Text Summarization Using Restricted Boltzmann Machine: Unsupervised Deep Learning Approach

Ashwini Ambekar[1], Kajol Shah[2], Minakshi Agrawal[3], Simica Pawar[4], Asma Shaikh[5]

[1, 2, 3, 4, 5] Dept of Computer Engineering

[1, 2, 3, 4, 5] Marathwada MitraMandal's College of Engineering, Pune.

Abstract- *Amount of information available on the internet is increasing day by day. A lot of time is required by the user to go through this information or documents. It is difficult for humans to manually summarize the information in these documents. Here, automatic text summarization comes in use. Text summarization is a method of automatically generating a compressed version of the original document. Based on this summary, the user can decide whether or not the document is worth reading and is relevant to his or her topic instead of going through a whole bunch of documents. In this paper, a generalized and query-oriented summary generation method for a single document is proposed.*

Keywords- Stemming, Stop-word removal, Part of Speech Tagging, Title Similarity, Sentence Matrix, Feature Vector Extraction, Inverse sentence frequency, Term weight, Positional feature, Restricted Boltzmann Machine (RBM), Sentence score.

I. INTRODUCTION

Text Summarization is a method which produces a compressed version of the original document using unsupervised Deep Learning. In unsupervised learning, the input data is not labeled. It is a type of machine learning algorithm, in which inferences are drawn from datasets consisting of unlabeled input data. Deep learning algorithm can be applied to unsupervised learning tasks. In this paper, Restricted Boltzmann Machine (RBM) is proposed which is a type of unsupervised deep learning. Using this method, user can save his or her time and read only the documents which are related to his or her topic of interest. There are two types:

1. **Extractive Text Summarization**
2. **Abstractive Text Summarization**

Extractive Text Summarization: In this method, text summarization based on some features like title similarity, term frequency, sentence ranking etc are extracted to form summary.

Abstractive Text Summarization: In this method, paraphrasing the sentences of the original document is done to generate the summary.

The method proposed in this paper is for Extractive Text Summarization and it is a single document, query oriented in which the user can enter his or her query to generate summary according to user's interest.

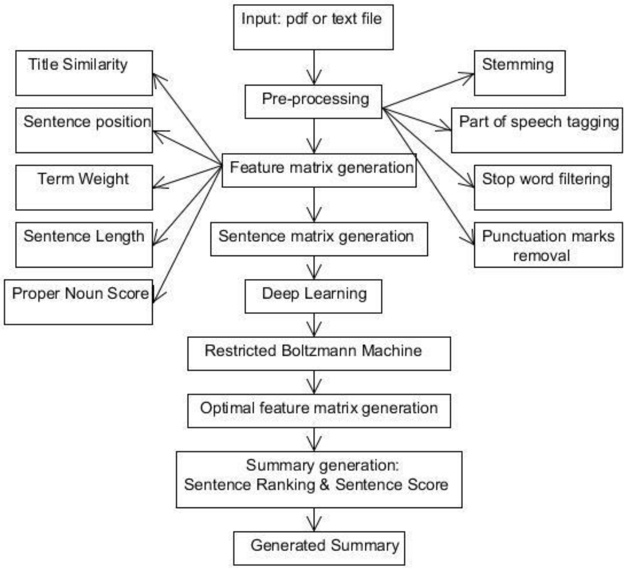


Fig 1: System Architecture

II. PREPROCESSING

In this phase, the words which are not important and unwanted are removed. This is done in order to structure the document by applying myriads of techniques by which the density of document is reduced and document is made light weight. This makes the further processing of document easier. Following are the techniques used here:

2.1 Stop-word removal

The words which are not important and don't have any meaning on their own are filtered in this step. These words are called stop-words. For a particular word to be a stop-word, there is no specific rule. It depends upon the situation and is completely subjective. Usually articles, prepositions, etc are considered as stop words and are removed. Here, we have considered words such as a, an, the, is, are, on, etc. as stop-words.

2.2 Part of Speech Tagging

Categorizing the words of text on the basis of part of speech (noun, adverb, verb, adjective etc.) is called as Part of Speech Tagging. It is difficult because some words can represent more than one part of speech at different times and some parts of speech are complex or unspoken. So, in this phase, the words are marked corresponding to a part of speech.

2.3 Stemming

In stemming, the word is brought to its base or root form. Example: using singular form of words like using boys as boy or swimming reduced to swim. Here, we have considered that except the words which belong to the proper nouns category, all other category's words can be stemmed The stem need not be identical to the morphological root of the word.

2.4 Punctuation Marks Removal

The punctuation marks like ", . : ; ? / " etc. from the document are removed in this pre-processing step, hence making the document light-weight which further simplifies the summary generation process.

III. FEATURE VECTOR EXTRACTION

The document which is made light weight in the preprocessing phase is now structured into a matrix. A sentence matrix M of order n*v contains the features for every sentence of a matrix. Here, 'n' is the number of sentences in the document and 'v' is the number of features.

Four features are extracted of a sentence of text document namely:

1. Title Similarity
2. Positional Feature
3. Term Weight
4. Sentence Length
5. Proper Noun Score

3.1 Title Similarity

A sentence in the document is said to be important for the summary if it is similar to title of the document. Similarity is calculated based on the common words occurring in the title of the documents and sentences of the document. The title similarity is calculated using the sentence score which is defined as the ratio of number of common words occurring between the title and the sentences in the document to the total number of words of the document. The feature sentence of a sentence is said to be good if it has maximum number of words common to the title. [2] Mathematically represented as:

$$S = (w_t \cap w_s)/ w_t$$

Where,

w_t = Number of words in the title
w_s = Number of words in the sentence
$w_t \cap w_s$ = Common words in the title and the sentence

3.1 Sentence Position

The position of the sentence can determine the relevance of the sentence for the summary. Usually the sentences that appear in the starting and ending of the text are of more importance. So, based on this the sentence score is calculated. The positional score in our case is calculated by considering the following conditions:

p2 = 1, if the sentence appears in the starting part of the text

p2 = 0, if the sentence appears in the middle part of the sentence

p2 = 1 if the sentence appears in the last part of the sentence
Here, we have considered 20% of sentences from the start and the end of the text as important and marked their p2 value as 1.

3.2 Term Weight

Term weight means the term frequency and its importance. The term frequency gives the total number of times the term has occurred in the whole document which depicts the importance of the term in a document. The term frequency is represented by tf(f,d) where, f is the frequency of the word and t is the text document. The inverse sentence frequency(isf) tells whether the term is common or rare across the document. The total term weight is calculated by computing the se two concepts:

1. Term Frequency, tf(f,d) and
2. Inverse Sentence Frequency represented as isf.

3.4 Sentence Length

The sentence length decides the importance of the sentence in summarization. The sentences that are too short do not give much information about the document. Whereas sentences that are too long will have unnecessary information about the document that will not be useful for summarization. We have discarded the sentences that have words less than 3 as they will not be able to prove useful for summarization.

3.5 Proper Noun Score

In the process of summary generation, important role is played by the Proper Nouns. It gives information regarding, to whom or to what the author is referring. Roles played by individuals or locations description will be different more number of times in a document. Here, the number of words which are proper noun are counted.

IV. FEATURE MATRIX GENERATION

The above calculated features' values are then stored in a matrix form where the columns represent the features and rows represent the sentences.

V. ALGORITHM FOR DEEP LEARNING

Here, we are using Restricted Boltzmann Machine(RBM) for deep learning. The sentence matrix containing a set of feature vectors is given as an input to the RBM phase as a visible layer.[2]

Let S be a set of sentences

$$S=(s1,s2,\ldots.,sn)$$

where,

$si = (f1,f2,\ldots\ldots..f4), i<= n$

where, n is the number of sentences in the document. Restricted Boltzmann Machine contains two hidden layers and for them two set of bias value is selected namely H0 and H1:

1. $H0=\{h1,h2,\ldots,hn\}$
2. $H1=\{h1,h2,\ldots,hn\}$

These are sets of bias values which are randomly selected, the whole operation is performed with these two sets During the first cycle of RBM, a new refined matrix is generated:

$$s'=(s'1,s'2,\ldots.,s'n)$$

It is calculated by: $\sum_{1}^{n}si + hi$

During step 2 the same procedure will be applied to this obtained refined set to get the more refined sentence matrix set with H1 and which is given by:

$$s''=(s''1,s''2,\ldots.,s''n)$$

After obtaining the refined sentence matrix from the RBM it is further tested on a particular randomly generated threshold value for each feature we have calculated.

Step 1:

$$[f1,f2,f3,f4]\ [f1,f2,f3,f4]\ [f1,f2,f3,f4]=> \sum_{1}^{n}si + hi(H0)=>s'=(s'1,s'2,\ldots.,s'n)$$

Step 2:

$$[f1,f2,f3,f4]\ [f1,f2,f3,f4]\ [f1,f2,f3,f4]=> \sum_{1}^{n}si + hi(H1)=>s''=(s''1,s''2,\ldots.,s''n)$$

Here, we can increase the number of hidden layers in RBM.

VI. ENHANCED FEATURE MATRIX

An Enhanced feature matrix is obatined from the deep learning phase which is now used for the further summary generation phase.

VII. GENERATION OF SUMMARY

Now, in two ways summary can be generated. One is a generalized summary of the whole document and the other way is based on the user query entered by the user(using the sentence score).

7.1 Sentence Score

This step is performed only for user-query based summary generation. Here, the ratio of the number of words which are common in the sentence and the query entered by the user to the total number of words in the document are calculated .[2]

$$S_c = (S_w \cap S_u)/T_w$$

where,

S_c =sentence score
S_w =words in the sentences
S_u =words in the user query
$S_w \cap S_u$ =number of words common in the sentence and user query

7.2 Sentence Ranking

The number of sentences which should be their in the summary is calculated by

N=0.3* total number of sentences in the text

Here, we are considering 30% sentences of the document in the summary.

Sentence ranking for generalized summary:

The sum of the feature values for each sentence in the enhanced matrix is calculated. Now based on these values, the sentences are arranged in descending order(sentences having higher feature sum). Now, the top N sentences are selected from this sorted order of sentences.

Sentence ranking for user-query based summary:

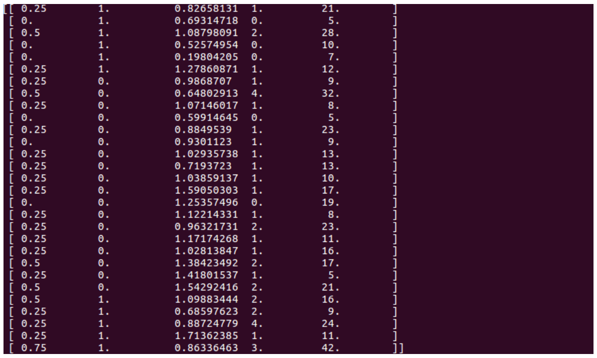
Here, based on the sentence score, sentences are arranged in descending order. Now, the top N sentences are selected from this sorted order of sentences.

7.3 Summary generation

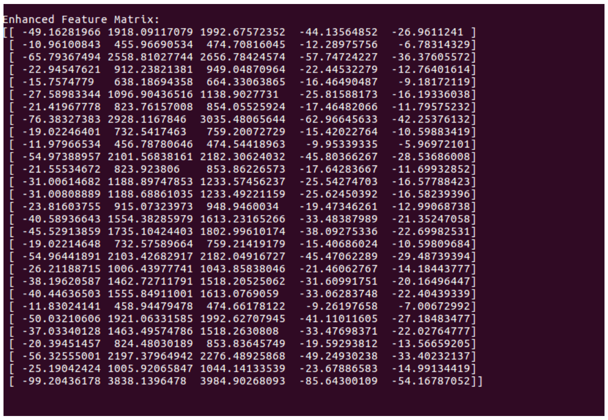
The N sentences selected are now arranged according to their position in the text and then displayed to the user. In this way, extractive summary is generated.

VIII. SCREEENSHOT OF FEATURE MATRIX AND ENHANCED MATRIX GENERATION FOR A SAMPLE TEXT

Feature matrix generated for a sample text:



```
[[ 0.25        1.          0.82658131  1.          21.         ]
 [ 0.          1.          0.69314718  0.           5.         ]
 [ 0.5         1.          1.08798091  2.          28.         ]
 [ 0.          1.          0.52574954  0.          10.         ]
 [ 0.          1.          0.19804205  0.           7.         ]
 [ 0.25        1.          1.27860871  1.          12.         ]
 [ 0.25        0.          0.9868707   1.           9.         ]
 [ 0.5         0.          0.64802913  4.          32.         ]
 [ 0.25        0.          1.07146017  1.           8.         ]
 [ 0.          0.          0.59914645  0.           5.         ]
 [ 0.25        0.          0.8849539   1.          23.         ]
 [ 0.          0.          0.9301123   1.           9.         ]
 [ 0.25        0.          1.02935738  1.          13.         ]
 [ 0.25        0.          0.7193723   1.          13.         ]
 [ 0.25        0.          1.03859137  1.          10.         ]
 [ 0.25        0.          1.59050303  1.          17.         ]
 [ 0.          0.          1.25357496  0.          19.         ]
 [ 0.25        0.          1.12214331  1.           8.         ]
 [ 0.25        0.          0.96321731  2.          23.         ]
 [ 0.25        0.          1.17174268  1.          11.         ]
 [ 0.25        0.          1.02813847  1.          16.         ]
 [ 0.5         0.          1.38423492  2.          17.         ]
 [ 0.25        0.          1.41801537  1.           5.         ]
 [ 0.5         0.          1.54292416  2.          21.         ]
 [ 0.5         1.          1.09883444  2.          16.         ]
 [ 0.25        1.          0.68597623  2.           9.         ]
 [ 0.25        1.          0.88724779  4.          24.         ]
 [ 0.25        1.          1.71362385  1.          11.         ]
 [ 0.75        1.          0.86336463  3.          42.         ]]
```



```
Enhanced Feature Matrix:
[[ -49.16281966 1918.09117079 1992.67572352  -44.13564852  -26.9611241 ]
 [ -10.96100843  455.96690534  474.70816045  -12.28975756   -6.78314329]
 [ -65.79367494 2558.81027744 2656.78424574  -57.74724227  -36.37605572]
 [ -22.94547621  912.23821381  949.04870964  -22.44532279  -12.76401614]
 [ -15.7574779   638.18694358  664.33063865  -16.46490487   -9.18172119]
 [ -27.58983344 1096.90436516 1138.9027731   -25.81588173  -16.19336038]
 [ -21.41967778  823.76157008  854.05525924  -17.46482066  -11.79575232]
 [ -76.38327383 2928.1167846  3035.48065644  -62.96645633  -42.25376132]
 [ -19.02246401  732.5417463   759.20072729  -15.42022764  -10.59883419]
 [ -11.97966534  456.78780646  474.54418963   -9.95339335   -5.96972101]
 [ -54.97388957 2101.56838161 2182.30624032  -45.80366267  -28.53686008]
 [ -21.55534672  823.923806    853.86226573  -17.64283667  -11.69932852]
 [ -31.00614682 1188.89747853 1233.57456237  -25.54274703  -16.57788423]
 [ -31.00808889 1188.68861035 1233.49221159  -25.62450392  -16.58239396]
 [ -23.81603755  915.07323973  948.9460034   -19.47346261  -12.99068738]
 [ -40.58936643 1554.38285979 1613.23165266  -33.48387989  -21.35247058]
 [ -45.52913859 1735.10424403 1802.99610174  -38.09275336  -22.69982531]
 [ -19.02214648  732.57589664  759.21419179  -15.40686024  -10.59809684]
 [ -54.96441891 2103.42682917 2182.04916727  -45.47062289  -29.48739394]
 [ -26.21188715 1006.43977741 1043.85838046  -21.46062767  -14.18443777]
 [ -38.19620587 1462.72711791 1518.20525062  -31.60991751  -20.16496447]
 [ -40.44636503 1555.84911001 1613.0769059   -33.06283748  -22.40439339]
 [ -11.83024141  458.94479478  474.66178122   -9.26197658   -7.00672992]
 [ -50.03210606 1921.06331585 1992.62707945  -41.11011605  -27.18483477]
 [ -37.03340128 1463.49574786 1518.2630808   -33.47698371  -22.02764777]
 [ -20.39451457  824.48030189  853.83645749  -19.59293812  -13.56659205]
 [ -56.32555001 2197.37964942 2276.48925868  -49.24930238  -33.40232137]
 [ -25.19042424 1005.92065847 1044.14133539  -23.67886583  -14.99134419]
 [ -99.20436178 3838.1396478  3984.90268093  -85.64300109  -54.16787052]]
```

Enhanced Feature matrix generated for a sample text:

IX. RESULT

The summary produced by the above text summarization method using unsupervised deep learning algorithm is better than the summary produced by methods using supervised learning techniques as no training dataset is required in the proposed approach. Hence, the time and cost required to train the dataset is saved. So, this method is more efficient.

X. EVALUATION METRICS

The proposed text summarization method can be evaluated using three basic evaluation criteria:

10.1 RECALL

The reliability of the proposed approach is measured using recall.It is ratio of the difference between retrieved sentences and relevant sentences to retrieved sentences.

$$\text{Recall} = (S_{ret\,-}\,S_{rel})/S_{ret}$$

where,
S_{ret} = number of sentences retrieved
S_{rel} =number of relevant sentences

10.2 PRECISION

It is used to measure how precise summary is produced by the proposed approach. It is ratio of the difference between retrieved sentences and relevant sentences to relevant sentences.

$$\text{Recall} = (S_{ret\,-}\,S_{rel})/S_{rel}$$

10.3 F-MEASURE

F-measure value can be calculated by:
F-measure=(2*Recall* Precision)/(Recall + Precision)

XI. CONCLUSION

In this way, two kinds of summary can be generated using the above mentioned approaches.

XII. FUTURE SCOPE

This method can be further improved for generating summaries of books of large volumes. Also, we can add some features to this summarization method to generate user-query based summary of medical reports.

REFERENCES

[1] M Yousefi-Azar, ”Text Summarization using unsupervised deep learning”, Science Direct, Volume:68, 2017, Page:93-105

[2] PadmaPriya G. and 2K. Duraiswamy , ”An Approach for Text Summarization using deep learning approach”, Journal of Computer Science, Volume:10, 2014, Page:1-9