**Chapter II**

**Literature Review**

Automated news categorization is news categorizer that use machine learning to categorize news and deploy as data virtualization on a website.

* Definition
* Machine Learning
* Information of software/hardware used in development of a system
* Related works

**2.1 Definition**

**2.1.1 News categorization**

The process of grouping news into it’s categories, for example sport news should be in a group of news about sport or politics news should be in a group of politics news.

**2.1.2 Automated news categorizer**

News categorizer that use machine learning algorithm like classification to categorize news into it’s categories automatically. The machine learning model will be naive bayes to be trained on training data set that is news articles from online news.

**2.1.3 Data Visualization**

The process that is used to visualize data. This process will be applied to illustrate the data that is news categories that has been categorized by machine learning. The project will use technique of data visualization to summarize the data of news and deploy on a web application.

**2.2 Machine Learning**

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs** that can access data and use it to learn for themselves.

**2.2.1 Classification**

Classification is supervised algorithm that learn from the given data to make new classification into its group. The project will use classification algorithm to classify news into its categories.

**2.2.2 Naive bayes classifier**

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes’ Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

**2.3 Tools**

**2.3.1 Code editor**

**2.3.1.1 Jupyter Notebook**

Jupyter Notebook is one of the most popular tools in the field of data science, which involves a lot of data management work. And still have to report research. Jupyter Notebook has been designed to meet the purpose of use, whether it is Access the library and write the code and see the results. Jupyter Notebook is designed to be more functional and readable than a normal program.

**2.3.2 Data Mining Libraries**

The libraries that help That control in the field of data management. Used to retrieve data from website and modify data.

**2.3.2.1 Requests**

Requests is a [Python](https://en.wikipedia.org/wiki/Python_(programming_language)" \o "Python (programming language)) HTTP [library](https://en.wikipedia.org/wiki/Software_library" \o "Software library), released under the [Apache License 2.0](https://en.wikipedia.org/wiki/Apache_License" \o "Apache License). The goal of the project is to make HTTP requests simpler and more human-friendly. The current version is 2.25.0

**2.3.2.2 Beautiful Soup 4**

Beautiful Soup is a library that makes it easy to scrape information from web pages. It sits atop an HTML or XML parser, providing Pythonic idioms for iterating, searching, and modifying the parse tree

**2.3.2.3 Numpy**

**\**

NumPy is a Python library that provides a simple yet powerful data structure: the **n-dimensional array**. This is the foundation on which almost all the power of Python’s data science toolkit is built.

**2.3.3 Machine Learning**

**2.3.3.1 Scikit-learn**

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

**2.3.4 Data Visualization**

The tools that is used to visualize data

**2.3.4.1 Django**

A [Python](https://en.wikipedia.org/wiki/Python_(programming_language)" \o "Python (programming language))-based [free and open-source](https://en.wikipedia.org/wiki/Free_and_open-source_software" \o "Free and open-source software) [web framework](https://en.wikipedia.org/wiki/Web_framework" \o "Web framework) that follows the model-template-views (MTV) [architectural pattern](https://en.wikipedia.org/wiki/Architectural_pattern_(computer_science)" \o "Architectural pattern (computer science)). The purpose of this frame work is to built website with python.

**2.3.4.2 Matplotlib**

a plotting library for the **Python** programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

**2.3.4.3 Seaborn**

A Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**2.4 Related Works**

There are several paper researches which related with news categorization using machine learning. The technique can be related to the project;

**2.4.1 AUTOMATIC SEMANTIC CATEGORIZATION OF NEWS HEADLINES USING ENSEMBLE MACHINE LEARNING: A COMPARATIVE STUDY.**

This project is about categorize news from news headline. Using machine learning to built-up application that automatically categorizing news into its categories by using machine learning model like SVM ( Support Vector Machine )

**2.4.2 AUTOMATIC TEXT TAGGING OF ARABIC NEWS**

**ARTICLES USING ENSEMBLE DEEP LEARNING MODELS**

This project is about categorize news in Arabic language by using deep learning.

**Conclusion**

Due to the focus on mobility and the internet in the recent years and to reduce the paper waste, many news companies went online and changed the traditional way of printing newspapers and articles. Because of that there‟s a huge number of different articles in the news website databases. However, categorizing news article in its respective category manually is very difficult and time consuming. Automatic categorization of the news corpus will profit society in several ways.