

### Case Study

International University of Applied Sciences

Data Science

**Development of a Spam Filter**

Henny Purwadi

32009177

Indonesia

henny.purwadi@iubh.de

Lecturer: Prof. Markus Pak

Date of submission: February 21th, 2023

**Abstract**

This project goal is to build a machine learning model to filter and classify messages as spam or not spam, developed by using the CRISP-DM (data understanding, data preparation, modeling, evaluation, and deployment) process, which includes stages of a proposal about how to build the project through the folder structure of a Git repository. The project objectives are to organize the project structure, analyze the provided data set, pre-process the data, vectorize using Tf-Idf, use L1 or L2 regularization, and cross validation to avoid overfitting, and then build the machine learning model. This study compares several algorithms, such as Decision Tree, Logistic Regression, and Naive Bayes, to determine which one is the best model for classifying and filtering messages as spam or not. Then the error analysis will be performed to understand the weaknesses and limitations of the model and raise confidence level from business partner about the model. The results of the analysis will be presented with suggestions for future steps in this area, and the graphical user interface will be proposed along with a concept plan for the model’s integration into daily work.

Github code: <https://github.com/hennypurwadi/Spam_Filter>

|  |  |
| --- | --- |
| **Table of Contents** |  |
| **Acknowledgement** | **ii** |
| **Abstract** | **iii** |
| **List of Figures** | **v** |
| **List of Tables** | **vi** |
| **List of Equations** | **vii** |
| **List of Abbreviations** | **viii** |
| **1 Introduction** | **1** |
| 1.1 Background and Motivation | 7 |
| 1.2 Research Questions and Objective | 8 |
| **2 Literature Review** | **13** |
| 2.1 Overview of Relevant Studies | 13 |
| 2.2 Terminology and Definitions | 19 |
| **3 Methodology** | **45** |
| 3.1 Train and Evaluate performance of Machine Learning Models | 45 |
| 3.2 Compare Accuracy of Different Machine Learning Models | 46 |
| **4 Results and Findings** | **74** |
| 4.1 Comparison of Model Performance on Labeled Dataset | 74 |
| 4.2 Sentiment Analysis Result and Location Comparison | 75 |
| 4.3 Hypothesis Testing Result | 75 |
| **5 Discussion** | **1** |
| 5.1 Result Interpretation | 7 |
| 5.2 Discussion of Limitations | 8 |
| **6 Conclusion** | **13** |
| 6.1 Summary | 13 |
| 6.2 Recommendations for Future Research | 19 |

# List of Figures

Figure 1. Sentiment analysis on a dataset of tweets about robot companions. 1

Figure 2. Agile methodologies used worldwide. 18

Figure 3. Business areas of JPI of total customer pool. 21

# List of Tables

Table 1. The State of Big Data/AI in 2019. 2

Table 2. Proportion of companies successfully transformed into data 2

Table 3. The main challenges during the digital transformation of business. 3

# List of Equations

Equation 2. Linear discriminant function. 19

Equation 3. Euclidean distance equation. 40

Equation 4. Winning neuron equation. 42

# List of Abbreviations

CRISP-DM CRISP-DM

ML Machine learning.

Tf-Idf T

**1 Introduction  
1.1 Background and Motivation**

This project was created by following CRISP-DM (Cross Industry Standard Process for Data Mining) for data analysis to answer business problems. The stages are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.

**1.2 Project Aim and Objective**

**Overall aim**

The purpose of this project is to build a Machine Learning model to classify spam or not spam.

**Project Objective**

Conduct a thorough error analysis to enhance the understanding of the limitations of the model among the stakeholders and boost their confidence in the model.

1. To organize the project through CRISP-DM.
2. To make a proposal about how to build the project through folder structure of a Git repository.
3. To evaluate the quality of provided data set and present the findings with visualization to the business partner.
4. To do pre-processing data, such as clean and filter the provided dataset, remove irrelevant data, and prepare the data to be ready for the next stage.
5. To vectorize dataset with Tf-Idf, which is turn texts into numbers, to make them understandable by the computer.
6. To use regularization. By including penalty terms in the loss function, regularization helps to avoid overfitting by preventing model from fitting too closely to the training data.
7. To use cross validation by folding the data into several smaller folds and selecting subsets as training and validation sets, cross validation contributes to the process of validating the model. Therefore, the model can be evaluated multiple times on smaller portions of data, to give better estimate of performance on unseen data.
8. To classify cleaned dataset as different label, either spam or ham (not spam). The selected best performing model will be used to predict labels in future new data.
9. To present and discuss the results of the analysis, including study limitations and error analysis to make business partner understand the weaknesses of the approach.
10. To propose daily work model integration through GUI (Graphical User Interface) proposal.
11. To provide a summary and conclusion of the study and suggestions for future steps in this area.

**2 Organize the Project**

**2.1 Folder Structure**

Below is the proposal of the folder structure of a Git repository for the project.

**Figure 1. Folder Structure Proposal**

Diagram

Description automatically generated

**Source: Own Representation**

The folder structure is used to help business partners to understand the project, and makes the project updatable and repeatable.

**2.2 Terminology and Definitions**

Terminology and definitions related to widely used machine learning algorithms for classification, such as Decision Tree, Logistic Regression, and Naive Bayes.  
  
**Decision Tree**

Decision Tree

## 

## Logistic Regression

Logistic regression is a statistical method which comparing the two sets of features to one another, to find the ideal parameters the coefficients (weight) that provide the best fit for the data. (Raschka, 2015)

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

**Figure 5. Logistic Regression**

Chart, line chart

Description automatically generated

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.

w is the weight vector, y is the target class, x is the feature, b is the bias term.

## Naïve Bayes

Naive Bayes is a probabilistic method based on Bayes' theorem. The Conditional probability equation determines text positivity.The key assumption is that each model attribute is independent and contributes equally. (Pati, & Pradhan, 2020, p. 7).

Formula: **P(Q|R) = P(Q)\*P(R|Q) / P(R),** where Q and R are events.

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

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

P(Q) and P(R) are the probabilities observing events Q & R

**Tokenization and** **Stopwords.**

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. (Raschka. (2015, p. 269).

**Stemming** is reducing words to their stem/ 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. (Mueller & Massaron, 2021, p. 355).

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

## 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 which assigns weight to each word in a text. (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) = Truly predicted as Positive.
* True Negative (TN) = Truly predicted as Negative.
* False Positive (FP) = Falsely predicted as Positive
* False Negative (FN) = Falsely predicted as Negative.

**Precision** is proportion of **correctly positive predictions** divided by the **total** number of **positive predictions**. Precision = TP/(TP+FP)

**Recall** is proportion of **correctly positive** **predictions** divided by the **total** number **of actual positive.** Recall(R) = TP/(TP+FN)

**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** is balancing precision and recall. The worst value is 0, and the best value is 1. 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.

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

## 3 Methodology

**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 2. Research methodology workflow:**

**Source: Own representation**

**3.2 Compare Performance** of Different Machine Learning Models

In a classification task, the performance of a model can be evaluated using a number of different metrics. Here is a brief explanation of some common evaluation metrics, such as Accuracy, Precision, Recall, and F1-Score.

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 with regard to human-machine relations.

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

**3.8 Expected Outcome**

* Researcher expects that logistic regression method will outperform other methods like Bayes and Linear SVC.
* Researcher believes that the sentiment 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 or not there will be dependencies between locations and human feelings about digital companions.

**4 Results and Findings  
4.1 Comparison of Model Performance on Labeled Dataset**

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





**Scoring Model Results**

A classification report was generated using a Linear SVC on a labeled dataset with 80000 rows for both training and testing purposes.

**Figure**



**Source: Own representation.**

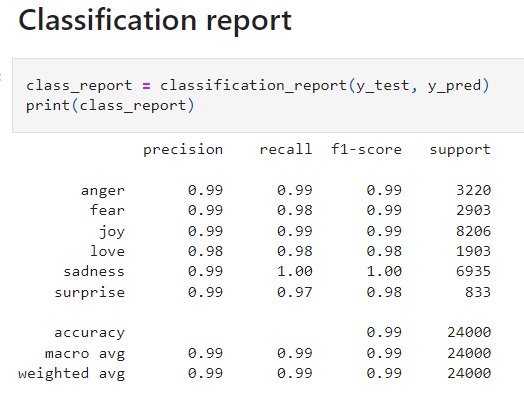
**The LinearSVC** model achieved an overall accuracy of 99%. This means that out of all the instances in the test set, the model **correctly** classified 99% of them.

**Table 3. Sentiment Analysis Model Accuracy**

|  |  |
| --- | --- |
| **Model** | **Accuracy Score** |
| BeroulliNB | 0.8300833333333333 |
| MultinomialNB | 0.9846666666666667 |
| Logistic Regression | 0.8886666666666667 |
| **Linear SVC** | **0.9922083333333334** |
| XGBoost | 0.9442083333333333 |

**Source: Own representation.**

**LinearSVC** model had the **highest accuracy score** among the models tested. Hence, we 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, using the snscrape Python library.



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 how often the model is right.

Precision is how often the model predict specific emotion, and it's **actually** that emotion.

Recall is how often the model predict all the times a specific emotion is there.

F1-score is how good is the model overall, mixed of precision and recall.

**For the "love" class**, the model’s precision is 98%, recall is 97%, and f1-score is 96%.

Precision interpretation: from the sample predicted as "love," 98% of them were **actually** "love," and 2% were actually not "love."

Recall interpretation: the model correctly classified 97% of the actual "love", and incorrectly classified only 3% of them.

The model looks to be performing really good, because it got a lot of correct predictions, with high scores for all three evaluations metrics, in all classes.

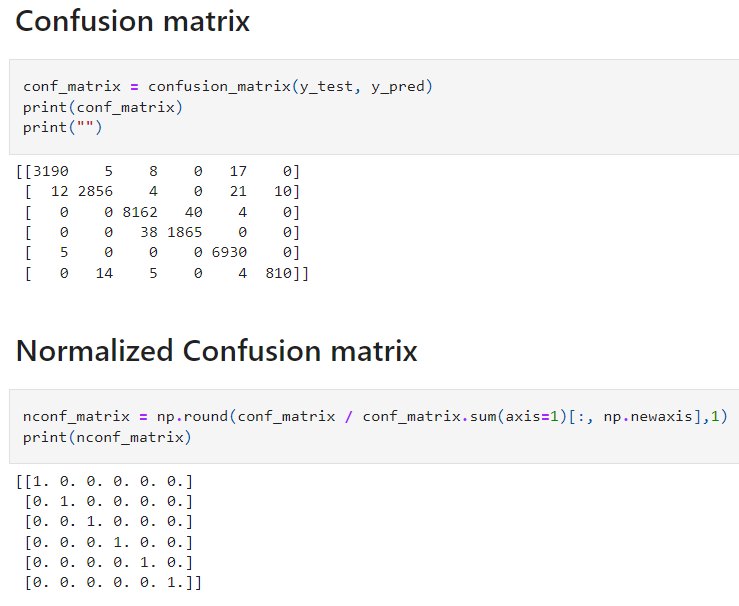


Figure 2. Heatmap of Normalized Confusion matrix

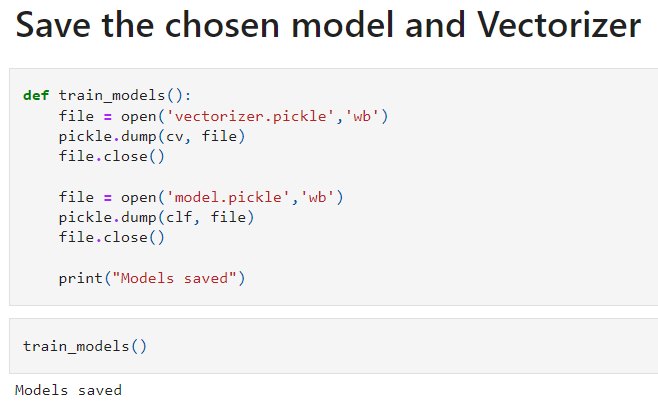
Source: Own representation.

Source: Own representation.

The model had some **difficulty with the "fear" class**, with the highest **21 incorrect predictions** out of 2879 total predictions for that class. The model performed the best on the "love" and "sadness" classes, with only 5 and 0 incorrect predictions out of 6935 and 6934 total predictions for those classes, respectively.

However, seems the model had a relatively high accuracy, with a total of 95 % out of 7200 total predictions in the test labeled dataset.

**Save model and vectorizer**



**Source: Own representation.**

**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.

Divide the number of positive thoughts (946) by the total number (946 + 256 + 603) and multiply by 100 to get a percentage, 51.1% of Americans are positive. To calculate each continent's positive (joy, love), neutral (surprise, sad), and negative (anger, fear) feelings, need to repeat this process.

America has the most favorable feelings (51.1%), followed by Europe (47.2%), Asia (53.3%), and Australia (42.9%). Since America and Asia have strong positive sentiments, AI robot digital companion company marketing have high possibility to be successful there.

The chi-square test findings show that America is associated with Asia and Australia, but not Europa.

America depends on Asia and Australia but not Europe. Europe and Asia have no major relationship, whereas Asia and Australia do. Asia depends on Australia, while Europa is autonomous.

The contingency table's independent variable is the continent, and the dependent variable is emotion. This suggests that the continent (America, Europe, Asia, or Australia) influences the emotion but not vice versa.

If America has more positive sentiments than Europe, this may be attributable to cultural differences, economic situations, or other continent-specific causes. The continent determines the emotion (positive or negative) in (America or Europe).

**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 the majority of people.
* The application of machine learning models for conducting sentiment analysis also comes with a number of 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.

Therefore, more study is needed in order 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.

**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

# References

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

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

# 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>

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

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

# 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.

# 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.

# 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>