka. (2015). Python Machine Learning Equation Reference. Retrieved from https://github.com/rasbt/python-machine-learning-book

Redjeki. & Widyarto. (2022). View of Comparison of Seven Machine Learning Algorithms in the Classification of Public Opinion. View of Comparison of Seven Machine Learning Algorithms in the Classification of Public Opinion. Retrieved from https://jurnal.ubd.ac.id/index.php/te/article/view/1046/526

Rozin. P.. & Royzman. E. B. (2001. November). Negativity Bias. Negativity Dominance. and Contagion. Personality and Social Psychology Review. 5(4). 296–320. https://doi.org/10.1207/s15327957pspr0504_2

Sanh. V.. Debut. L.. Chaumond. J.. & Wolf. T. (2020. January 1). *[PDF] DistilBERT. a distilled version of BERT: smaller. faster. cheaper and lighter* | *Semantic Scholar*. [PDF] DistilBERT. a Distilled Version of BERT: Smaller. Faster. Cheaper and Lighter | Semantic Scholar. https://www.semanticscholar.org/paper/DistilBERT%2C-a-distilled-version-of-BERT%3A-smaller%2C-Sanh-Debut/a54b56af24bb4873ed0163b77df63b92bd018ddc

Saying Goodbye To My Emo Robot. (2022. November 9). YouTube. Retrieved from https://www.youtube.com/watch?v=JDQM6E4Vnbs

Siemon. Strohmann. Khosrawi-Rad. Elshan. de Vreede. & Meyer. (2022. July 11). Why Do We Turn to Virtual Companions? A Text Mining Analysis of Replika Reviews. AIS Electronic Library (AISeL) - AMCIS 2022 Proceedings: Why Do We Turn to Virtual Companions? A Text Mining Analysis of Replika Reviews. Retrieved from https://aisel.aisnet.org/amcis2022/sig_hci/sig_hci/10/

Sun. C.. Qiu. X.. Xu. Y.. & Huang. X. (2019. May 14). How to Fine-Tune BERT for Text Classification? arXiv.org. https://arxiv.org/abs/1905.05583v3

Tabel Chi Square Atau Chi Square Table Dalam Excel. (2012. July 1). Uji Statistik. Retrieved from https://www.statistikian.com/2012/07/chi-square-tabel-dalam-excel.html

The 6 dimensions model of national culture by Geert Hofstede. (n.d.). Geert Hofstede. Retrieved from https://geerthofstede.com/culture-geert-hofstede-gert-jan-hofstede/6d-model-of-national-culture/

TheBloke/vicuna-7B-1.1-HF · Hugging Face. (n.d.). TheBloke/vicuna-7B-1.1-HF · Hugging Face. https://huggingface.co/TheBloke/vicuna-7B-1.1-HF

T. (2023. April 18). GitHub - tloen/alpaca-lora: Instruct-tune LLaMA on consumer hardware. GitHub. https://github.com/tloen/alpaca-lora

Touvron. H.. Lavril. T.. Izacard. G.. Martinet. X.. Lachaux. M. A.. Lacroix. T.. Rozière. B.. Goyal. N.. Hambro. E.. Azhar. F.. Rodriguez. A.. Joulin. A.. Grave. E.. & Lample. G. (2023. February 27). *LLaMA: Open and Efficient Foundation Language Models*. arXiv.org. https://arxiv.org/abs/2302.13971v1

udiprod. (2007. February 5). SVM with polynomial kernel visualization. YouTube. Retrieved from https://www.youtube.com/watch?v=3liCbRZPrZA

Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality | LMSYS Org. (n.d.). Vicuna: An Open-Source Chatbot Impressing GPT-4 With 90%* ChatGPT Quality | LMSYS Org. https://lmsys.org/blog/2023-03-30-vicuna

Vincent. (2020. January 5). Anki's toy robots are being saved from a digital death. Anki's Toy Robots Are Being Saved From a Digital Death - the Verge. https://www.theverge.com/2020/1/5/21050378/anki-vector-saved-shutdown-servers-assets-bought

Vaswani. A.. Shazeer. N.. Parmar. N.. Uszkoreit. J.. Jones. L.. Gomez. A. N.. Kaiser. U.. & Polosukhin. I. (2017). *Attention is All you Need*. Attention Is All You Need. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Appendices

Appendix directory

Appendix A: Sample codes

Sample codes

INTERNATIONAL
UNIVERSITY OF
APPLIED SCIENCES

Declaration of Authenticity

I hereby declare that I have completed this Bachelors/ Master's thesis on my own and without any additional external assistance. I have made use of only those sources and aids specified and I have listed all the sources from which I have extracted text and content. This thesis or parts thereof have never been presented to another examination board. I agree to a plagiarism check of my thesis via a plagiarism detection service.

Indonesia, 28th May 2023

Place, Date

HENNY PURWADI

Student signature