

American International University-Bangladesh (AIUB)

# A Comparative Analysis On Instruction Detection From Text

# 

# 

***Abdur Rahman Emon (18-36603-1)***

***Md Shakibur Rahman(18-36598-1)***

***Gazi Ehsanul Haque(18-36572-1)***

***Nabila Akter Nowshi (17-35635-3)***

*A Thesis submitted for the degree of Bachelor of Science (BSc) in Computer Science and Engineering (CSE) at*

*American International University Bangladesh in July, 2022*

Faculty of Science and Technology (FST)

## Abstract

In today’s era we read various kinds of papers, articles, vlogs, magazines, text, reports and many readable papers. Day by day the resource and the techniques are getting more and more popular. But the main obstacles of our daily life is time. People are busier than before and our time is very limited. So, they want some shortcut in every way. They want to find the accurate summary and main instructions in a paper. It is especially seen between the undergrad students. The education system became open book. There is no specific syllabus for the students so if we want to find some instructions or steps on this internet site it would seem quite difficult.  But there are some tools and machines that change the perspectives of the literature review. So we have designed and trained a machine which can detect the instruction from a text. The trained data machine is specially designed to find the instruction, command, and request. The most popular way to extract the findings is NLP (Natural Language process). NLP is a machine learning language processing system which allows text analysis by using Artificial Intelligence. It’s a model of mathematical representation in which the input key is given a text. Unlike algorithm programming, a machine learning model is able to generalize and deal with novel classes. Machine learning for NLP and text analysis involves a set of statistical techniques for identifying parts of speech, entities, sentiment, and other aspects of text. Text data requires a special approach to machine learning. Because text data can handle thousand and thousand dimensions but tends to be sparse. However, the common approaches are KNN, CNN, SVM and many others.  We have done 4 algorithms in our python training code. All of the algorithms, logistic regression has achieved the highest number of successful results. The key result of our resource is to find our targeted instruction within a short amount of time. We have created our own dataset made by our group members. The core concentration of our resource was to minimize the time complexity and maximize the output. By using logistic regression this problem can be solved with a result of more than 76%. The significance of our main result was it was worth it to get a good result.

## Declaration by author

From the beginning of our thesis we have always become honest, sincere .We ensured that this work is completely our original work, no contents or material previously published or written by another person except where due reference has been made in the text. Thus this thesis done by various authors we have clearly mentioned the contribution as a whole, including conceptualization, formal analysis, investigation, methodology, Implementation, theoretical derivations, preparation of figures, writing original draft with review & editing and other research work used in our thesis.

We acknowledge we gathered permission from the copyright holder in where we have used their material.

## Approval

The thesis titled **“A comparative analysis on Instruction Detection From Text”** has been submitted to the following respected members of the board of examiners of the department of computer science in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science on **(date of defence)** and has been accepted as satisfactory.

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

**Tanvir Ahmed**

   Lecturer

Department of Computer Science

American International University-Bangladesh

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

**Farzana Bente Alam**

  Lecturer

Department of Computer Science

American International University-Bangladesh

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

**Dr. Md. Abdullah - Al - Jubair**

*Assistant Professor, Head-In-Charge (Undergraduate)*

Department of Computer Science

American International University-Bangladesh

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

**Professor Dr. Tafazzal Hossain**

*Dean*

Faculty of Science & Information Technology American International University-Bangladesh

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

**Dr. Carmen Z. Lamagna**

Vice Chancellor

American International University-Bangladesh

## Publications included in this thesis

No publications included

## Submitted manuscripts included in this thesis

No manuscripts included

## Other publications during candidature

No other publications included

## Research involving human or animal subjects

No animal or human subjects were involved in this research.

## Contributions by authors to the thesis

List the significant and substantial inputs made by different authors to this research, work and writing represented and/or reported in the thesis. These could include significant contributions to: the conception and design of the project; non-routine technical work; analysis and interpretation of research data; drafting significant parts of the work or critically revising it so as to contribute to the interpretation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Md Shakibur Rahman** | **Abdur Rahman Emon** | **Gazi Ehsanul Haque** | **Nabila Akter Nowshi** | **Contribution (%)** |
|  | *18-36598-1* | *18-36603-1* | *18-36572-1* | *17-35635-3* |
| Conceptualization | 25% | 25% | 25% | 25% | 100 % |
| Formal analysis | 30% | 20% | 20% | 30% | 100 % |
| Investigation | 20% | 30% | 30% | 20% | 100 % |
| Methodology | 30% | 30% | 20% | 20% | 100 % |
| Implementation | 30% | 20% | 25% | 20% | 100 % |
| Theoretical derivations | 30% | 20% | 30% | 20% | 100 % |
| Preparation of figures | 25% | 25% | 25% | 25% | 100 % |
| Writing – original draft | 25% | 25% | 25% | 25% | 100 % |
| Writing – review & editing | 25% | 25% | 25% | 25% | 100 % |

If your task breakdown requires further clarification, do so here. Do not exceed a single page.

## Acknowledgments

First of all, our honorable thesis supervisor Lecturer TANVIR AHMED from the Faculty of Science and Technology, American International University-Bangladesh, motivated us a lot from the very beginning of our thesis. As a result, this thesis has been done successfully. He has provided us the right direction from time to time about what we should do and the meetings has been conducted in a certain time period when we discussed our progress, challenges, and next goals. We would like to thank him for his commitment, precious advice, and supervision in this thesis. We are very grateful to Carmen Z. Lamagna, Vice-Chancellor, American International University-Bangladesh, for providing us with a beautiful study environment in our university. Finally, we are thankful to FARZANA BENTE ALAM, who guided us as an external thesis supervisor.

## Keywords

machine learning, deep learning, natural language processing, text classification, instruction detection, data mining

**Table of Contents**

[A Comparative Analysis On Instruction Detection From Text 1](#_Toc108598581)

[Abstract 2](#_Toc108598582)

[Declaration by author 3](#_Toc108598583)

[Approval 4](#_Toc108598584)

[Publications included in this thesis 5](#_Toc108598585)

[Submitted manuscripts included in this thesis 5](#_Toc108598586)

[Other publications during candidature 5](#_Toc108598587)

[Research involving human or animal subjects 5](#_Toc108598588)

[Contributions by authors to the thesis 6](#_Toc108598589)

[Acknowledgments 7](#_Toc108598590)

[Keywords 8](#_Toc108598591)

[List of Tables 11](#_Toc108598592)

[List of Abbreviations and Symbols 12](#_Toc108598593)

[Introduction 1](#_Toc108598594)

[1.1 Thesis topic 2](#_Toc108598595)

[1.2 History 2](#_Toc108598596)

[Literature review 4](#_Toc108598597)

[2.1 Introduction 4](#_Toc108598598)

[2.2 Impact 5](#_Toc108598599)

[2.3 Definition 5](#_Toc108598600)

[2.4 Sources 5](#_Toc108598601)

[Methods 7](#_Toc108598602)

[3.1 KNN algorithm 9](#_Toc108598603)

[3.2 Support Vector Machine (SVM) 10](#_Toc108598604)

[3.3 Naïve Bayes 12](#_Toc108598605)

[3.4 Logistic Regression 12](#_Toc108598606)

[Results or findings 14](#_Toc108598607)

[Discussion 15](#_Toc108598608)

[Conclusion 16](#_Toc108598609)

[Bibliography 17](#_Toc108598610)

**List of Figures**

Figure 1: Architecture of our system 8

Figure 2: Labeled texts 9

Figure 3: K-Nearest Neighbor 10

Figure 4: Support Vector Machine 11

Figure 5: Instructional Dataset 14



# List of Tables



Table 1: Accuracy of Models 14

# List of Abbreviations and Symbols





Abbreviations



KNN K Nearest Neighbor

SVM Support Vector Machine

NLP Natural Language Processing

*etc. etc.*



**Chapter 1**



# Introduction



An instruction text is a piece of non-fiction text that gives instructions on how to complete a task[1]. Also called an instructional text, or an instructive text. Instructional texts are associated with manuals. The instructional text is didactic and is based on the clarity of its exposition The goal is for the reader to comprehend the explanations and be able to use the knowledge without aid or help from others. Word recognition has been the main emphasis of reading teaching for the majority of American history, whether through more or less explicit phonics training or an indirect, inductive method [Joanna P. Williams, 2018]. In our daily life we hear different kinds of instruction even if everyone gives instructions to complete various tasks. People use social media like Facebook, twitter, Instagram, YouTube etc. Here we learn many things based on instruction. For example, a cooking recipe instruction makes us learn how to cook, a signal instructs whether we go right or left, during exams instructions are given how to answer questions, different kinds of manuals give instructions on how to complete specific tasks, instructions for taking medicine etc. Command, request also known as Instructive text. There are many instruction texts including various manuals or texts. System does not understand whether it is instructive or not. This research paper shows instructive text detection/extraction from text. Many of us did not get instructive text. The reason of this research is very simple .This paper shows how to detect instructions so that machines or people who apply these methods easily can understand whether it's instruction or not. The research question answered throughout this paper is what models or method used, how implementation was done and the necessity of this research.We use text classification methods,data mining ,text mining. The process of determining the class to which a text document belongs is called text classification. [Soucy, P., & Mineau, G. W. (2001, November)]. Different kinds of algorithm used for text classification like KNN, SVM [Y Lin, J Wang - 2014].Emotion recognition from text also used Natural language processing[S Shaheen, W El-Hajj, H Hajj -2014] .This paper also analysis NLP and various algorithm .This research provides full details on instruction detection also shows implementation accuracy benefits of doing this research.

## 1.1 Thesis topic

The thesis topic is based on Instruction detection from text. In this paper we try to cover and implement how to detect instructions using NLP and various kinds of algorithms. We use text classification models. We have instructions like everywhere, every work done based on instruction. Without instruction, completing tasks is not easy. So, if you want to complete a task perfectly you have to follow instructions. Before following instruction first, you need to find out which is the instruction from text. This research shows how easily a model finds instruction from text that’s the main goal of our topic.

## 1.2 History

The creation of computer techniques for natural language understanding has been one of the main objectives of AI from its inception.

With examples like translating the word pen correctly in "The box is in the pen" versus "The pen is in the box," early studies in machine translation demonstrated the challenges of this endeavor (Bar-Hillel 1964). It was soon realized that comprehension of language required knowledge of not just lexical and grammatical concepts but also semantic, pragmatic, and general world knowledge. However, in the 1970s, artificial intelligence (AI) systems were created that showed intriguing aspects of language understanding in constrained contexts, such as the blocks world (Winograd 1972), or responses to queries about a database of knowledge about moon rocks (Woods 1977), or even aircraft maintenance (Waltz 1978). The development of natural language systems utilizing manually coded symbolic grammars and knowledge bases continued throughout the 1980s (Allen 1987). Nevertheless, creating these systems remained challenging and necessitated a significant amount of domain-specific knowledge engineering. Additionally, the systems were fragile and could not perform well outside of the limited duties for which they were intended. In the 1950s, when behaviorism was flourishing in psychology (Skinner 1957) and information theory was just being introduced in electrical engineering, empirical and statistical analyses of natural language were previously popular (Shannon 1951). To develop an algorithmic and objective method for determining a language's structure, linguists researched techniques for automatically learning lexical and syntactic information from corpora.

The National Conference on Artificial Intelligence (AAAI-92), the International Joint Conference on Artificial Intelligence (IJCAI-91, IJCAI-95), the AAAI Fall Symposia (1992), the Annual Meeting of the Association for Computational Linguistics (ACL-93-ACL-96), and the International Conference on Computational Ling have all hosted numerous specialized workshops on topics related to empirical natural language processing in recent years. The First Conference on Empirical Natural Language Processing (EMNLP-96) took place at the University of Pennsylvania the previous year (Brill and Church 1996), and the Second Conference on EMNLP just took place in conjunction with AAAI-97.

**Chapter 2**



# Literature review



## 2.1 Introduction

Monitoring that identifies instructive behaviors and sends out notifications when they are found is known as an intrusion detection. Instructive text detection is finding whether a text is instructive or not .Serving to instruct of enlighten or inform is a kind of instructive text[D. Fried, R. Hu, V. CirikA. Rohrbach, J. Andreas, L.-P.  Morency,T.  Berg-Kirkpatrick, K.  Saenko, D.  Klein, and T. Darrell (2018)]. In an instructive text the texter deliver command or direction for doing something but many people cannot detect or catch instruction from text . This is probably because many of the readers fail to make the inferences. The objective of this research is to find out instruction-based text so that the reader can easily get if a text is instructive or not. People fail to inference text instruction. These readers are unable to get information from specific portions of the text because they do not link to those parts of the text. In this research we have extracted our core instructions from any input given as a text file. Then train machine to develop a sequence of instructions word by word and locate them from the provided text file. We have used text classification methods, data mining, text mining. We have implemented four algorithms here: Naive Bayes, logistic regression, SVM & KNN. We used Text classification in NLP. The field of research known as natural language processing (NLP) is huge. Tokenization of text is where NLP starts in the context of this study ([A.Pai](https://www.analyticsvidhya.com/blog/author/aravindpai/), 2020). The process of breaking the input text up into tokens is known as tokenization. According to Manning and Schutze, each token is often a component of a phrase that adds to its semantics. A text is made up of words, each of which might be considered a token. Text clarification is the process of categorizing the text into a group of words.

## 2.2 Impact

Research has shown thatText analytics can create reports, identify significant patterns, and offer other ways that businesses can make data-driven decisions. Natural language processing is a key technology used by text mining and text analytics (Garner, R. 1986). Analyzing text properly requires specialized solutions. Text analytics technology introduces an automated approach to analyzing and visualizing unstructured text data for qualitative measurements (Tanner-Smith, E. E., & Tipton, E.,2014).text analysis helps gaining actionable insights from every social media post, email, chat message, trouble ticket, and survey. Text analytics allows to learn more about what people are talking about, thinking, and feeling when engaging with your products and services.

## 2.3 Definition

The text analytics process starts by gathering large text data sets. Depending on the scope of research and the resources that are available. Reading is one of the primary ways people are introduced to new information. Authors of informational text often use devices to present and organize information in ways that will help their readers understand it better. Text structure is one such device (Kintsch, 1974;Meyer, 1975). Knowing how authors decided to structure a text may provide readers with valuable information about how to approach the text and assist them in identifying important information to remember from the text. Meyer (1985). Text analytics delivers many advantages to people, as it’s a critical part of extracting meaning from the unstructured data sets that are otherwise unable to process.

## 2.4 Sources

Using NLP, text classification can automatically analyze text and then assign a set of predefined tags or categories based on its context. We have used text classification algorithms to find text instructions.

Logistic Regression: Logistic regression is a calculation used to predict a binary outcome: either something happens, or does not. This can be exhibited either Yes or No. Using this algorithm, it is inferenced if the text is instructional or not

Naive Bayes: It calculates the possibility of whether a data point belongs within a certain category or does not. In text analysis, it is used to categorize words as belonging to a preset tag (classification) or not.

SVM: A support vector machine (SVM) uses algorithms to train and classify data

KNN: In this is a pattern recognition algorithm that uses training datasets to find the k closest relatives.

**Chapter 3**



# Methods



Over the past few decades, text categorization issues have received much study and have been addressed in numerous practical applications. Many researchers are increasingly interested in creating applications that make use of text categorization techniques, particularly in light of recent advances in Natural Language Processing (NLP) and text mining. It is possible to break down the majority of text classification and document categorization systems into the four steps of feature extraction, dimension reduction, classifier selection, and evaluations.

Machine learning is an approach that allows computers to learn from empirical data that has been approved by professionals [Subroto IM, Selamat A. 2014]. Machine learning is an application (subfield) of artificial intelligence that gives computers the capacity to learn on their own and get better with time without external programming. Machine learning task is divided into supervised learning and unsupervised learning. Supervised learning is a type of machine learning that uses labeled data to train machine learning models. In labeled data, the output is already known. The model just needs to map the inputs to the respective outputs. Supervised Learning methods need external supervision to train machine learning models. Hence, the name supervised. They need guidance and additional information to return the desired result. Models that can predict labels based on labeled training data. Supervised machine learning has two parts classification and regression. In Unsupervised Learning, the machine uses unlabeled data and learns on itself without any supervision. The machine tries to find a pattern in the unlabeled data and gives a response. That means, it uses unlabeled data to train machines. Unlabeled data doesn’t have a fixed output variable. The model learns from the data, discovers the patterns and features in the data, and returns the output. Unsupervised learning finds patterns and understands the trends in the data to discover the output. So, the model tries to label the data based on the features of the input data. One of the most common methods/tasks of unsupervised learning is clustering.

In this research we used supervised machine learning. Text classification falls under supervised machine learning. In supervised machine learning system, the system is trained with examples for which class labels are known. After the training, the system is tested on a data set with unknown class labels. The architecture of our system is shown in Figure 1. At first, we prepare our dataset to make it suitable for training and testing data. In figure we have that module, in this module data are processed so that they are ready to feed into the classifier algorithm. We make instructional dataset manually where we insert various instruction. After that we manually label the instructions. Then we apply preprocessing. Preprocessing is a crucial stage in classification since it has a significant impact on how accurately a system uses natural language [Ruhwinaningsih L, Djatna T. 2016]. The initial stage of text preprocessing is tokenization. Tokenization involves breaking the text down into a stream of tokens, which can be words, phrases, symbols, or other significant units. Tokenization may remove some symbols like punctuation in addition to breaking text into tokens.

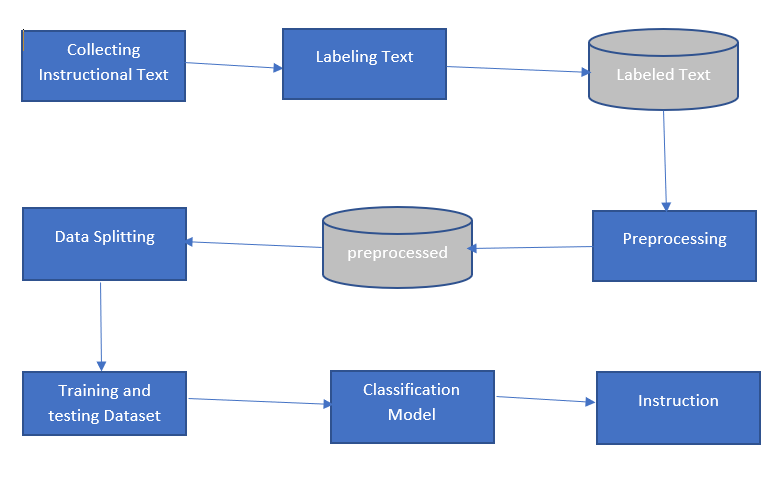


Figure 1: Architecture of our system

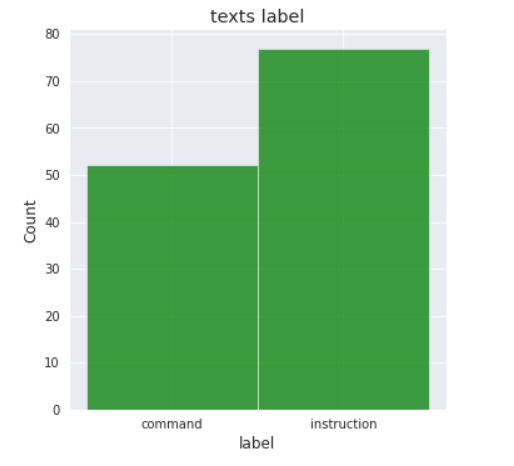


Figure 2: Labeled texts

The main task instruction extraction done by this system. We split data for two purpose one is training another one is testing data. After training and testing we create text classification model. Four classification model is used in this research KNN, SVM, Naïve Bayes, Logistic Regression. After train these classification models we can extraction our desire instruction. These algorithm structures are given below:

## 3.1 KNN algorithm

KNN is a non-parametric algorithm, which means it does not rely on underlying data to make assumptions. It is also referred to as a lazy learner algorithm since it saves the training set before performing an action on it during classification rather than instantly learning from it.

One of the most significant non-parameter algorithms in the field of pattern recognition is KNN, which is a supervised learning predictable classification method [Belur V, Dasarathy, 1991]. KNN training samples alone, without the aid of new data, develop the classification rules. KNN classification method predicts the test sample’s category according to the K training samples which are the nearest neighbors to the test sample, and judge it to the category which has the largest category probability. Steps for performing KNN algorithm:

* Step 1: Start by selecting neighbor number K.
* Step 2: Select the K-nearest neighbors of the new data point based on the Euclidean distance.
* Step 3: Count the number of data points among the K neighbors in each category.
* •Step 4: Assign the new data point where you counted the most neighbors after you've finished counting.

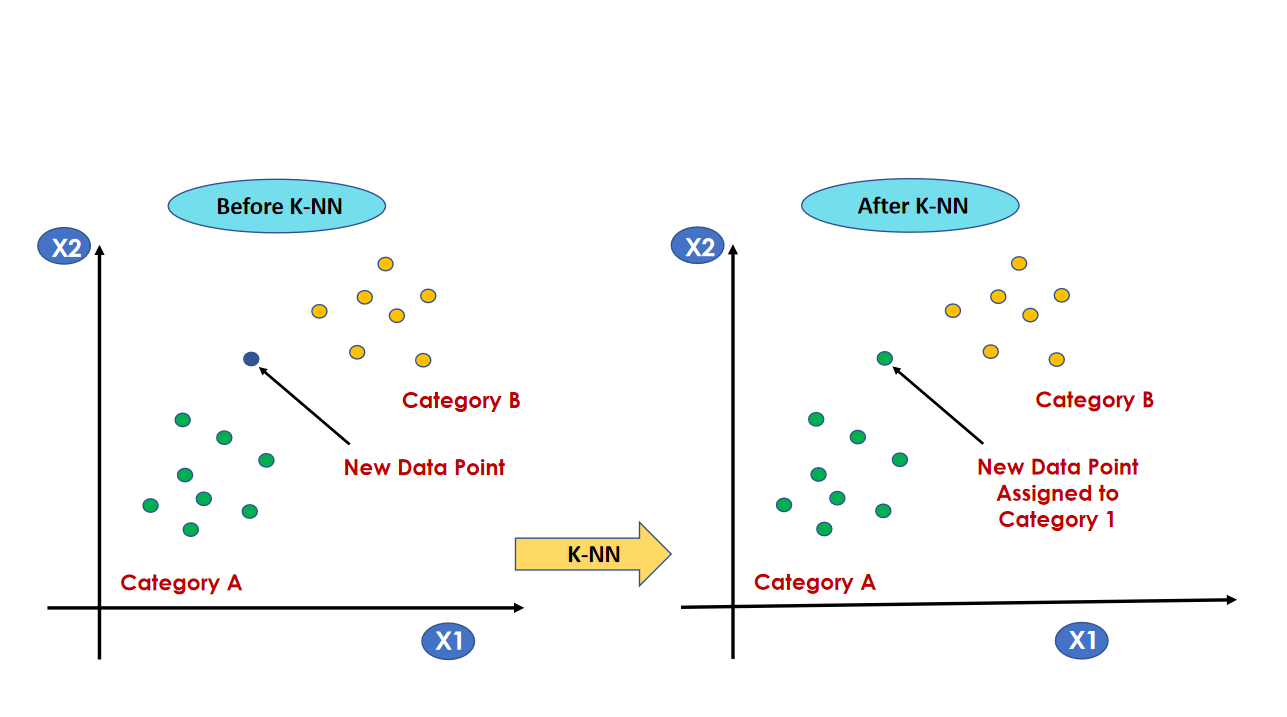


Figure 3: K-Nearest Neighbor

After splitting data for training and testing we train KNN classification model. This model accuracy is 53%.

## 3.2 Support Vector Machine (SVM)

A fundamental and well-known technique for completing machine learning tasks is the Support Vector Machine (SVM). Based on the structural risk minimization concept from computational learning theory, SVM is a potent supervised learning paradigm. It is a cutting-edge classification algorithm that has a reputation for being effective in a wide range of applications. The method's great generalization capacity makes it especially well-suited for high dimensional data, like text. [Wang, Z. Q., Sun, X., Zhang, D. X., and Li (2006)]. When used for text classification, support vector machines (SVM) offer good precision but low recall. One way to boost recall using SVM customization is to change the threshold that goes with the SVM. There are two different segments that are classified using a decision boundary or hyperplane.

**Linear SVM**:  These are frequently suggested for text classification because the majority of these kinds of classification issues can be solved linearly.

When there are numerous instances, characteristics, and text classifications, the linear kernel performs exceptionally well. The many properties of linear kernel functions make them smoother, quicker, and more flexible in terms of the parameters that may be optimized.

The following function identifies the linear kernel:

f(X) = w^T \* X + b

**Non-Linear SVM**: Due to its lower processing efficiency and less precise predictions compared to other kernels, the polynomial kernel isn't frequently utilized in practice.

Here's the function for a polynomial kernel:

f(X1, X2) = (a + X1^T \* X2) ^ b

This is one of the simpler polynomial kernel equations you can use. **f(X1, X2)** represents the polynomial decision boundary that will separate your data. **X1** and **X2** represent our dataset.

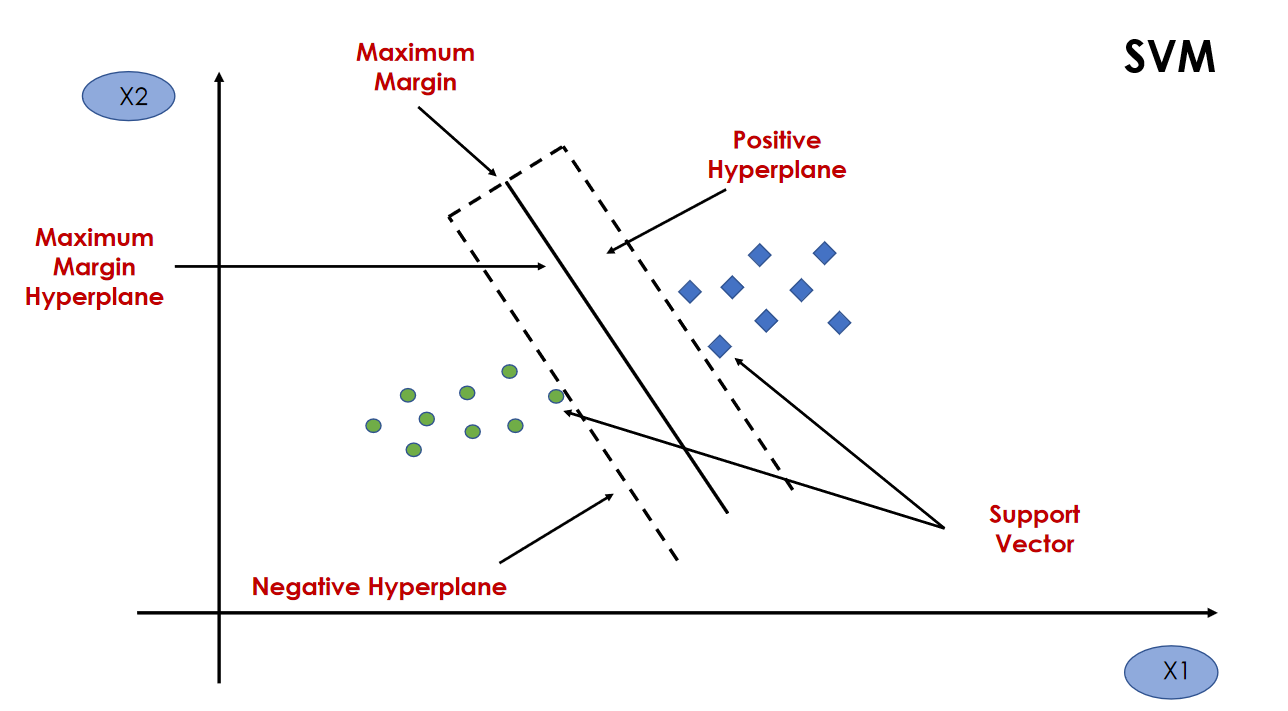


Figure 4: Support Vector Machine

Accuracy of this model is 57%.

## 3.3 Naïve Bayes

Naïve Bayes is a straightforward learning algorithm that makes use of the Bayes rule and the fundamental presumption that, given the class, the attributes are conditionally independent. Despite the fact that in practice this independence assumption is frequently broken, naive Bayes frequently produces competitive classification accuracy. The Naive Bayes Classifier is one of the most straightforward and efficient classification algorithms available today. It aids in the development of quick machine learning models capable of making accurate predictions. Being a probabilistic classifier, it makes predictions based on the likelihood that an object will occur. This approach for categorization works well for both binary and multiclass classification. Compared to numerical input variables, naive Bayes performs better in cases of categorical input variables. It is helpful for anticipating data and making predictions based on past outcomes.

A type of classifier that makes use of the Bayes Theorem is naive Bayes. It forecasts membership probabilities for each class, such as the likelihood that a specific record or piece of data falls under a specific class. The most likely class is that which has the highest likelihood. A group of supervised learning algorithms known as naive Bayes methods utilize Bayes' theorem with the "naive" assumption that each pair of features is conditionally independent given the value of the class variable.

Accuracy of this model is 69%.

## 3.4 Logistic Regression

The approach for supervised classification includes logistic regression. Recent years have seen an increase in the use and importance of this method. According to the logistic function, individuals are classified using this algorithm [Liu YY, Yang M, Ramsay M, Li XS, Coid JW (2011)].A statistical analysis method called logistic regression uses previous observations from a data set to predict a binary outcome, such as yes or no. By examining the correlation between one or more already present independent variables, a logistic regression model forecasts a dependent data variable. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Because the result is a probability, the dependent variable's range is limited to 0 and 1.

In logistic regression, the odds—that is, the probability of success divided by the probability of failure—are transformed using the logit formula. This logistic function is expressed by the following formulas and is also frequently referred to as the log odds or the natural logarithm of odds:

Logit(pi) = 1/ (1+ exp(-pi))

ln(pi/(1-pi)) = Beta\_0 + Beta\_1\*X\_1 + … + B\_k\*K\_k

Logit(pi) is the dependent or response variable in this logistic regression equation while x is the independent variable. The most frequent method for estimating the beta parameter, or coefficient, in this model is maximum likelihood estimation (MLE). In order to find the best fit for the log odds, this approach iteratively evaluates various beta values. The log likelihood function is created after each of these iterations, and logistic regression aims to maximize this function to get the most accurate parameter estimate. The conditional probabilities for each observation can be calculated, logged, and added together to produce a forecast probability once the best coefficient (or coefficients, if there are multiple independent variables) has been identified. If the categorization is binary, a probability of less than .5 predicts 0 and a probability of more than 0 predicts 1. It is recommended to assess the model's goodness of fit, or how well it predicts the dependent variable, once the model has been computed. A common technique for evaluating model fit is the Hosmer-Lemeshow test [12].

Accuracy of this model is 76%.

**Chapter 4**

# Results or findings

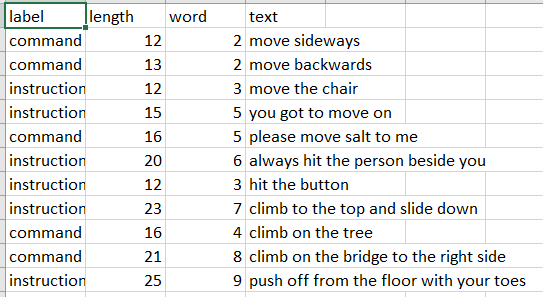


Figure 5: Instructional Dataset

The dataset in the above we make manually. We collected various instruction and make this dataset. Then we labeled this data after that we select four algorithms to train this dataset. We applied KNN, Naïve Bayes, SVM, Logistic Regression. The result we got from this model shown below:

|  |  |
| --- | --- |
| MODELS | ACCURACY |
| KNN | 53% |
| Naïve Bayes | 69% |
| SVM | 57% |
| Logistic Regression | 76% |

Table 1: Accuracy of Models

Here in the table shown that after training testing through these models we can finally say that Logistic Regression got better accuracy than other models.

**Chapter 5**



# Discussion



We have used four different methods on our algorithm. Among four different methods logistic regressions have the best output in our trained machine. However, there are some limitations also. Since we have used our own made dataset and our dataset is not too huge like other datasets there, we can get some unusual results. We have trained our dataset machine model based on the instructions and there are not too many prepared datasets exists in this field. So, we had to make our own datasets. The problems with making our own datasets are inaccuracy, inefficiency, system failure, loss of systems. If we work on a small dataset there exists some inefficiency such as we couldn’t get the accurate results and finds will be tempered. If small datasets all others limitations are ignored yet it will create some inaccuracy to the system. The machine and coding language we have used on they are python-based machine and NLP natural learning process has some basic limitations too in the system. Since python is an open-source platform and the code is pretty fast. Yet while working on a dataset it creates some difficulties and the machine gets slower where NLP will execute. We could use some more data to get the perfect accuracy of our trained machine. The next challenge of our data sets and training machine was we only have used four modules or methods. If we could use some more modules or methods, we could get some more better results and we can get a better accuracy. Finally, the biggest challenge was our dataset was handmade and biased. If we use some random datasets, we could get a better solution of our system. Every machine has some advantages and disadvantages as well. We could ignore the disadvantages and can focus our advantages and further work which will lead us to the future work procedure.

**Chapter 6**



# Conclusion



We can conclude that our research was successful based on all of the strategies we have described thus far and the techniques we have employed. One might utilize them for both personal and professional usage by adhering to the appropriate context and necessity. We have achieved text-based instruction detection in this paper. We suggest using a machine learning model based on a distinct dataset that has been trained to analyze text. We applied four distinct techniques to an algorithm that predicts instructions from text using natural language using a data set that we created our own. As we created our own dataset and it is not as large as other datasets, we can obtain some unexpected findings. There aren't many preset datasets available in this sector so we have to made our own for training our machine. Future research is needed to explore more algorithmic method.  Four different methods were utilized to extract instructions from text. All of the approaches presented here have different capabilities depending on the situation. More or other algorithms may be investigated in order to provide more accurate results for future work.

Therefore, there is a great opportunity to develop an AI-based solution that can actually derive instructions from the text by learning all the features alone.

# Bibliography

1. twinkl.com 'What is an Instruction Text?'[online]. Available: https://www.twinkl.com/teaching-wiki/instruction-text.
2. whatdoesmean.net 'What is instructional text?'[online].Available: https://whatdoesmean.net/what-is-instructional-text/
3. Shaheen, S., El-Hajj, W., Hajj, H., & Elbassuoni, S. (2014, December). Emotion recognition from text based on automatically generated rules. In 2014 IEEE International Conference on Data Mining Workshop (pp. 383-392). IEEE.
4. Williams, J. P. (2018). Text structure instruction: the research is moving forward. Reading and Writing, 31(9), 1923-1935.
5. Lin, Y., & Wang, J. (2014, June). Research on text classification based on SVM-KNN. In 2014 IEEE 5th International Conference on Software Engineering and Service Science (pp. 842-844). IEEE.
6. Soucy, P., & Mineau, G. W. (2001, November). A simple KNN algorithm for text categorization. In Proceedings 2001 IEEE international conference on data mining (pp. 647-648). IEEE.
7. Subroto IM, Selamat A. Plagiarism Detection through Internet using Hybrid Artificial Neural Network and Support Vectors Machine. TELKOMNIKA (Telecommunication Computing Electronics and Control). 2014; 12(1): 209-18.
8. Ruhwinaningsih L, Djatna T. A Sentiment Knowledge Discovery Model in Twitter’s TV Content Using Stochastic Gradient Descent Algorithm. TELKOMNIKA (Telecommunication Computing Electronics and Control). 2016 Sep 1; 14(3):1067-76.
9. Belur V, Dasarathy, “Nearest Neighbor (NN) Norms˖ NN Pattern Classification Techniques”, Mc Graw-Hill Computer Science Series, IEEE Computer Society Press, Las Alamitos, California, 1991,pp.217-224
10. Wang, Z. Q., Sun, X., Zhang, D. X., & Li, X. (2006, August). An optimal SVM-based text classification algorithm. In 2006 International Conference on Machine Learning and Cybernetics (pp. 1378-1381). IEEE.
11. Liu YY, Yang M, Ramsay M, Li XS, Coid JW (2011) A comparison of logistic regression, classification and regression tree,and neural networks models in predicting violent re-offending.J Quant Criminol 27(4):547–553
12. ibm.com, 'What is logistic regression'[online] Available:https://www.ibm.com/topics/logistic-regression.
13. D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency,T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell, “Speaker-follower models for vision-and-language navigation,” inAdvances inNeural Information Processing Systems, 2018, pp. 3314–3325.
14. Kintsch, W. (1974).The representation of meaning in memory. Hillsdale,NJ: Erlbaum.
15. Meyer, B. J. F. (1985). Prose analysis: Purposes, procedures, and problems.In B. K. Britton & J. Black (Eds.),Understanding expository text: Atheoretical and practical handbook for analyzing explanatory text(pp.269–304). Hillsdale, NJ: Erlbaum.
16. Meyer, B. J. F. (1975).The organization of prose and its effects onmemory. Amsterdam, Netherlands: North-Holland Publishing.
17. Garner, R., Alexander, P., Slater, W., Hare, V. C., Smith, T., & Reis, R.(1986). Children’s knowledge of structural properties of expository text.Journal of Educational Psychology, 78,411–416.http://dx.doi.org/10.1037/0022-0663.78.6.411
18. Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation withdependent effect sizes: Practical considerations including a softwaretutorial in Stata and SPSS.Research Synthesis Methods, 5,13–30.http://dx.doi.org/10.1002/jrsm.1091
19. Renu, R. S., & Mocko, G. (2015, August). Text analysis of assembly work instructions. In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 57052, p. V01BT02A003). American Society of Mechanical Engineers.
20. Venkatesh, S. G., Biswas, A., Upadrashta, R., Srinivasan, V., Talukdar, P., & Amrutur, B. (2021, May). Spatial reasoning from natural language instructions for robot manipulation. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 11196-11202). IEEE.
21. Bar-Hillel, Y. 1964. Language and Information. Reading, Mass.: Addison-Wesley.
22. Winograd, T. 1972. Understanding Natural Language. San Diego, Calif.: Academic.
23. Woods, W. A. 1977. Lunar Rocks in Natural English: Explorations in Natural Language Question Answering. In Linguistic Structures Processing, ed. A. Zampoli. New York: Elsevier.
24. Waltz, D. L. 1978. An English Language Question Answering System for a Large Relational Database. Communications of the Association for Computing Machinery 21(7): 526–539.
25. Allen, J. F. 1987. Natural Language Understanding. Menlo Park, Calif.: Benjamin/Cummings. Anderson, J. R. 1977. Induction of Augmented Transition Networks. Cognitive Science 1:125–157.
26. Skinner, B. F. 1957. Verbal Behavior. New York: Appleton-Century-Crofts.
27. Shannon, C. 1951. Prediction and Entropy of Printed English. Bell Systems Technical Journal 30:50–64. Simmons, R. F., and Yu, Y. 1992. The Acquisition and Use of Context-Dependent Grammars for English. Computational Linguistics 18(4): 391– 418.