AUTOMATED NEWS CATEGORIZER

MARWAN YEEDENG

NIHENG MAE

FATONI UNIVERSITY

1442/2021

AUTOMATED NEWS CATEGORIZER

MARWAN YEEDENG

611431016

NIHENG MAE

601431012

FATONI UNIVERSITY

1442/2021

FATONI UNIVERSITY

FACULTY OF SCIENCE AND TECHNOLOGY

DEAPARMENT OF INFORMATION TECHNOLOGY

TITLE

AUTOMATED NEWS CATEGORIZER

PRESENT BY

MARWAN YEEDENG

611431016

NIHENG MAE

601431012

…………..………..………. ADVISOR

(MR. KHOLED LANGSARI)

DATE: …. /…. /1442

…. /…. /2021

DEPARMENT OF INFORMATION TECHNOLOGY APPROVES THIS

PROJECT REPORT AS PARTIAL FULFILMENT OF THE REQUIREMENT

FOR THE DEGREE OF BACHERLOR OF INFORMATION TECHNOLOGY

ACADEMIC YEAR 1436/2015

…………..………..………. HEAD DEPARTMENT

(MR. FAUZAN MAPA)

DATE: …. /…. /1442

…. /…. /2021

……….………..………..………..…….……….

(MR. SOBREE HAYEEMAD)

DATE: …. /…. /1442

…. /…. /202

Title: Automated News categorizer

Authors: Marwan Yeedeng 611431016, Niheng Mae 601431012

Department: Information Technology

Academic year: 2021

ABSTARCT

Nowadays, there are massive amount of data on the internet. The internet represents the contents delivery. People use internet to send data to one another. News agencies use deliver contents of news over the internet to people via web application. People read news over website a lot. News has been sent over the internet every day. With this massive amount of data will make a big problem that is to take a big amount of time to categorize. With this problem we found the solution is that to use machine learning to classify news automatically. Our project will use Naive Bayes machine learning to classify news. Once after classification. The data will be sent into SQL database. The output of the project is to visualize data on the website that we received from the database. The visualization will be in form of Bubbles visualization to visualize the amount of data.

หัวข้อ:

ผู้เขียน:

สาขา:

ปีการศึกษา: 2564

บทคัดย่อ

การเปรียบว่า “ชีวิตเหมือนการเดินทาง” เป็นอะไรที่เหมาะสมดี ซึ่งในคำว่า “การเดินทาง” นี้ ก็มีอีกหลากหลายแง่มุม จึงมักมี คติ คำคม หรือประโยคเปรียบเปรยต่าง ๆ ให้คิด ให้ทบทวน บทความสั้น เรื่องนี้ก็เช่นกัน

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

1.0. PROJECT OVERVIEW

Due to widespread availability of Internet, there are a huge resource that produce massive amount of daily news. Every day, people read a lot of news. Wither in social media like Twitter, Facebook or online news website. The news has been delivered to the people with different resources and even worse the news is represented differently or not talking about the same thing.

Because of this, it makes people being confused that what exactly happened lately? This makes a lot of people having no idea what are mostly people of the news throughout a month?

Unfortunately, the most of situations that happened in the country or in the world have been announced through news. Whenever people miss the news. They seem to be unprepared for the situations. This is why people should have known about what is the trend of the news in their daily life.

From the problem. We find out that it can be solved by using the technique of data mining to find out what is the trend of the news of each month? By using NLP (Natural Language Processing) to process the human language that is been written in news articles along with Naïve Bayes to categorize news into categories and summarize by using the technique of data visualization on a web application. The summarization will be about what is the trend of news of each month?

1.1. PROBLEM STATEMENTS

The main problem to the people who are reading online news are they cannot keep up with the news because there many online news sources or the internet and another problem is, they will miss up the news because there are variety of online news article on the internet like politics, economy, sport and so on. For those follow the only sport news they might up technology news or the others. Form those problem may cause the unexpected and unprepared situation problem. For instance, people have missed up the technology news. Once the technology changes the world. They will be unprepared for the situation.

From the problem above. The solution is having a news visualization to summaries the news form online news article. The output will be a web application that visualize trend of news of each month. The news will be shown in form of timeline visualization. The web application use news categorization with machine learning to categorize into its categories and visualize on a web application as a grouplike politics, sport, economy and so on.

1.2. OBJECTIVE

- To use the machine learning to categorize news and deployment by using technique of data visualization and data storytelling.
- To visualize the categorized news into news timeline visualization in form of data storytelling on a web application.
- QQQ

1.3. SIGNIFICANCE OF STUDY

1.3.1. DEVELOPER

- To gets technique skill of data analysis in term of using machine learning to apply with news categorizing.
- To deployment the result on web application by using python.
- QQQ

1.3.2. USER

- User can follow the trend of news are viral in social during that time.
- User gets quick access to relevant news and topics of interest.
- QQQ

1.4. SCOPE

The scope of the project is to have a web application that summarize the trend of news of each month in form of data visualization the visualization will use date storytelling technique to make beautiful data visualization and also the visualization can be interactive to users the form of visualization will can be a timeline of each month with describe the trend of news since the first data of the month to the last data of the month. The implementation of this project is useful for anyone. They will know the situation that are happening so that they will be prepared when the situation happened.

1.5. SIGNIFICANCE OF STUDY

1.5.1. SOFTWARE REQUIREMENT

1.5.1.1. Resource

- Bangkok Post Agency Website

1.5.1.2. Code Editor

- Jupyter Notebook
- Weka
- Visual Studio Code
- GitHub

1.5.1.3. Programing Language

- Python
- HTML
- JavaScript

- CSS

1.5.1.4. Python Libraries

- Requests
- Beautiful Soup 4
- Pandas
- NumPy
- re (Regular expression)
- Matplotlib
- Seaborn

1.5.1.5. Web Framework

- Bootstrap
- Django

1.5.1.6. Data Base

- MySQL

1.5.1.7. QQQ

1.5.2. HARDWARE REQUIREME

1.5.2.1. Personal Computer

- Asus VivoBook 15 x512da
- HP Pavilion Power 15-cb035TX

1.6. CONCLUSION

QQQ

CHAPTER TWO

LITERATURE REVIEW

2.0. INTRODUCTION

2.1. DEFINITION

2.1.1. NEWS CATEGORIZATION

The task of automatically classify the news documents into their predefined classes based on their content with the confidence learned from the training news dataset.

2.1.2. AUTOMATED NEWS CATEGORIZER

News categorizer that uses machine learning algorithm like classification to categorize news into its 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.

2.1.4. 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.

2.1.4.1. 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.

2.1.4.2. NAÏVE 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.

2.2. TOOLS USED

2.2.1. Tools used Related 1

2.2.2. Tools used Related 2

2.2.3. Tools used Related 3

2.3. RELATED WORK

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

Due to widespread availability of Internet, there are a huge of sources that produce massive amounts of daily news. Moreover, the need for information by users has been increasing unprecedently, so it is critical that the news is automatically classified to permit users to access the required news instantly and effectively. One of the major problems with online news sets is the categorization of the vast number news and articles. In order to solve this problem, the machine learning model along with the Natural Language Processing (NLP) is widely used for automatic news classification to categorize topics of untracked news and individual opinion based on the user's prior interests. However, the existing studies mostly rely on NLP but uses huge documents to train the prediction model, thus it is hard to classify a short text without using semantics. Few studies focus on exploring classifying the news headlines using the semantics. Therefore, this paper attempts to use semantics and ensemble learning to improve the short text classification. The proposed methodology starts with preprocessing stage then applying feature engineering using word2vec with TF-IDF vectorizer. Afterwards, the classification model was developed with different classifier KNN, SVM, Naïve Bayes and Gradient boosting. The experimental results verify that Multinomial Naïve Bayes shows the best performance with an accuracy of 90.12% and recall 90%. (Bogery et al. 2019)

2.3.2. A COMPARATIVE ANALYSIS OF NEWS CATEGORIZATION USING MACHINE LEARNING APPROACHES

The rapid growth of print and digital media increased the reach of one and all in terms of information, resulting in more amount of Text data to be mined. This data is nothing but a heap of unclassified information which when kept together means nothing. This means that there is a need to tag all these data i.e., News Classification. News classification is the task of automatically classify the news documents into their predefined classes based on their content with the confidence learned from the training news dataset. This research evaluates some most widely used machine learning techniques, mainly Naive Bayes, Random Forest, Decision Tree, SVM and Neural Networks, for automatic news classification problem. To experiment the system, a dataset from BBC that have two columns, one has the news headlines and the other contains the type it belongs to. There are 2225 rows in the data set is used. The average results show that the Naive Bayes is performing better than the other four algorithms with the classification accuracy of 96.8 %. Then follows the Random Forest with accuracy 94.1 %, Support Vector Machine (SVM) with accuracy 96.4 %, Neural Networks with accuracy 96.4 % and the Decision Tree with accuracy 83.2 %. (Deb et al. 2020)

2.3.3. TEXT MINING APPROACH TO CLASSIFY TECHNICAL RESEARCH DOCUMENT USING NAÏVE BAYES

World Wide Web is the store house of abundant information available in various electronic forms. Since past few years, the increase in the performance of computers in handling large quantity of text data has led researchers to focus on reliable and optimal retrieval of visible and implied information that exist in the huge resources. In text mining, one of the challenging and growing importance’s is given to the task of document classification or text characterization. In this process, reliable text extraction, robust methodologies and efficient algorithms such as Naive Bayes and other made the task of document classification to perform consistently well. Classifying text documents using Bayesian classifiers are among the most successful known algorithms for machine learning. This paper describes implementations of Naïve Bayesian (NB) approach for the automatic classification of Documents restricted to Technical Research documents based on their text contents and its results analysis. We also discuss a comparative analysis of Weighted Bayesian classifier approach with the Naive Bayes classifier. (M et al. 2015)

2.4. CONCLUSION

QQQ

CHAPTER THREE

METHODOLOGY

3.0. INTRODUCTION

The CRoss Industry Standard Process for Data Mining (CRISP-DM) is a process model with six phases that naturally describes the data science life cycle. It's like a set of guardrails to help you plan, organize, and implement your data science (or machine learning) project. (Saltz and Hotz 2020)

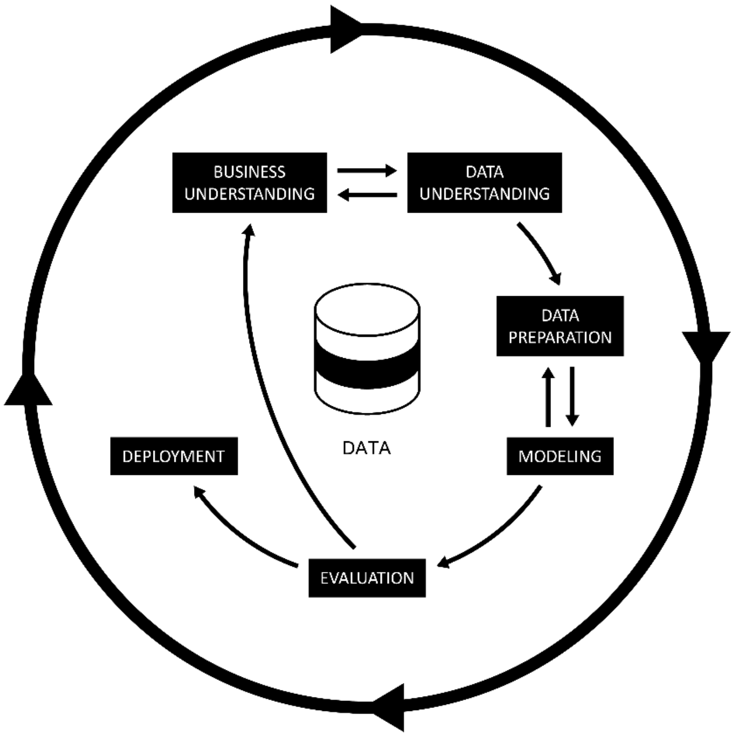

Figure 3.1: CRoss Industry Standard Process for Data Mining (CRISP-DM)

3.1. BUSINESS UNDERSTADING

3.2. DATA UNDERSTANDING

3.3. DATA PREPARATION

3.4. MODELING

3.5. EVALUATION

3.6. DEPLOYMENT

3.7. PLANING

3.8. ANALYSIS

3.9. DESIGNING

3.10. CONCLUSION

QQQ

CHAPTER FOUR

FINDING AND IMPLEMENTATION

4.0. INTRODUCTION

4.1. FINDING

4.2. TESTING

4.3. CONCLUSION

QQQ

CHAPTER FIVE

CONCLUSION

5.0. INTRODUCTION

5.1. RESULT

5.2. RECOMMENDATION

5.3. CONCLUTION

QQQ