ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

Analysis of Different Classification Algorithms on Question-Pairs and 20 news-Bydate Datasets

Govardhana Babu Kolli[1], Dr. B Prajna[2]

M. Tech Student, Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India[1]

Professor, Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India[2]

ABSTRACT: This paper presents a system which finds text similarity and text classification with two datasets question-pairs and 20news-bydate respectively. In the former case, we determine how the multiple question pairs are closer to each other based on various classification techniques with bag of words to classify question-pairs as duplicate or different. In the latter case, the task of assigning a set of predefined categories to news according to its content so that users can identify the popular news groups uses different classification techniques with Term Frequency-Inverse Document Frequency (TF-IDF). Our proposed approach was evaluated using two question-pairs and four groups of 20news-bydate datasets implemented with three classification techniques such as K-Nearest Neighbour (KNN), Naive Bayes and Support Vector Machine (SVM) to predict whether the text is similar or not and to classify news into various groups. The classification accuracy was obtained as 65.15% and 92.08% for question-pairs and 20Newsgroup datasets respectively. Better results are obtained while compared with other classification techniques.

KEYWORDS: Text similarity, text classification, KNN, Naïve Bayes, SVM, Bag-of-Words, TF-IDF.

I. INTRODUCTION

Text similarity plays a vital role in retrieving useful information out of large amounts of data and is also used for question answer portals Information Retrieval. Quora and Stack Overflow are two social media platforms where people ask questions and can connect to the actual experts who contribute unique insights and quality answers. It has an enormous user base and over 100 million people visit every month, so that many people may ask similar worded questions. Multiple questions with the same intent can cause readers to spend more time finding the best answer to their question, and make writers feel they need to answer multiple versions of the same question. It would be beneficial if the redundant question can be reduced. This will help the users of the site in getting knowledge pertaining to a topic at one place rather than searching to same material over different questions which results in wastage of time and resources.

Text classification is the task of assigning a set of predefined tags or categories to news according to its content. It is one of the fundamental tasks in Natural Language Processing (NLP).Text classification methods have drawn much attention in recent years and have been widely used in many programs. Text mining is the process of finding some information which is previously unknown by extracting that information from large sets of unstructured text. Today, this un-structured data is growing as mostly the information is available in an electronic form such as emails, on World Wide Web, electronic publications and other documents. This un-structured information cannot be used for further processing by computers. The computers typically handle text as simple sequences of character string and are unable to provide useful information from the given text, without any process performed on it. Therefore, specific processing and pre-processing methods are required in order to extract useful patterns and information from the unstructured text.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

II. PROBLEM DEFINITION

The main objective is to determine whether a text is similar or dissimilar and assigning a set of predefined tags or categories to free news according to its content based on classification techniques. The input to the classifier is text based and output is finding text either similar or dissimilar and assigning categories to according to its content.

III. DATASET

We performed our experiments on two datasets viz. question-pairs and 20news-bydate datasets.

3.1. Question-pair dataset:

The data is hosted by Kaggle. It consists of about404351 training examples having 6 attributes:

- id - the id question pair
- qid1, qid2 - unique ids of each question
- question1, question2 - the full text of each question
- is_duplicate - the target variable, set to 1 ifquestion1 and question2 have the same meaning, and 0 otherwise.

Here are some examples of duplicate and non-duplicate questions from the data set.
The question pairs below are duplicates:

- Why did Trump win the Presidency?
- How did Donald Trump win the 2016 Presidential Election?

The question pairs below are non-duplicates:

- How can I start an online shopping (ecommerce) website?
- Which web technology is best suitable for building a big E-Commerce website?

3.2. 20news-bydate dataset:

The 20 Newsgroups dataset consists of 20,000 newsgroup documents, partitioned into 20 different newsgroups. The dataset contains over 18000 messages with assigned topic labels, and is split into train and test parts. The training dataset contains around 11300 labeled messages and the test dataset contains 7500 messages without label(the task is to predict these labels).Here, we have taken four categories such as 'alt.atheism', 'soc.religion.christian', 'comp.graphics', 'sci.med'.
The train part having 3 attributes:

- Id- the id of 20news-bydate
- Message- messages of different topics
- Topic - the topic which belongs to group id

The test part having 2 attributes:

- Id- the id 20news-bydate
- Message- messages of different topics

IV. METHODOLOGY

Analysis of different classification algorithms on two datasets follows 3 steps:

4.1. Split Dataset into Train and Test part:

Figure 4.1.shows the workflow of the model. The dataset is initially divided into training and testing data using a 70:30 split. A supervised approach is used to train the model. Being a feature based approach; the next step is to extract the features from the data.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

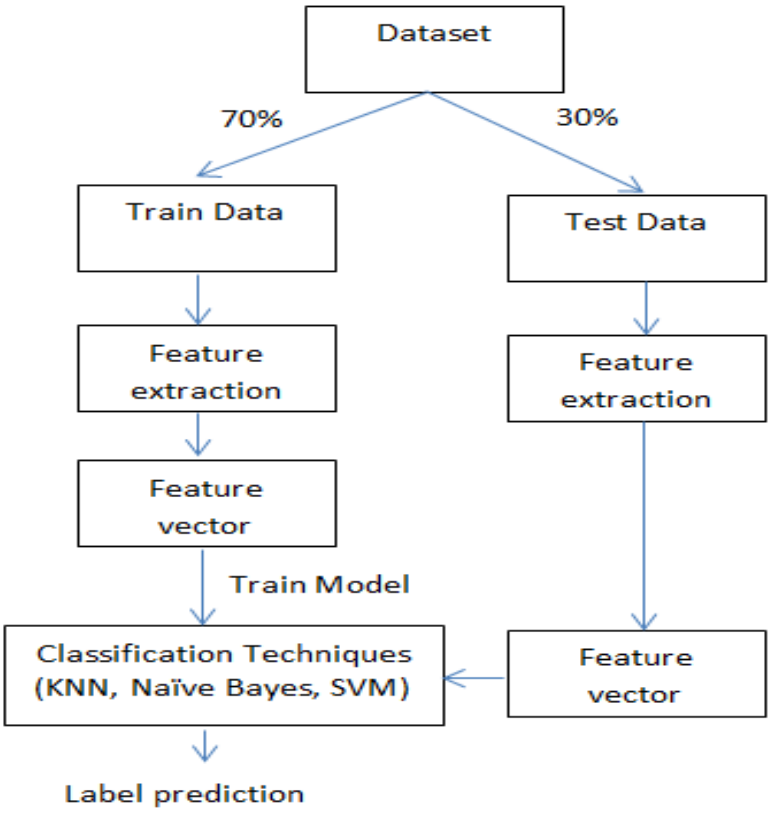

Figure 4.1. Workflow Model

4.2. Feature Extraction and Feature Vector:

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to manageable groups for processing. The large dataset require a lot of computing resources to process. In this method that select and or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set. A feature vector is a vector that stores the features for a particular observation in a specific order. We will use a bag of words and TF-IDF representation to convert reviews into sparse vector.

4.3. Classification:

To build the classifier we use three different supervised techniques: K-Nearest Neighbour, Naive Bayes, and Support-Vector Machine.

V. IMPLEMENTATION

5.1. Workflow:

- Load the dataset.
- Split data into train and test.
- Convert dataset into vectors using bag of words or TF-IDF.
- Apply classification algorithm to predict similarity or to classify the news into various groupsof train data

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

- Find accuracy of the model

The model was implemented in Python using sklearn package. The other packages also included nltk, pandas and numpy.

VI. RESULTS

The evaluation metric used is the score, which is nothing but the mean accuracy on the given test set and labels. The predictions made by the trained model on the test split are compared with the corresponding labels and the mean accuracy is computed.

Accuracy=(TP+TN)/(TP+TN+FP+FN)
Where TP = True Positive
TN = True Negative
FP = False Positive
FN= False Negative

6.1 Results

In this section, we discuss about the output screens that shows the flow of process. Here we explain two datasets, question-pairs and 20news-bydate. The proposed system in this paper got an accuracy score of 65.15%for question-pairs dataset when trained on a subset of 1000question-pairs sampled from the training data and 92.08% for 20news-bydate. It was achieved better accuracy as compared to the other classification techniques.

6.1.1. The results showing accuracy of various classifiers for Question-pairs dataset:

```
Accuracy Score is: 65.15%.

Confusion Matrix:
[[182  29]
 [ 86  33]]

Classification Report:
             precision    recall  f1-score   support

          0       0.68      0.86      0.76       211
          1       0.53      0.28      0.36       119

avg / total       0.63      0.65      0.62       330
```

Figure 6.1. BOW + KNN Model

```
Accuracy Score is: 59.7%.

Confusion Matrix:
[[132  79]
 [ 54  65]]

Classification Report:
             precision    recall  f1-score   support

          0       0.71      0.63      0.66       211
          1       0.45      0.55      0.49       119

avg / total       0.62      0.60      0.60       330
```

Figure 6.2.BOW + Naïve Bayes Model

```
Accuracy Score is: 60.61%.

Confusion Matrix:
[[142  69]
 [ 61  58]]

Classification Report:
             precision    recall  f1-score   support

          0       0.70      0.67      0.69       211
          1       0.46      0.49      0.47       119

avg / total       0.61      0.61      0.61       330
```

Figure 6.3. BOW + Support Vector Machine

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

6.1.2 The summary results showing accuracy of various classifiers for Question-pairs dataset:

SNo.	Classifiers	Accuracy	Precision	Recall
1	K-Nearest Neighbours using Bag of Words Vector Representation	65.15%	63.0%	65.0%
2	Naive Bayes using Bag of Words Vector Representation	59.7%	63.0%	60.0%
3	Support Vector Machines using Bag of Words Vector Representation	60.61%	61.0%	61.0%

Table.6.1. Results

6.2.1.The results showing accuracy of various classifiers for 20news-bydate dataset

```
Accuracy Score is 83.49%.

Confusion Matrix:
[[192   2   6 119]
 [  2 347   4  36]
 [  2  11 322  61]
 [  2   2   1 393]]

Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.60      0.74       319
           1       0.96      0.89      0.92       389
           2       0.97      0.81      0.88       396
           3       0.65      0.99      0.78       398

avg / total       0.88      0.83      0.84      1502
```

Figure 6.4.TF-IDF + Naive Bayes

```
Accuracy Score is 92.08%.

Confusion Matrix:
[[264   6  13  36]
 [  3 375   6   5]
 [  4  27 361   4]
 [  3   9   3 383]]

Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.83      0.89       319
           1       0.90      0.96      0.93       389
           2       0.94      0.91      0.93       396
           3       0.89      0.96      0.93       398

avg / total       0.92      0.92      0.92      1502
```

Figure 6.5. TF-IDF + SVM

```
Accuracy Score is 77.23%.

Confusion Matrix:
[[280   3   9  27]
 [ 50 312  12  15]
 [ 83  23 243  47]
 [ 63   2   8 325]]

Classification Report:
              precision    recall  f1-score   support

           0       0.59      0.88      0.70       319
           1       0.92      0.80      0.86       389
           2       0.89      0.61      0.73       396
           3       0.79      0.82      0.80       398

avg / total       0.81      0.77      0.78      1502
```

Figure 6.6.TF-IDF + KNN

6.2.2. The summary results showing accuracy of various classifiers for 20news-bydate dataset

SNo.	Classifiers	Accuracy	Precision	Recall
1	TF-IDF + Naive Bayes	83.49%.	88.0%	83.0%
2	TF-IDF + Support Vector Machines	92.08%	92.0%	92.0%
3	TF-IDF + K-Nearest Neighbours	77.23%.	81.0%	77.0%

Table.6.2. Results

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 8, August 2019

Table 6.1 shows accuracy of various classifiers in analysis of text similarity. The results of question-pairsdataset implemented with K-Nearest Neighbours plus Bag of Words Vector Representation given better accuracy compares to other classifiers in predicting the text similarities. Table 6.2.shows the 20news-bydate dataset implemented withTF-IDF + Support Vector Machines has given excellent accuracy compares to remaining techniques in classifying the news into different groups.

VII. CONCLUSION

This paper presented a set of classification techniques compared for text similarity measuresand to classify the news into various groups. It also demonstrated how effective pre-processing and post processing are done to enhance the performance of the classifier.The objective is to predict whether the text is similar or not and to assigning news into various groups according to its content. The experimental results indicate that the techniques used are very promising in performing all these machine learning classifiers to obtain the best results based on accuracy as metric.

REFERENCES

[1] Shashi Shankar, Aniket Shenoy "*Identifying question pairs having the same intent*"
[2] Seyyed Mohammad Hossein Dadgar, Mohammad Shirzad Araghi, Morteza Mastery Farahani, *"A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification"*(ICETECH), March 2016, India.
[3] Shashank Pathak, 2ayush Sharma, Shashank Shekhar Shukla, *"Semantic String Similarity for Quora Question Pairs"* IJAST, Issn(P): 2321 – 8991, Issn(E): 2321 –9009
[4] J. Jeon, W. B. Croft, and J. H. Lee, *"Finding semantically similar questions based on their answers,"*in Proceedings of the 28th ACM SIGIR Conference, ser. SIGIR '05. New York, NY, USA: ACM, 2005, pp. 617–618.
[5] M. Chen, X. Jin, and D. Shen, *"Short text classification improved by learning multi-granularity topics,"* in IJCAI,2011, pp. 1776–1781.
[6] V. Govindaraju and K. Ramanathan, "Similar document search and recommendation," Journal of Emerging Technologies in Web Intelligence, vol. 4, no. 1, pp. 84–93, 2012.
[7] Torsten Zesch et al. 2012. UKP: *"Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures"*