AUTOMATED NEWS CATEGORIZING

MR. MARWAN YEEDENG

MR. NIHENG MAE

ADVISER

MR. KHOLED LANGSAREE

FATONI UNIVERSITY

1.1.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 SVM (Support Vector Machine) 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?

1.2.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.

1.3.OBJECTIVE OF RESEARCH

- 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.

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. 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.

1.5.2. 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.

1.6.SOFTWARE AND HARDWARE ARE REQUIREMENT

1.6.1.1. SOFTWARE REQUIREMENT

1.6.1.1.1. Resource

- News from Bangkok post agency website.

1.6.1.1.2. Data Mining Process

- Python
- Jupyter Notebook
- Data Mining Libraries
  - Requests
  - Beautiful Soup 4
  - Pandas
  - NumPy
  - re (Regular expression)

1.6.1.1.3. Data Visualization Process

- Website Language
  - HTML
  - JavaScript
  - CSS
  - Python
- Data Visualization Libraries
  - Django
  - Matplotlib
  - Seaborn

1.6.1.1.4. Data Base

- MySQL
- Data Integration Format
  - JSON

1.6.1.2. HARDWARE REQUIREME

1.6.1.2.1. Personal Computer

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

1.7.RELATE WORK

1.7.1. 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)

1.7.2. AUTOMATIC TEXT TAGGING OF ARABIC NEWS ARTICLES USING ENSEMBLE DEEP LEARNING MODELS.

Automatic document categorization gains more importance in view of the plethora of textual documents added constantly on the web. Text categorization or classification is the process of automatically tagging a textual document with most relevant label. Text categorization for Arabic language become more challenging in the absence of large and free datasets. We propose new, rich and unbiased dataset for the single-label (SANAD) text classification, which is made freely available to the research community on Arabic computational linguistics. In contrast to the majority of the available categorization systems of Arabic text, we offer several deep learning classifiers. With deep learning, we eliminate the heavy pre-processing phase usually used to on the data. The experimental results showed solid performance on SANAD corpus with a minimum accuracy of 93.43%, achieved by CGRU, and top performance of 95.81%, achieved by HANGRU. In pursuit of superior performance, we implemented an ensemble model to combine best deep learning models together in a majority-voting paradigm. (Elnagar et al., 2019)

1.8.METHODOLOGY

CRIST-DM

Business Understanding	- Categorizing news automatically by using machine learning SVM (Support Vector Machine).

Data Underdtanding	- Resource
	  - Bangkok post agency website.
	- File extension
	  - JSON.

Data Preparation	- NLP (Natural Language Processing).
Modelling	- SVM (Support Vector Machine).
Evaluation	- Model Evaluation.
Deployment	- Web Application.