Text Classification using XLNet with Infomap Automatic Labeling Process

Presented By: Alvi Ahmmed Nabil
Roll: 1707009

Department of Computer Science and Engineering

Khulna University of Engineering & Technology, Khulna, Bangladesh

Author: Triana Dewi Salma, Gusti Ayu Putri Saptawati, Yanti Rusmawati

Published On: 2021 IEEE 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)

Agenda

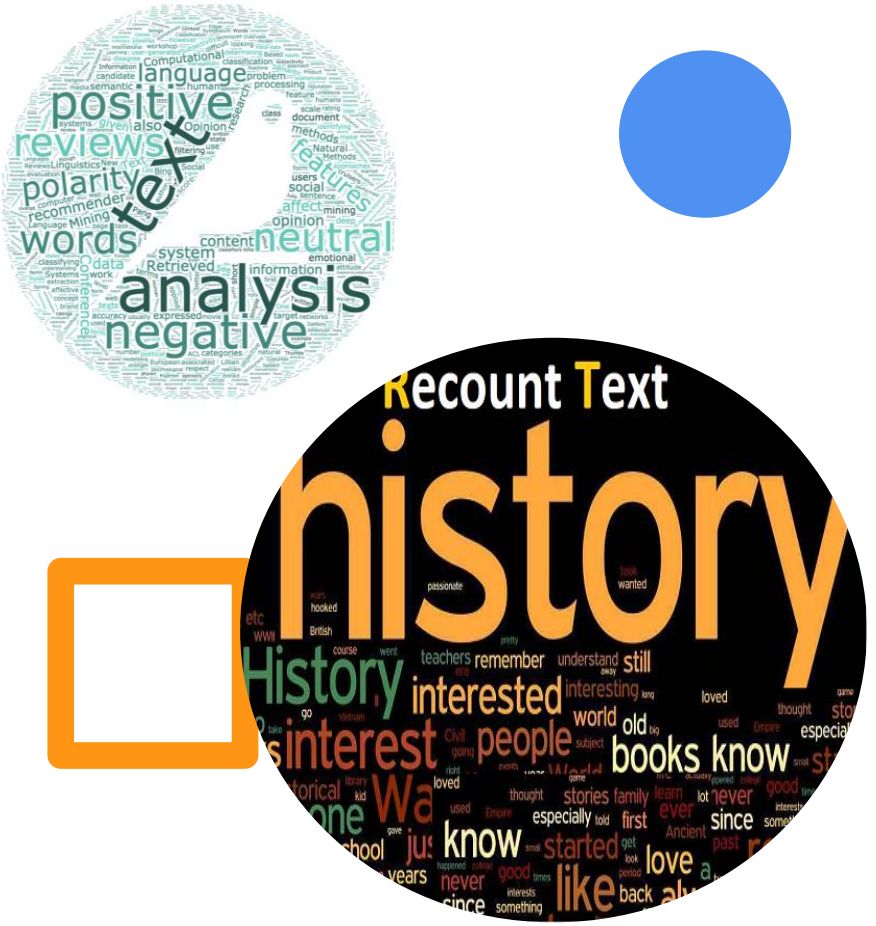
- Objectives
- Introduction
- Method
- Experiment and Analysis
- Conclusion
- Future Work

Objective

- To automatically label the the training data using community detection with infomap.
- Then, to make classification model based on the training data produced in first step using XLNet.

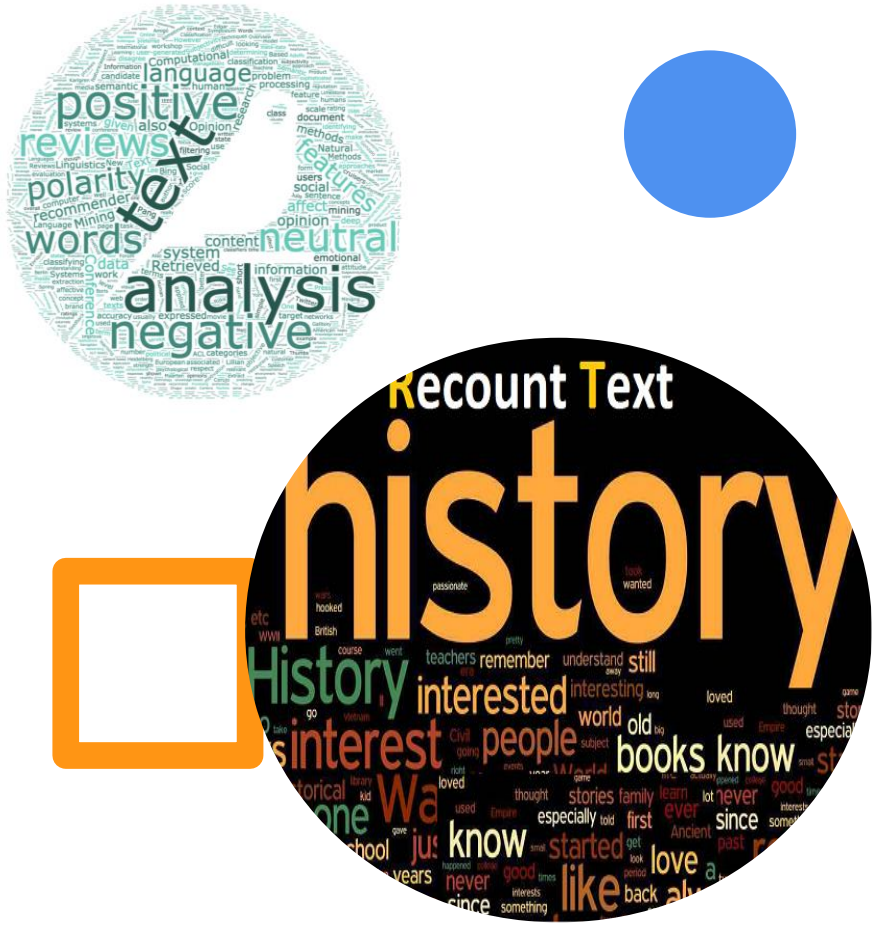
Introduction

- One of the most important task in Natural Language Processing is Text Classification.
- Training data in supervised learning must be labeled manually by humans who are experts in their fields, resulting in high costs and limited high-quality training data.
- Manual data labeling by humans is prone to mislabeling the data which plays a big role on the quality of the model trained in classification.



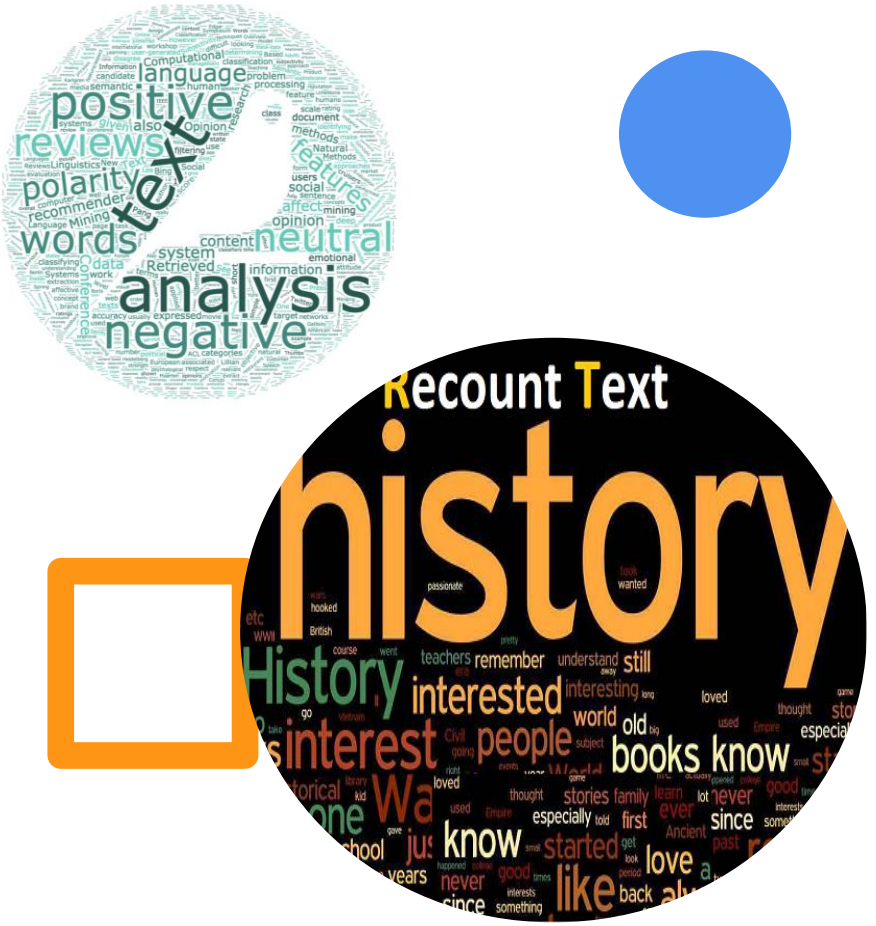
Introduction

- Community Detection with infomap can deal with this problem of labeling.
- We can automate the process of labeling where humans do not have to label the data manually.



Introduction

- In recent years, Deep Learning has become the center of attention for various fields such as image processing, NLP and computer vision.
- Deep Learning can take advantage of large datasets to achieve a higher level of accuracy than previous classification techniques. One of the newest deep learning models, XLNet has received a state-of-the art predicate for 18 NLP tasks.



Method

Method

- The whole process is divided into 2 stages.
  1. Automatic Labeling Stage of training data using Infomap.
  2. Text Classification stage using the XLNet model.

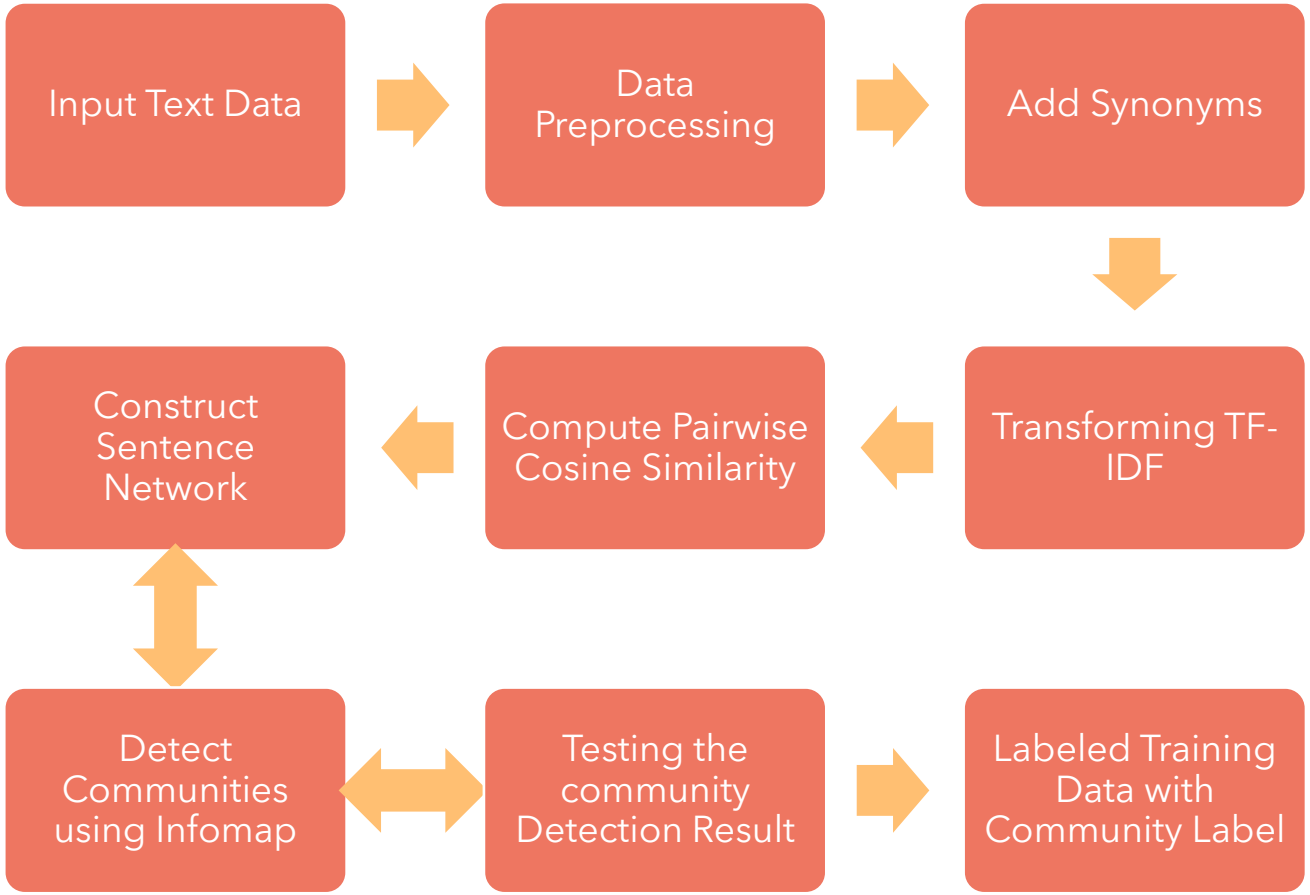
Method

- Data: The data used in this study was taken from TREC-QA, one of the open domain question-answer dataset. The data set consists of two columns named Questions and Class Label and 5452 rows for training and 500 rows for test.

Questions	Class Label
What fowl grabs the spotlight after the Chinese Year of the Monkey ?	1(Entity)
How can I find a list of celebrities ' real names ?	0(Description)
What sprawling U.S. state boasts the most airports ?	5(Location)
How many Jews were executed in concentration camps during WWII ?	4(Number)
What does the abbreviation AIDS stand for ?	2(Abbreviation)
Name 7 famous martyrs .	1(Human)

Method

❑Automatic Labeling Stage of training data using Infomap:



Method

- Automatic Labeling Stage of training data using Infomap:
  - **Preprocessing**: This Steps includes deleting URLs, removing punctuations, deleting stopwords and stemming.
  - **Addition of Synonyms**: The addition of synonyms aims to increase the variety of words and is expected to improve the understanding of the resulting classification model for different words but describe the same meaning.

Method

- Automatic Labeling Stage of training data using Infomap:
  - **TF-IDF Transformation**: Each sentence is transformed into a vector representation by transforming TF-IDF calculation. It considers the word or Term Frequency(TF) in the document as well as how unique or \infrequent(IDF) a word is in the corpus. It gives a higher score for unique words and devalues common words.
  - The TF-IDF transformation consisted of two types of Ngrams, namely bigram and trigram.
  -

$$w_{t,d} = tf_{t,d} * x * \log_{10}\left(\frac{N}{df_t}\right)$$

Where,

$tf_{t,d}$ is the frequency of the word t in the document d
N is the number of document in the collection
$df_{t,d}$ is the number of documents where the word t appears

Method

- Automatic Labeling Stage of training data using Infomap:
  - **Calculation of Cosine similarity**: When the data has been represented in a vector using TF-IDF, the equation between the two vectors in each training data is calculated using cosine similarity. Cosine similarity has a value range of 0 to 1 where 0 means no similarity between sentences and 1 mean the sentences are equal. This step is achieved using Closeness Matrix.
  - **Formation of Sentence Network**: The sentence network is formed from the nodes of each sentences in the training data and the weights obtained from the calculation of the similarity matrix. This step is achived by using the newtorkx library in python.

Method

- Automatic Labeling Stage of training data using Infomap:
  - **Infomap Community Detection**: The first step is that each community will be coded according to the community level, and then in each community the node will be coded based on the node level. By integrating these two aspects, the coding of one node is confirmed by community-coding and node coding. Therefore, the problem of community detection can be replaced by the problem of coding compression, which makes the length of the encoding the shortest. The best method for community detection would be to obtain the maximum amount of encoding compression.

Method

- Automatic Labeling Stage of training data using Infomap:
  - **Testing of Automatic Labeling Result**: One test can be carried out by using Newman Girvan Modularity which has a value range of 0 to 1. This can be calculated from following equation:

$$Q(S) = \frac{1}{m} \sum_{c \in S} (m_s - \frac{(2m_s + l_s)}{4m})$$

where,

m is total number of edges in graph

m_s is number of edges in community

l_s is number of edges from node S to nodes outside of S

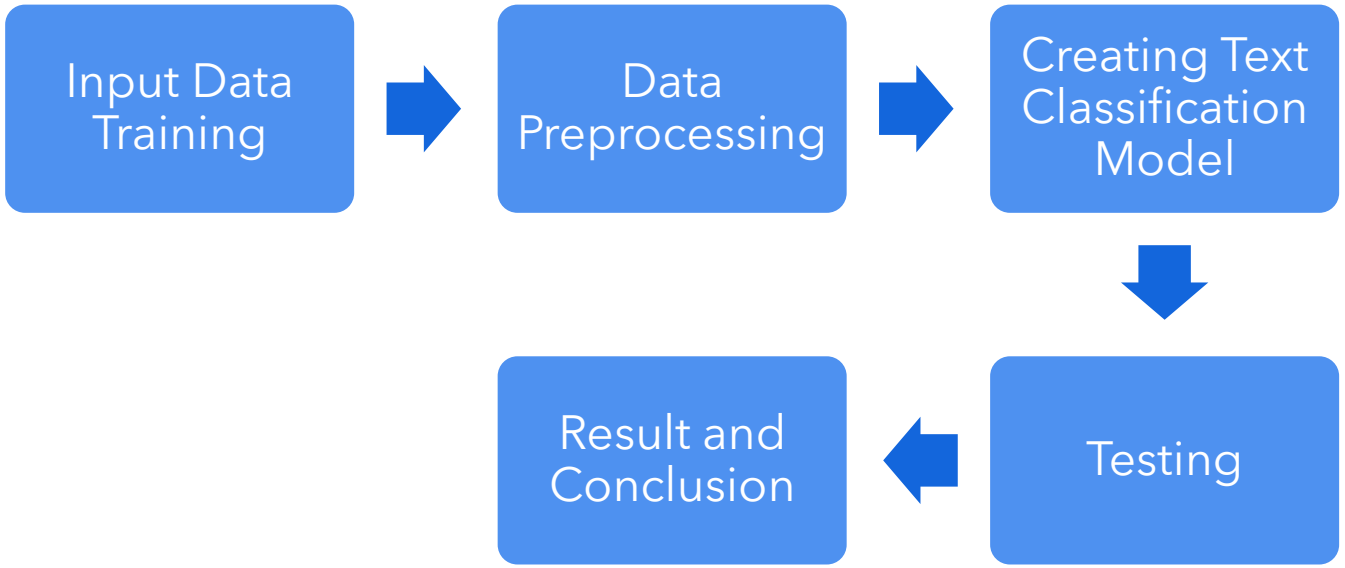
Method

❑**Text Classification**:

The XLNet model is pre-trained on BooksCorpus, English Wikipedia, Giga5, ClueWeb and Common Crawl. Classification is done by dividing the training data into 80% as training and 20% as validation data. Text classification is implemented based on a model that has passed the pre-train process with the XLNet model for sequence classification base cased. This model is then fine-tuned to better understand the training data that is owned and the number of epochs is set, namely 10, learning rate 3e-5, batch size in the range 1-5, max_len 64, and AdamW optimizer.

Method

- Text Classification:



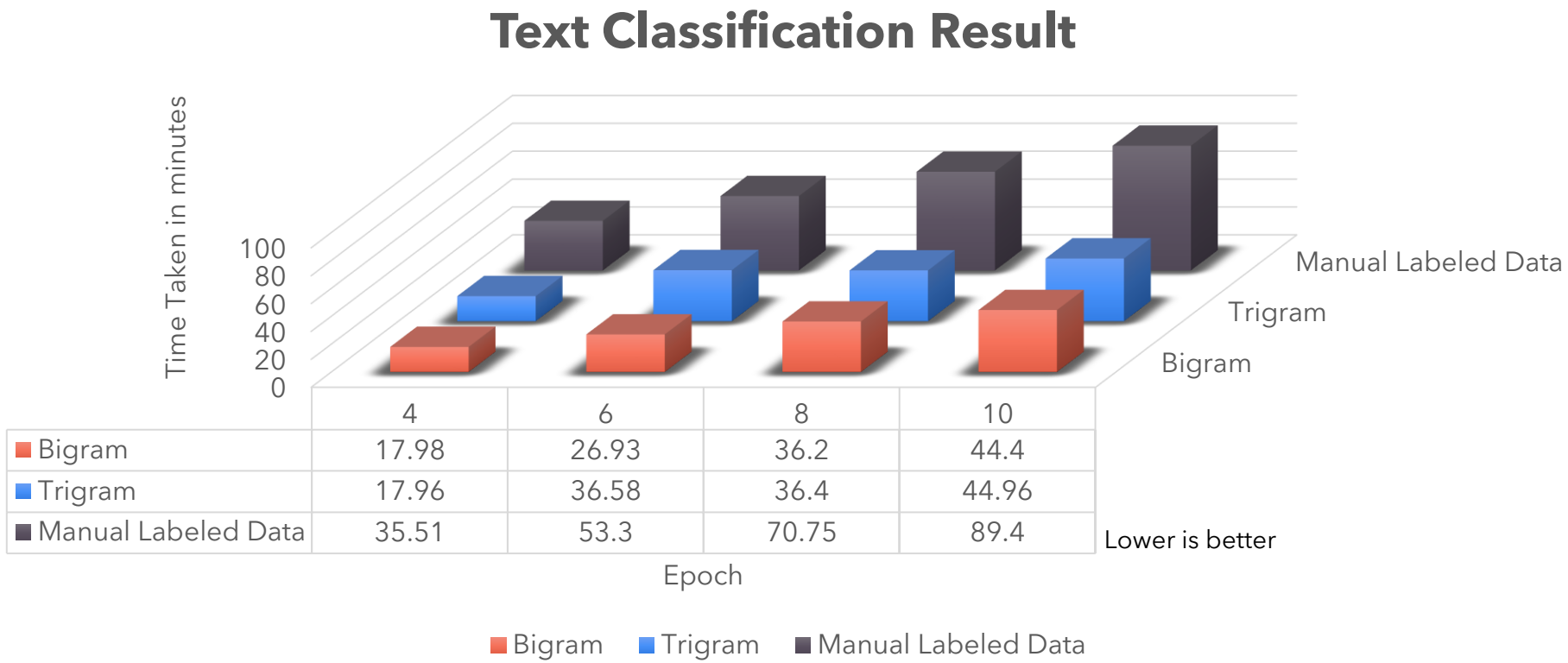
Expriment And Analysis

Experiment And Analysis

Result of Automatic Labeling

Ngram	Number of Community Formed	Modularity	Class Split	Class Merge
Bigram	1007	0.559	365.833	2.171
Trigram	706	0.5793	309.330	2.626

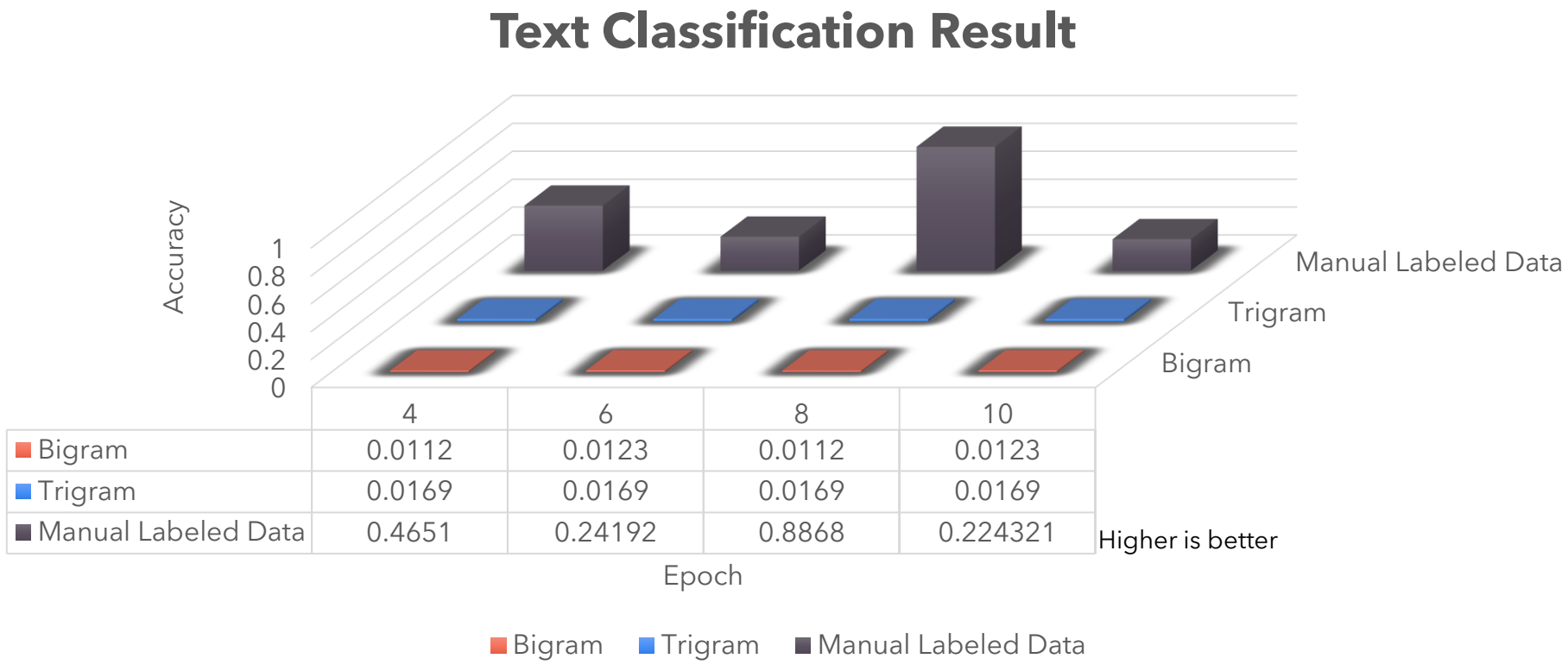
Experiment And Analysis



	4	6	8	10
Bigram	17.98	26.93	36.2	44.4
Trigram	17.96	36.58	36.4	44.96
Manual Labeled Data	35.51	53.3	70.75	89.4

Lower is better

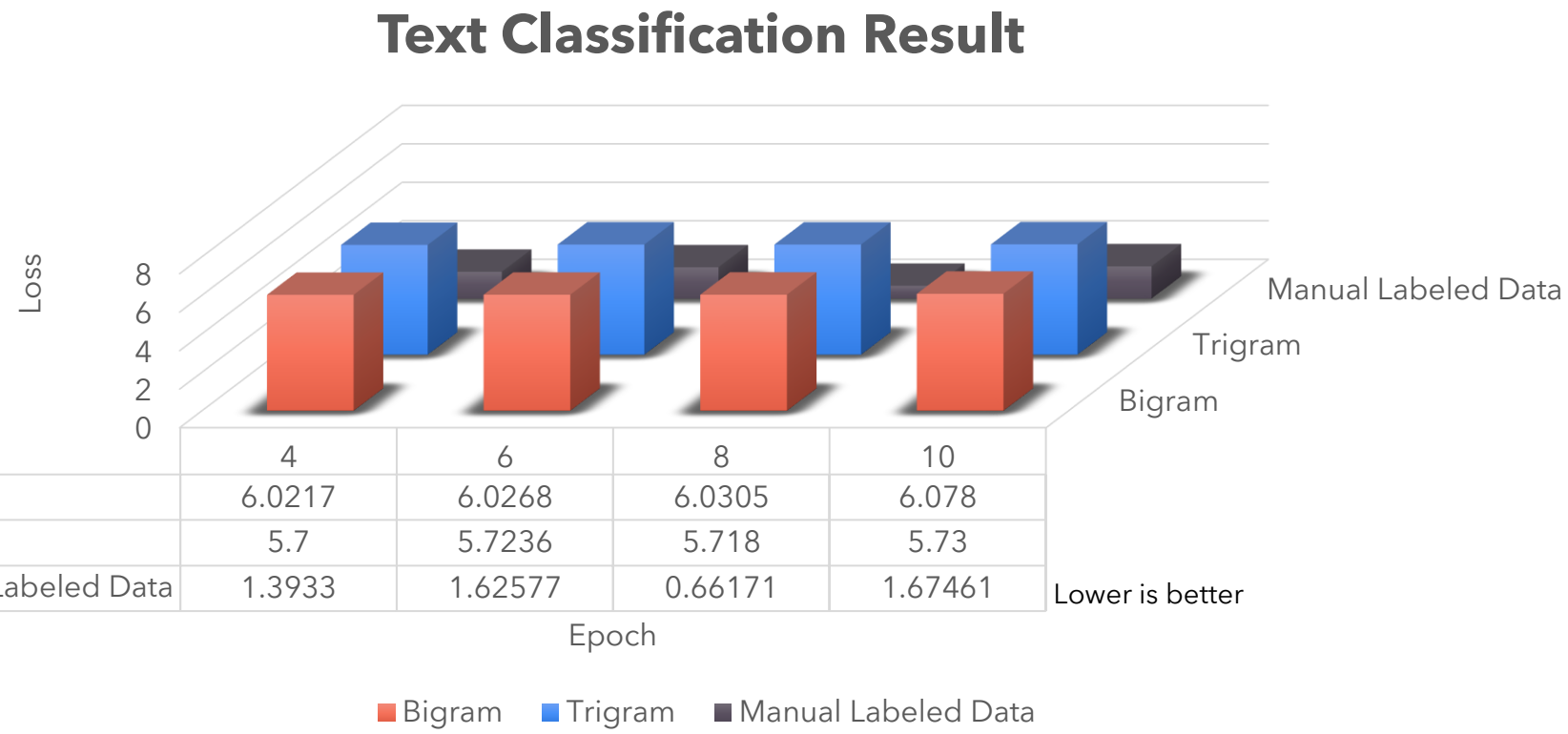
Experiment And Analysis



	4	6	8	10
Bigram	0.0112	0.0123	0.0112	0.0123
Trigram	0.0169	0.0169	0.0169	0.0169
Manual Labeled Data	0.4651	0.24192	0.8868	0.224321

Higher is better

Experiment And Analysis



	4	6	8	10
Bigram	6.0217	6.0268	6.0305	6.078
Trigram	5.7	5.7236	5.718	5.73
Manual Labeled Data	1.3933	1.62577	0.66171	1.67461

Lower is better

Experiment And Analysis

Threshold Test Result on Bigram

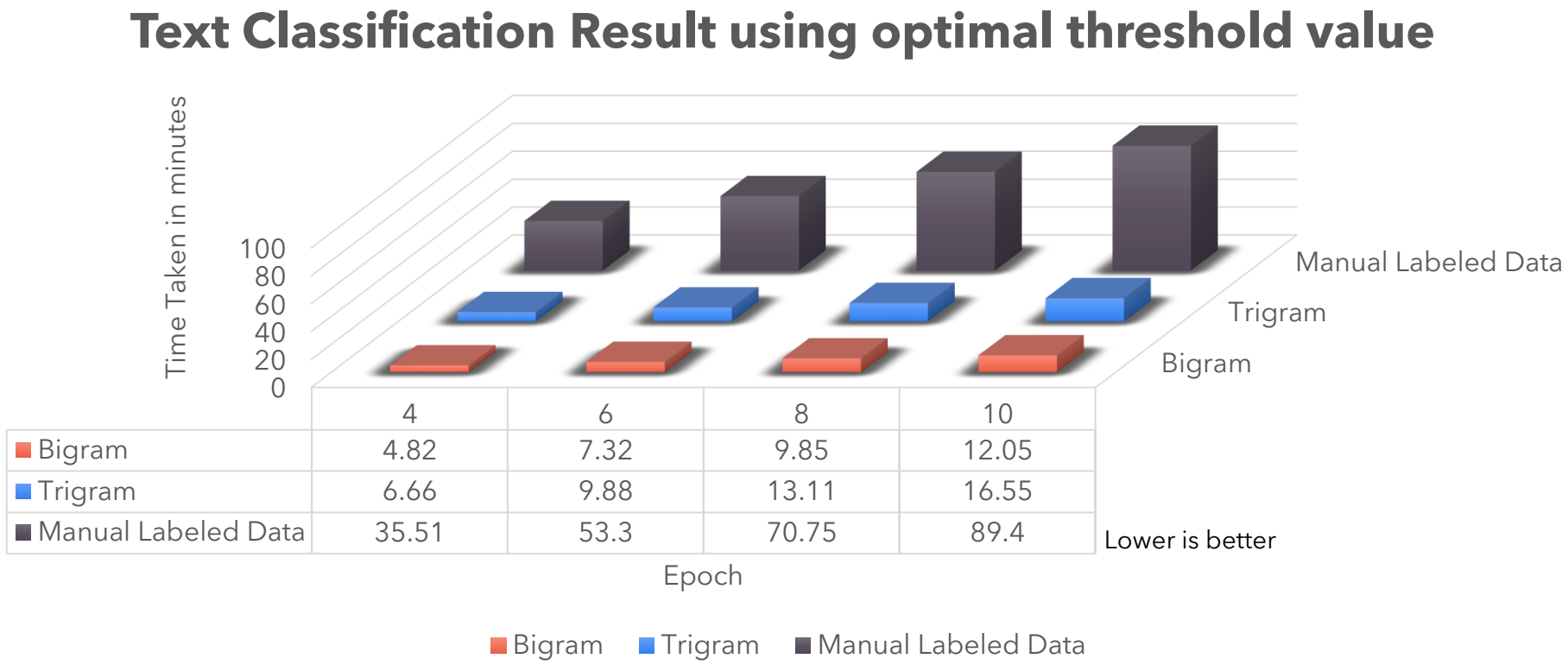
Threshold	Number of Community Formed	Modularity	Class Split	Class Merge
0.0	1007	0.559	365.833	2.171
0.1	311	0.69251	161.1666	3.090
0.2	194	0.827	101.833	3.118
0.3	350	0.8009	210	3.582
0.4	348	0.624	134.834	2.3074
0.5	97	0.425	35.5	2.134
0.6	22	0.2698	13.666	3.4545

Experiment And Analysis

Threshold Test Result on Trigram

Threshold	Number of Community Formed	Modularity	Class Split	Class Merge
0.0	706	0.5793	309.330	2.626
0.1	283	0.7637	144.334	3.03886
0.2	447	0.82031	286	3.8255
0.3	789	0.6225	236	1.787
0.4	83	0.4702	42.664	3.012
0.5	55	0.2837	21	2.1818
0.6	12	0.168	9.666	4.333

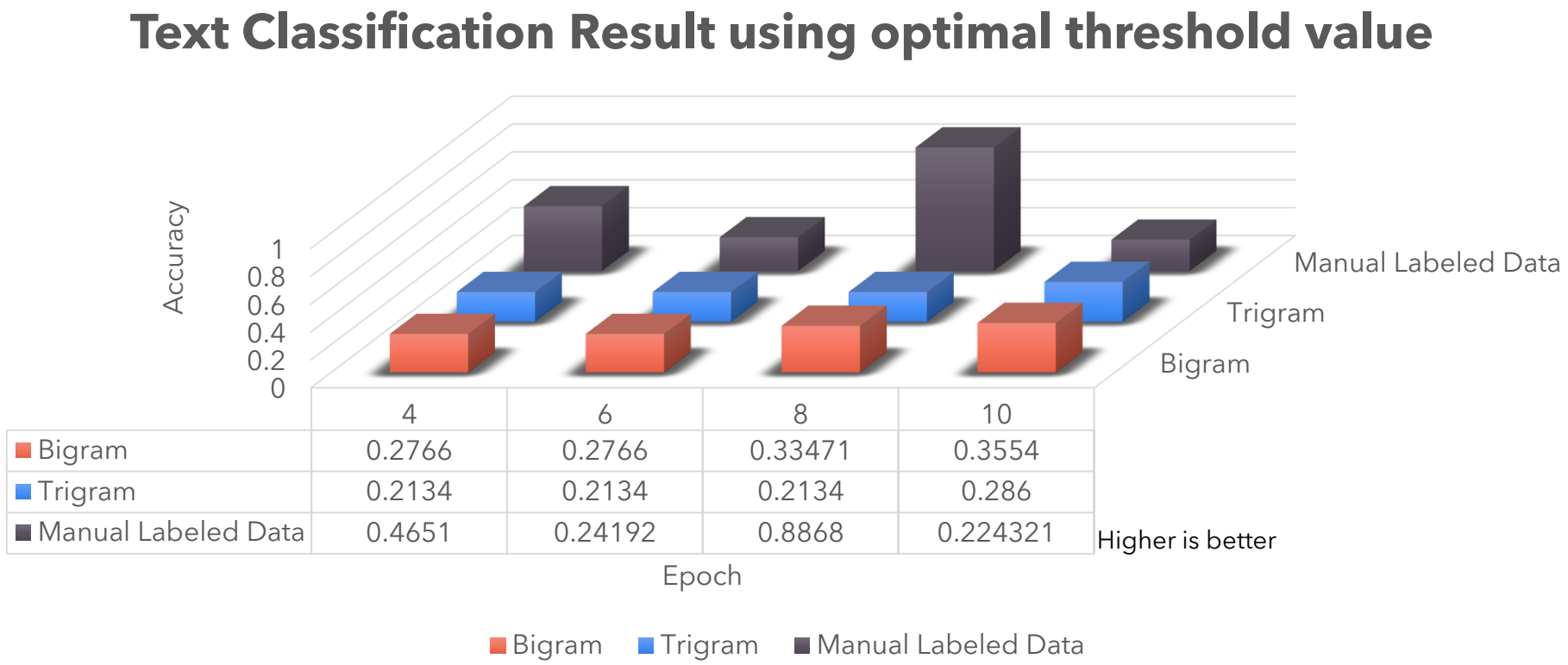
Experiment And Analysis



	4	6	8	10
Bigram	4.82	7.32	9.85	12.05
Trigram	6.66	9.88	13.11	16.55
Manual Labeled Data	35.51	53.3	70.75	89.4

Lower is better

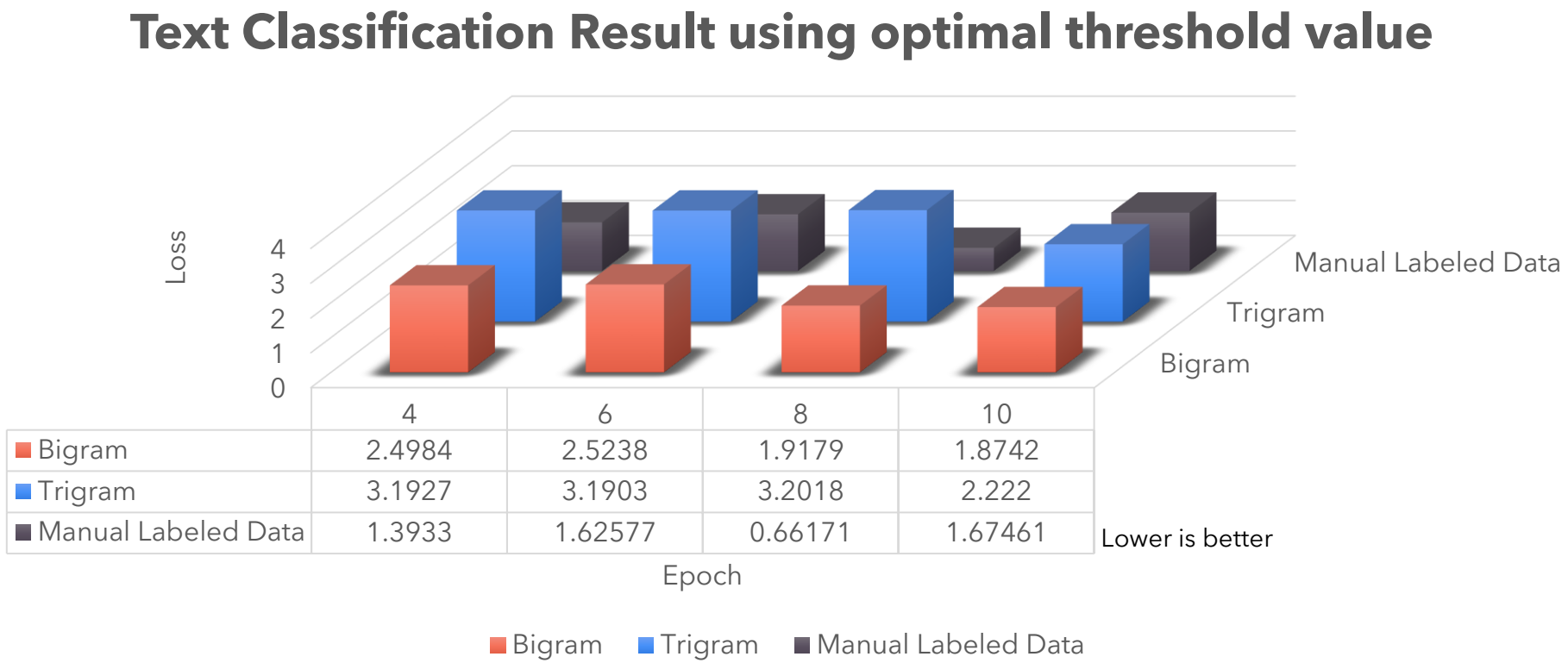
Experiment And Analysis



	4	6	8	10
Bigram	0.2766	0.2766	0.33471	0.3554
Trigram	0.2134	0.2134	0.2134	0.286
Manual Labeled Data	0.4651	0.24192	0.8868	0.224321

Higher is better

Experiment And Analysis



	4	6	8	10
Bigram	2.4984	2.5238	1.9179	1.8742
Trigram	3.1927	3.1903	3.2018	2.222
Manual Labeled Data	1.3933	1.62577	0.66171	1.67461

Lower is better

Experiment and Analysis

- Other parameters that have effects on the outcome
  - Nodes without community:

    Another thing that needs to be considered when using automatic labeling is that there are nodes that do not have a community, so a significant amount of data is lost or we can call it there is reduction in community.

    When using the optimum threshold, it is found that training data that has a community is 1446 on bigram and 2010 data on the trigram of the total training data of 5452. This means that 73.4% of the training data is on bigram, and 63.13% of the training data is on the trigram has no community.

Experiment and Analysis

- Other parameters that have effects on the outcome
  - ➢Node with many community:

    With the class merge value of 2.623188 on bigram data and 3.219 on trigram data, there are still communities that group 6 manual labels in one community. These nodes do not have a close cosine similarity when explored further, but by the Infomap algorithm, they are grouped into one community.
  - ➢Different Threshold Value in bigram and trigram.
  - ➢Use of keywords.
  - ➢Use of synonyms.

Conclusion

Using community detection for automatic labeling allows us to use data that is not community based. Therefore, the experiment significantly reduced the amount of training data, i.e., 73.4% of the training data on bigram and 63.13% of the training data on the trigram.

As a result, manually labeled classification still has higher accuracy.

Future Work

If we can minimize the class split and class merge we can significantly improve the performance of this model.

Therefore, data handling techniques that does not have community by considering semantic analysis so that sentences can be labeled as whole might improve this model beyond what has been accomplished.

Thank you Everyone

Presenter name: Alvi Ahmmed Nabil

Roll No: 1707009

Department of Computer Science and Engineering

Khulna University of Engineering & Technology, Khulna, Bangladesh