iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4209

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Document Classification Using NLP Techniques

Angelin Florence[1], Chinmay Vaidya[2], Devendra Panchal[3], Lokesh Negi[4]
Assistant Professor[1], Student[2], Computer Engineering Department, ST John College of Engineering and Management, Palghar, Maharashtra, India

Abstract: Natural language processing, refers to an artificial intelligence method concerned with the automatic manipulation of natural language between humans and computers, specifically to program the computers for processing natural language data. This project will classify unknown set of documents into their respective categories with help of TF-IDF algorithm and Vector space model (VSM).
Keywords: NLP, TF-IDF, VSM, Document classification

I. INTRODUCTION

Document classification is an issue in library science, computer science and information science. It refers to assigning a document to a particular label or category. It can be done "manually or algorithmically. The document which needs to be classified may be images, texts, music, etc. Every kind of document has its own classification problems.
This project proposes an approach for automatically classifying documents into a set of categories using Term Frequency and Inverse Document Frequency (TF-IDF) and Vector Space Model (VSM). In this approach, input is a set of example documents. Pre-processing of the input documents is done by parsing and removing the stop words, doing stemming and extracting noun as keywords. The unknown documents that requires classification are classified by applying vector space model technique based on the derived feature sets.

II. LITERATURE SURVEY

The author [1] implements TF-IDF algorithm which counts the word weight by considering frequency of the word (TF) and in how many files the word can be found (IDF). Since the IDF could see in how many files a term can be found, it can control the weight of each word. When a word can be found in so many files, it will be considered as an unimportant word. TF-IDF has been proven to create a classifier that could classify news articles in Bahasa Indonesia in a high accuracy; 98.3 %. The goal of this research is clear, to create an online news article classifier that can classify online news article into fifteen predefined categories; beauty, business, football, economy, entertainment, health, Yogyakarta, Indonesia food and drink, lifestyle, automotive, education, politics, property, sport, technology and travel. The main process is divided into two parts; the pre-processing phase and the processing phase. In the pre-processing phase, the word and weight dictionary will be created and, in the processing, phase the uncategorized articles will be categorized based on their topics. To create the word and weight dictionary, there are seven steps that have to be conducted, and those are; tokenization, bigram creation, duplicate removal, stop-words removal, word filtering based on the term frequency, supervised word removal to create a word dictionary and (TF-IDF) implementation to get the weight of each word.
The author [2] deals with supervised machine learning and content-based document classification of textual documents that are confined to four educational departments: - Civil, Computer Science, Mechanical and Electrical Engineering by using TF-IDF algorithm along with Natural Language Processing for feature selection and ID3 algorithm as a classifier. The paper focuses on the development and evaluation of a system that proves to be helpful in various educational sectors of the country. The results show 80% accuracy.

III.PROPOSED SYSTEM

The proposed system will have unknown set of documents classified into their respective categories. The process starts with a set of known documents as input to the system for creating pre-defined categories. The first step is to calculate the text document training processor (TDTP) of each document. Then, Term Frequency is calculated along with Inverse Document Frequency (TF-IDF algorithm), which will generate the feature matrix of document. Then Vector space model (VSM) will classify the documents using the feature matrix

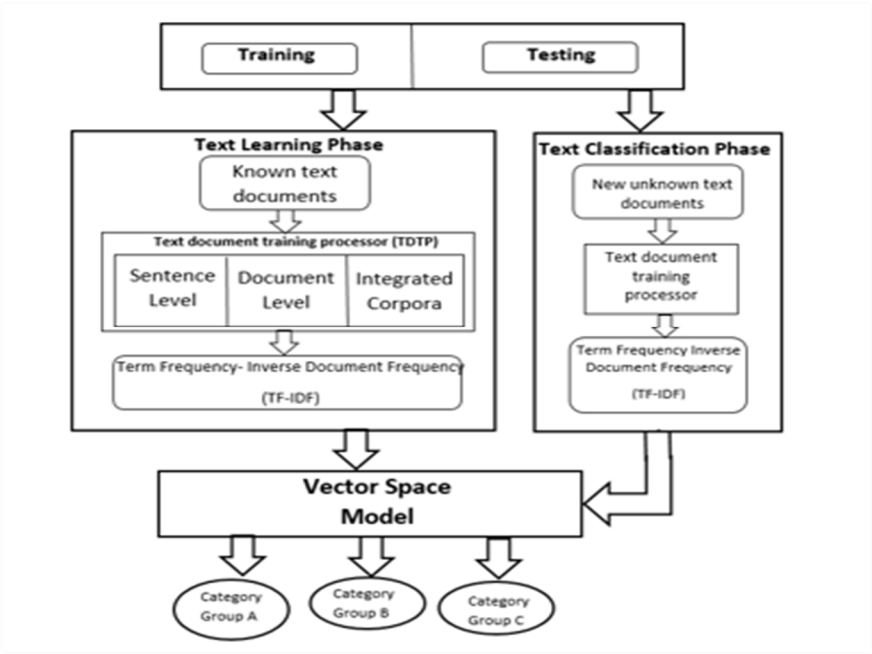


Figure 1: Architecture Diagram

A. Text Document Training Processor(TDTP): This processor prepares the document by performing many operations on the sentence, document, and integrated corpora levels and proceeds at sentence level then integrated & refined to make available for integrated corpora. To create a good classifier, it should be trained by using several known documents so that categories are created and the classifier classifies the articles precisely. The documents input to the system go through sentence extraction, POS tagging (Parts-of-speech), feature vector reduction then its result are sent to TF-IDF.

B. Term Frequency-Inverse Document Frequency(TF-IDF): In data recovery, TF-IDF, is a numerical measurement that is proposed to reflect how vital a word is to a record in an accumulation or corpus. It is frequently utilized as a weighting factor in ventures of data recovery, content mining, and client demonstrating. The tf-idf esteem expands relatively to the circumstances a word shows up in the report and is counterbalanced by the recurrence of the word in the corpus, which alters for the way that a few words seem all the more every now and again as a rule.

TF is the measure of how regularly a word shows up in a record and IDF is the measure of the uncommonness of a word inside the inquiry file. Joining TF-IDF is utilized to quantify the factual quality of the given word in reference to the inquiry. Scientifically, TFi = ni/(Σknk) where, ni is the quantity of events of the considered terms and nk is the quantity of events of all terms in the given record.

$$IDFi = \frac{\log N}{Dfi}$$

where, N = the quantity of events of the considered terms and dfi is the quantity of reports that contain term I.

$$TF\text{-}IDF = TFi \times IDFi.$$

More regular terms in an archive are more critical, i.e. more characteristic of the subject. fij = recurrence of term I in record j.May need to standardize term recurrence (tf) by separating by the recurrence of the most widely recognized term in the archive: tfij = fij/maxi{fij}. Terms that show up in a wide range of records are less demonstrative of general point. df I = archive recurrence of term I = number of reports containing term I idfi = backwards record recurrence of term I, = log2 (N/df I) (N: add up to number of reports). A run of the mill joined term significance pointer is tf-idf weighting: wij = tfij idfi = tfij log2 (N/dfi) • A term happening every now and again in the record however once in a while in whatever is left of the gathering is given high weight.

C. Vector Space Model(VSM): Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings. Documents and queries are both vectors. Each term, i, in a document or

query, j, is given a real-valued weight, wij. Both documents and queries are expressed as t- dimensional vectors: dj = (w1j, w2j, …, wtj).

IV. CONCLUSIONS

This method achieves good results by contrasting with traditional classification methods, and at the same time, this method has stronger generalization promotion ability, is recommended as a practical text classification method. The task is of classify the document into three predefined classes: one class which represents the information about sports, second class represents information about technology and third class represents education category. Documents which are not related to any of these categories will be classifies in another category. We use the accuracy metric to study the performance of our different classifiers.

REFERENCES

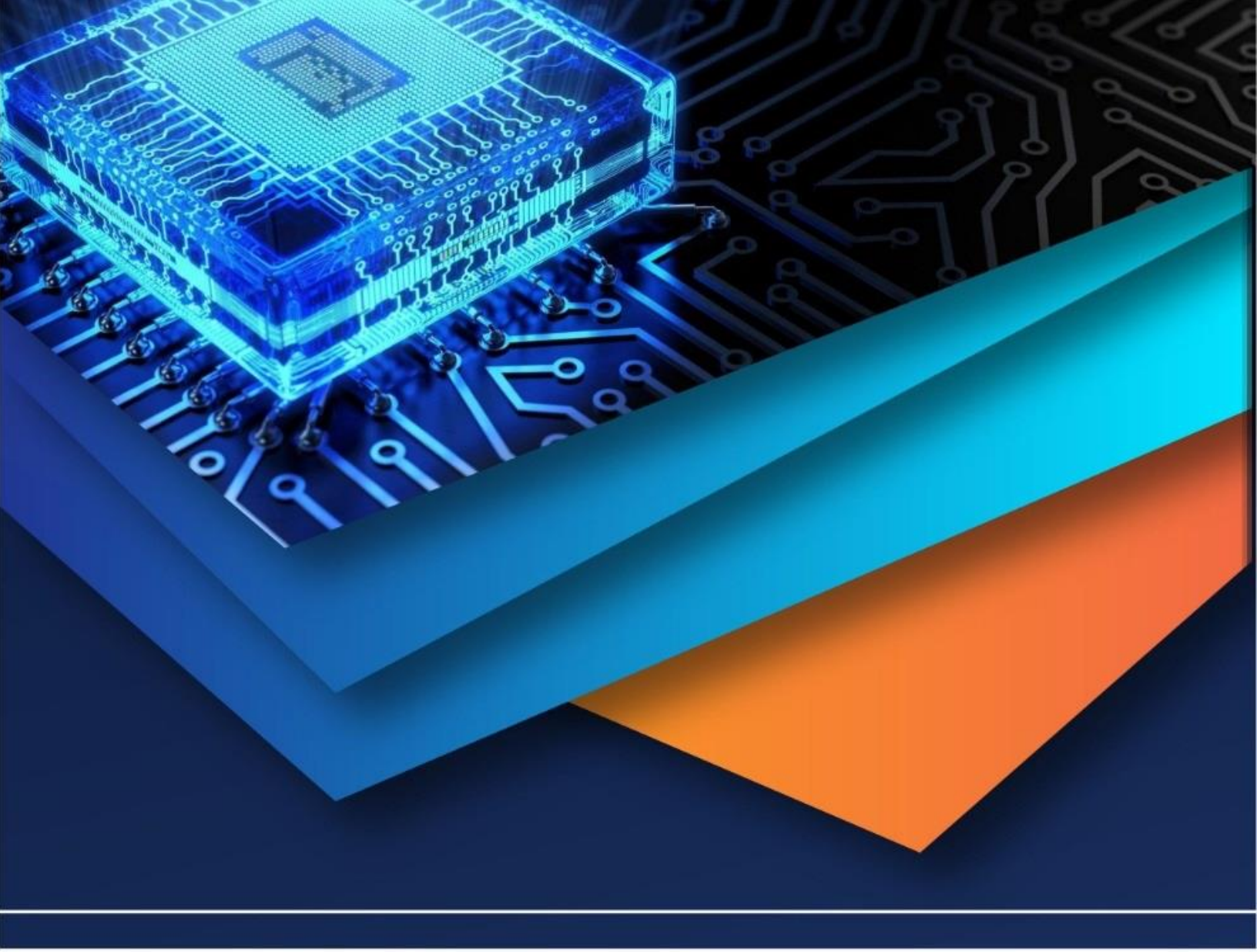
[1] Ari Aulia Hakim, Alva Erwin, Kho I Eng, Maulahikmah Galinium, Wahyu Muliady, "Automated Document Classification for News Article in Bahasa Indonesia based on Term Frequency Inverse Document Frequency (TF-IDF) Approach" 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia.

[2] Spoorthi M, Srilekha K, Sanjana J, Kushal Kumar B N "Educational Document Classification Using Natural Language Processing", International Journal of Engineering Research, Volume 5, Issue: Special 4, May-2016.

[3] Parsa Gaffari, Text Analysis 101: Document classification, Jan 2015. [Online]. Available:https://www.kdnuggets.com/2015/01/text-analysis-101-document-classification.html. [Accessed: 28- Oct- 2017].

[4] Jasiliu A.Kadiri And Niran A. Adetoro, "Information Explosion And The Challenges Of Information Andcommunication Technology Utilization In Nigerian Libraries Andinformation Centres," Ozean Journal Of Social Sciences 5, 2012.

[5] [2] Chee-Hong Chan, Aixin Sun, And Ee-Peng Lim, "Automated Online News Classification with Personalization," 4th International Confrence Of Asian Digital Library (ICADL), Pp. 320-329, December 2001.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY