KnE Social Sciences

3rd ICTVET 2018
3rd UNJ International Conference on
Technical and Vocational Education and Training 2018
Volume 2019

Knowledge E
Engaging minds

Conference Paper

A Systematic Literature Review of Short Text Classification on Twitter

Bambang Prasetya Adhi, Dea Saskiah, and Widodo Widodo

Informatics Education Universitas Negeri Jakarta

Abstract

Twitter is a microblogging service that allows people to communicate via messages containing only 140 characters briefly. With these limits, Twitter can be categorized as a short text document. And with the limited number of words makes the tweet it difficult to classify. This study aims to generate classification maps and find out the best method to classify short text documents, especially on Twitter by analyzing literary data using a systematic literature review analysis method. The process of collecting literature data is done by searching on several digital libraries with search strings that have been made based on the existing research question with the publication limit between 2013-2017. The results of this research indicate that from 1253 literature, 41 works of literature deserve to be analyzed. And based on 41 existing literature found that there are 21 methods of classification used for twitter classification. With the most widely used method is Support Vector Machine (SVM) and the best method is Word2Vec Logistic Regression with an accuracy of 95,8%.

Keywords: short text, systematic literature review, mapping research, classification, twitter

Corresponding Author:
Bambang Prasetya Adhi
bambangpadhi@unj.ac.id

Received: 11 January 2019
Accepted: 14 February 2019
Published: 25 March 2019

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the 3rd ICTVET 2018 Conference Committee.

1. Introduction

The text is an object to find information. There are two kinds of text, namely long text such as a paragraph and short text such as a news title. The long text itself or in this example a paragraph contains several sentences consisting of topic sentences, support sentences, and conclusion. But it is different from the short text that has limitations in writing, which is only about 100 characters. And it is often found that short text becomes an essential object in the topic of discussion, for example, the example that the researcher has mentioned before which is the news title.

As has been explained that a paragraph has several sentences that can be more easily classified into specific topics due to the number of dominant words in a paragraph. Unlike the short text that only has a few limited characters, so there is no dominant word

in a short text. This causes the short text difficult to be classified statistically compared to long text.

One example of a short text in addition to the news title is twitter. Twitter is one of the social media to send and receive short messages. This conversation is usually called a tweet. The tweet itself is a short message that only has a length of 140 characters. Tweets from people who have their own Twitter account are also various topics and sentiments. Some are about education, economics, technology, and others. Some are positive, neutral or negative tweets. So that makes people difficult in choosing information from the tweet. Besides, tweets on Twitter tend to have unstructured words.

Research on the short text document itself has not been classified based on its category and has not been comprehensively done because it is difficult to classify short text documents that do not have dominant words. [1] Therefore, a mapping is needed in the calcification of short text documents on Twitter so that it is easier for researchers to find the best method using data in the form of literature. This literature review aims to generate classification maps and find out the best method beetween 2013 and 2017.

2. Methodology

2.1. Research method

The method used in this study is a method of systematic literature review. [2] Where researchers analyze research that has been done previously through research literature relating to the short text classification on twitter.

2.2. Data and data source

The data used is in the form of research documents in the form of literature obtained by downloading the trusted digital library. The literature used is the result of previous research that is relevant to what the author needs in this study.

2.3. Technique and procedures of data collection

Data collection for this study uses data from the IEEE eXplore, ScienceDirect, and EBSCOhost [3] digital libraries in the search for related literature. The reason for the selection of the three digital libraries is because it is a digital library available at the

Figure 1: Research Flow Chart.

Jakarta State University and also the absence of specific criteria or characteristics in digital library selection.

2.4. Data analysis procedure

In this study, a binding rule was made. The rules are the literature that discusses the method of classifying short text documents on twitter, literature with the 2013-2017 publication year limit, and the English-language literature and when searching in digital libraries using search strings: (short text OR text) AND (classification) AND (twitter).

2.5. Data validity check

Examination of the validity of the data was carried out based on the relevance of the title of the available literature with the topic of discussion, namely the short text classification on twitter, the contents of abstracts and discussions with those who have competence and interest in research on short text documents on Twitter using the method of systematic literature review. In this case, the intended supervisor is a lecturer.[4]

3. Results and Analysis

3.1. Planning

As already stated of the research question, first the researcher must make a PICOC factor which is a primary factor in the formation of a research question. The PICOC factor can be seen in table 1.[5]

Table 1: PICOC Factors.

Population	Short text document
Intervention	Classification method
Comparison	Classification method and level of success
Outcomes	Classification method and level of success
Context	*Twitter*

Of the five PICOC factors above, a research question can be formulated which can be seen in table 2.[6–8]

Table 2: Research Question.

ID	Research Question
RQ1	What method that used in classifying short text documents on Twitter?
RQ2	What method that most often used in classifying short text documents on twitter?
RQ3	What is the best classification method in classifying short text documents on twitter?
RQ4	What journals/proceedings that most often contain research on classifying short text documents on twitter?
RQ5	What research topic chosen most often in classifying short text documents on twitter?
RQ6	Who are the authors who play an active role in classifying short text documents on twitter?

3.2. Conducting

Search or identification on a digital library is generated from a search string that was previously predetermined. The use of search strings is consistent in every digital library. From the search results, there are 1523 English literature from the digital library that have previously been determined, IEEE eXplore, EBSCOhost, and Science Direct.

Of the 1523 existing literature based on the results of identification, the literature was chosen again based on titles relevant to the topic and results into 123 literature. But in this stage, the literature is not necessarily used for research. Because the writer must first examine the quality of the literature.

Of the 123 existing literature, the literature must be reviewed or analyzed based on abstracts and the contents of the literature. In this case, the researcher reads all of the 123 available works of literature. From the screening process based on abstract and content, the final results were 41 literature. Of the 41 existing literature, 21 literature are international journals, 17 literature are the results of conferences, and three literature are manuscripts.

After reviewing or analyzing the data, the results of the data analysis are entered into the data extraction form which is made based on the research question that has been formulated previously.

3.2.1. Method classification on twitter

From Figure 2 it can be seen that there are 21 types of classification methods used in classifying short texts on twitter.

3.2.2. Frequently used classification methods

From Figure 3, it can be seen that several algorithms are often used in short or twitter text classification analysis, namely Support Vector Machine (SVM), Naive Bayes (NB), Multinominal Naive Bayes (MNB), k-Nearest Neighbor (k-NN), and Decision Tree (DT). But the most widely used algorithm of the 41 available literature is Support Vector Machine (SVM) which is 25 literature.

3.2.3. The best classification method

Based on the level of success, the research entitled Classifying Short Unstructured Data Using the Apache Spark Platform gets the highest accuracy value compared to other studies. This study uses an Associative Classifier with Entity Resolution (AssocER) and Logistic Regression classifier with Word2Vec (Word2vecLR). AssocER gets an accuracy rate of 91.8% while Word2vecLR receives the highest accuracy rate of 95.8%.

3.2.4. The most influential publication

There are three publications that are quite active in discussing short text classifications on twitter. The publication of the Journal of Theoretical and Applied Information Technology discusses as many as five literatures, the Journal of Information Science

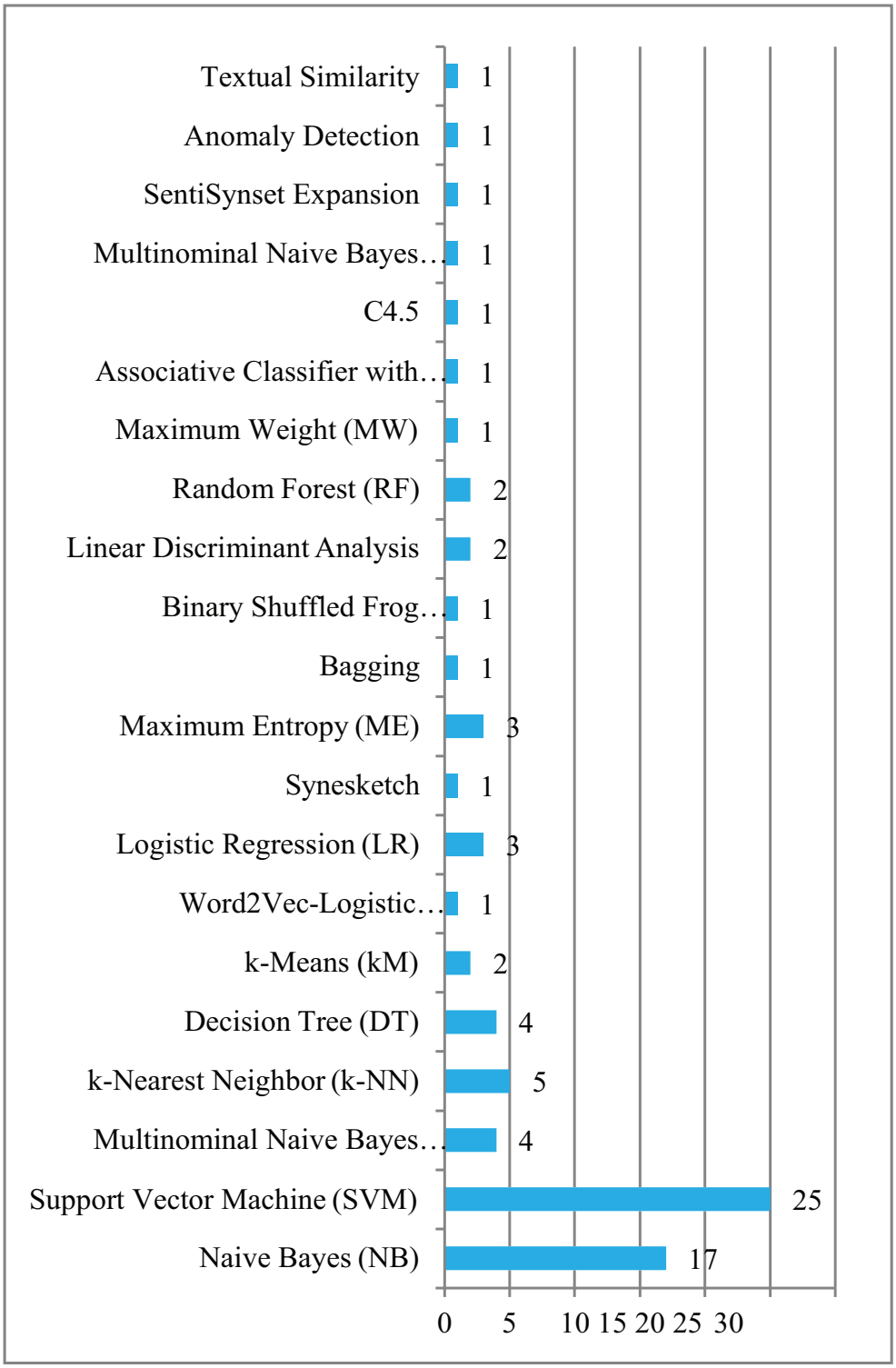

Figure 2: Various Methods in Twitter Classification.

discusses three literatures, and Applied Soft Computing discusses as many as two literatures.

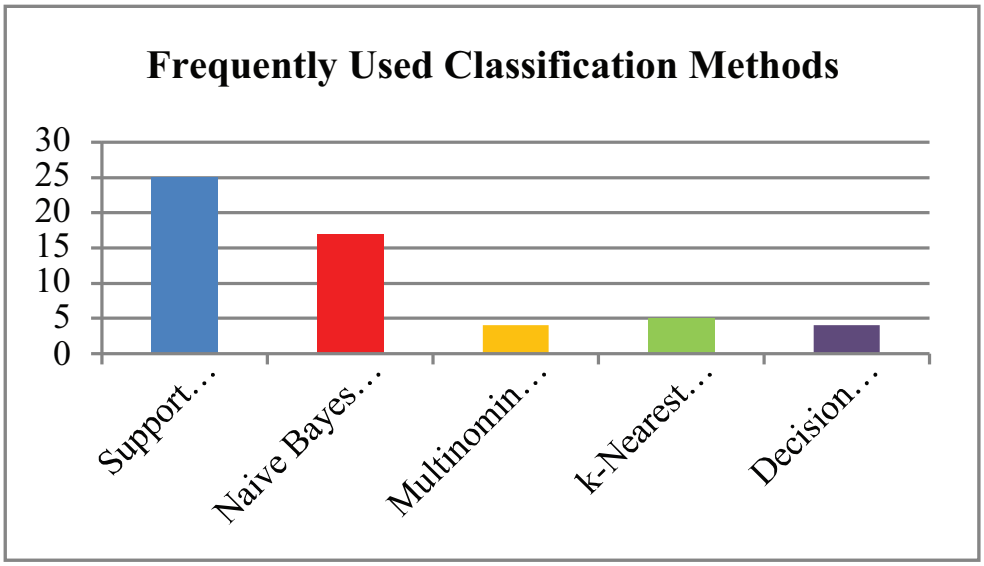

Figure 3: Frequently Used Classification Methods.

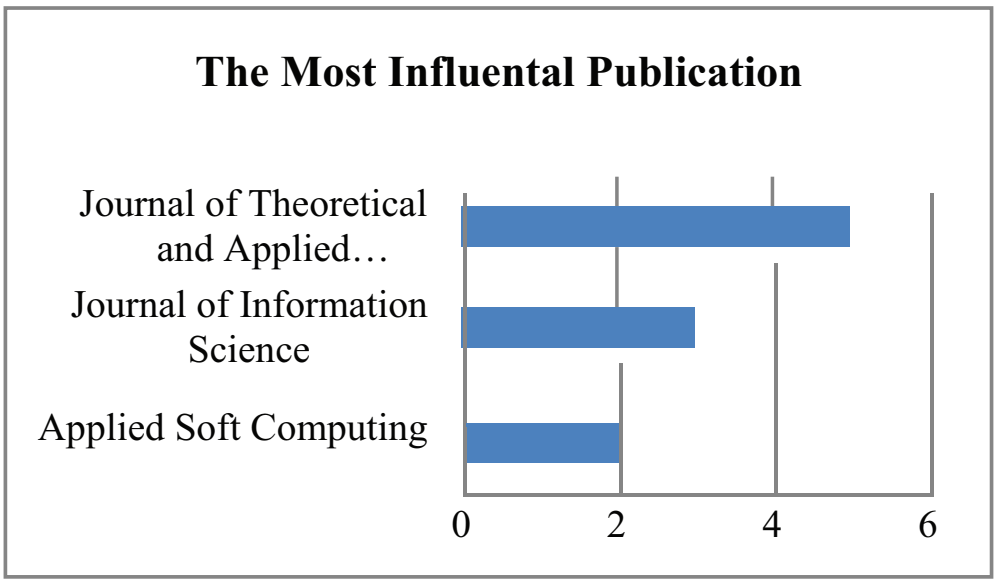

Figure 4: The Most Influential Publication.

3.2.5. Topics in classification literature on twitter

In the 41 literature that has been obtained, several topics are often discussed as safety pins for conducting research. Of the three topics, the topic most often taken as research material is Classification Sentiment.

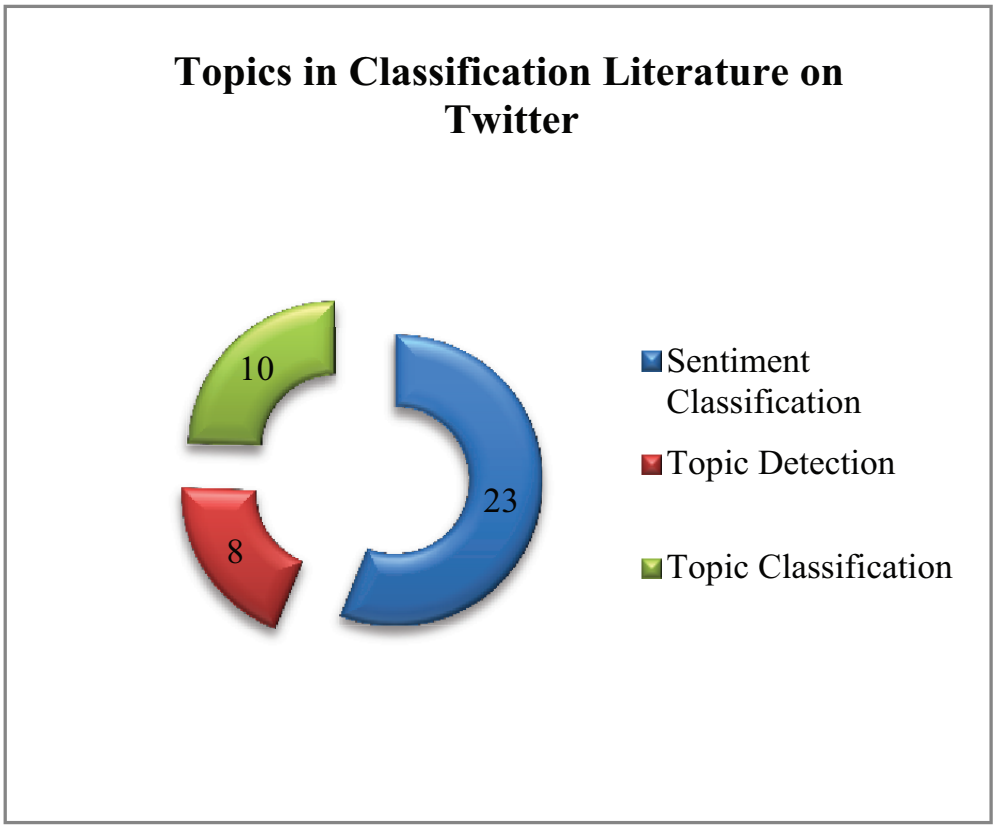

Figure 5: Topics in Classification Literature on Twitter.

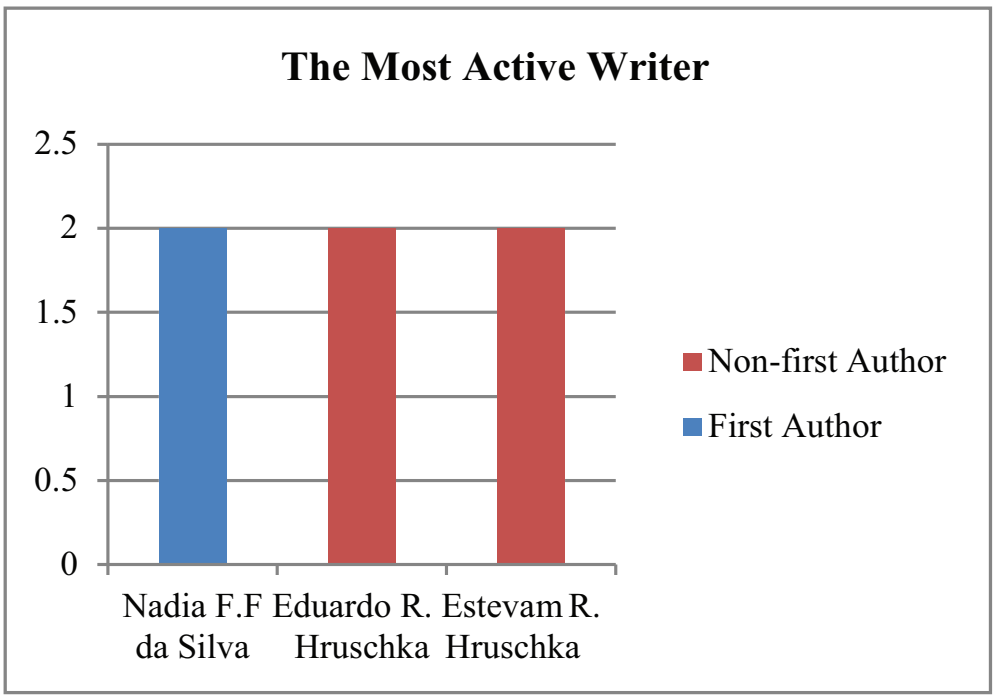

Figure 6: The Most Active Writer.

3.2.6. The most active writer

From the 41 literature taken, 131 researchers contributed to the research related to the short text classification on Twitter, with details of which 128 of them contributed to 1 study and three others contributed to 2 studies. Figure 7 shows the most active researcher and influences the research related to the short text classification on Twitter based on the number of studies conducted by the researcher.

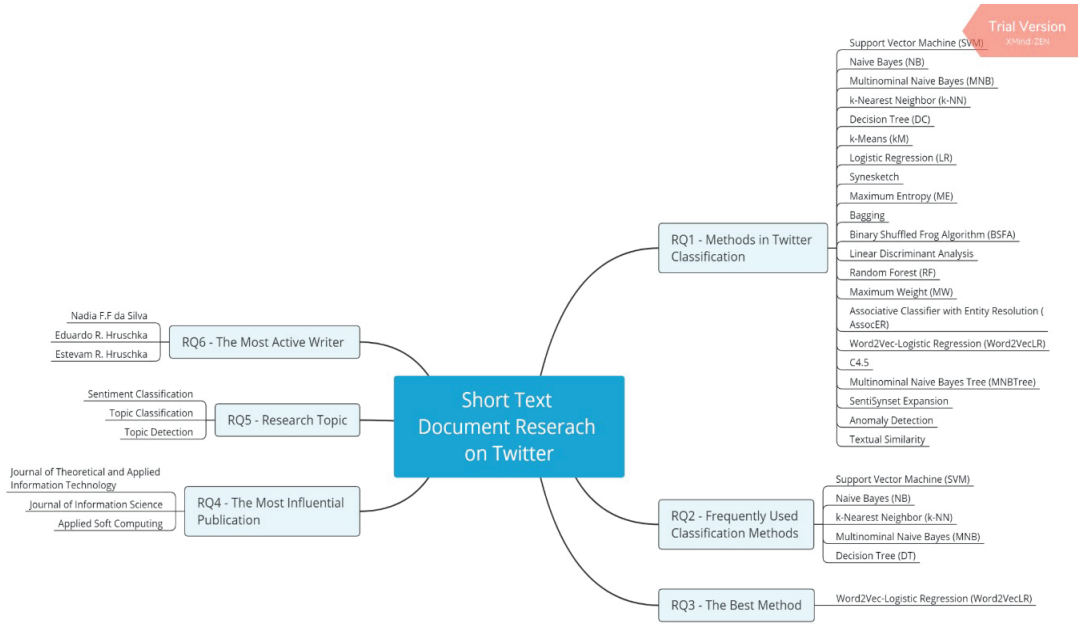

Figure 7: Short Text Document Research Map on Twitter.

3.3. Discussion

From the analysis of 41 existing literature, 21 types of classification methods were used for short text classification analysis. 21 methods are: Naive Bayes (NB), Support Vector Machine (SVM), Multinominal Naive Bayes (MNB), k-Nearest Neighbor (k-NN), Decision Tree (DT), k-Means (kM), Word2Vec- Logistic Regression (Word2VecLR), Logistic Regression (LR), Synesketch, Maximum Entropy (ME), Bagging, Binary Shuffled Frog Algorithm (BSFA), Linear Discriminant Analysis (LDA), Random Forest (RF), Maximum Weight (MW), Associative Classifiers with Entity Resolution (AssocER), C4.5, Multinominal Naive Bayes Tree (MNBTree), SentiSynset Expansion, Anomaly Detection, and Textual Similarity.

Of the 21 types of classification methods, it was found that Support Vector Machine (SVM) is most often used for short text classification. Of the 41 literature obtained 25 literature using the SVM classification method. SVM is often used because it is

considered more effective and does not take much time in classification. SVM also often provides the best level of accuracy compared to other algorithms.

Of the 41 literature used, the literature entitled Classifying Short Unstructured Data Using the Apache Spark Platform gets the highest accuracy value compared to other studies, using the Spark platform to implement two classification strategies for large-scale processing data. This study uses an Associative classifier with Entity Resolution (AssocER) and Logistic Regression classifier with Word2Vec (Word2vecLR). AssocER is effective for product datasets with an accuracy rate of 91.8%, but the results are not suitable for tweet dataset. While Word2vecLR is effective for tweet datasets with an accuracy of 95.8%, but the level of accuracy in the product dataset is not good. AssocER is effective and efficient for classifying large numbers of words while Word2VecLR is better used for informal data contexts such as tweets on Twitter in large numbers.

4. Conclusions and Suggestions

The conclusion obtained to map the research of short text documents on Twitter after going through all the step in systematic literature review method can be seen in picture 4.6 which is a short text research map on Twitter. And the following is an explanation of the research map in picture 4.6:

1. Of the 41 existing literature, there are 21 classification methods used in classifying short texts on twitter.
2. The Support Vector Machine (SVM) method is the most widely used method in classifying short texts on twitter. Followed by Naive Bayes (NB), k-Nearest Neighbor (k-NN), Multinominal Naive Bayes (MNB), and Decision Tree (DT). Support Vector Machine (SVM) is often used because it is considered more effective and faster during runtime.
3. The best method for classifying short text on twitter is the Word2Vec Logistic Regression method (Word2VecLR) with an accuracy rate of 95.8%.
4. The topic of discussion that is often used in classifying short texts on twitter is sentiment classification.

References

[1] Han, Jiawei dkk. 2012. *Data Mining Concepts and Techniques*. Ed ke-3. USA: Elsevier Inc

[2] . Wahono, R. S. (2015). *A* Systematic Literature Review of Software Defect Prediction: Research Trends, Datasets, Methods and Frameworks. *Journal of Software Engineering*, 1:1

[3] Kitchenham&Charters. (2007). Guidelines in Performing Systematic Literature Reviews in Software Engineering. *EBSE Technical Report 33(5): 20.*

[4] Hall, T., Beecham S., Bowes D., Gray D., & Counsell, S. (2012). A Systematic Literature Review on Fault Prediction Performance in Software Engineering. IEEE Transactions on Software Engineering, 38:1276-1304.

[5] Leedy, Paul.D. & Jeanne.E. Ormrod. (2014). *Practical Research: Planning and Design a Research*. Ed ke-10. London: Pearson Education Limited.

[6] Heck, P. & Zaidman, A. (2016). A Systematic Literature Review on Quality Criteria for Agile Requirements Specifications. Link Springer, 26:127-160.

[7] Paris, D. L., Bahari, M., Iahad, N., & Ismail, W. (2016). Systematic Literature Review of e-Commerce Implementations Studies. *Search EBSCOhost*, 89:422-438.

[8] Wen, J., Li, S., Lin, Z., Hu, Y., & Huang, C. (2012). Systematic Literature Review of Machine Learning (ML) Based Software Development Effort Estimation Models. *Search EBSCOhost*, 54:41-59.