Title Generation of Talk Shows

Sobia Dastgeer
01-243182-019

A thesis submitted in fulfilment of the
requirements for the award of degree of
Masters of Science (Computer Science)

Department of Computer Science

BAHRIA UNIVERSITY ISLAMABAD

OCTOBER 2020

Approval for Examination

Scholar's Name : Sobia Dastgeer Registration Number: 59365

Enrollment: 01-243182-019

Program of Study: MSCS

Thesis Title: Title Generation of Talk shows

It is to certify that the above scholar's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for examination. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index 15 %. that is within the permissible limit set by the HEC for the MS/M.Phil degree thesis. I have also found the thesis in a format recognized by the BU for the MS/M.Phil thesis.

Principal Supervisor Signature:

Date: 15-10-2020

Name: Dr Muhammad Asfandeyar

Author's Declaration

I, Sobia Dastgeer hereby state that my MS thesis titled

"Title Generation of Talk shows

"

is my own work and has not been submitted previously by me for taking any degree from this university.

Bahria University Islamabad

or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw/cancel my MS degree.

Name of Scholar: Sobia Dastgeer

Date: 15-10-2020

Plagiarism Undertaking

I, solemnly declare that research work presented in the thesis titled

„ Title Generation of Talk shows

_____

_____”

is solely my research work with no significant contribution from any other person. Small contribution / help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Bahria University towards plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred / cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS/M.Phil degree, the university reserves the right to withdraw / revoke my MS/M.Phil degree and that HEC and the University has the right to publish my name on the HEC / University website on which names of scholars are placed who submitted plagiarized thesis.

Scholar/Author's Sign: Sobia

Name of Scholar: Sobia Dastgeer

Dedication

To my beloved mother and father

Acknowledgements

In the name of Allah, the most beneficial, the most merciful. First of all I would like to thank Allah Almighty for giving me strength and courage to complete this research and achieve this milestone. No doubt without his will I was not be able to do anything.It is my pleasure to acknowledge my deepest thanks and gratitude to my Supervisor Dr. Muhammad Asfand-e Yar for his guidance and kind supervision during my research.

I am also very Thankful to Dr Awais Ahmad and Dr Ibrar for their encouragement and moral support throughout my MS work.
I would like to thank the Department of Computer Science for providing me the environment and resources for completing the thesis tasks.Finally, I would like to express my very profound gratitude to my parents for always encouraging me to reach higher in order to achieve my goals.

Abstract

An extraordinary video title describes the most notable occasion compactly and catches the watcher's consideration.The Talk shows have manually generated titles and may be the title of video is not matched with the content of video. Due to incorrect titles or the content mismatch mostly the views are less.This research study proposes a technique for title generation. This research study propose a technique for generating the title using NLP, and Deep learning techniques. The lda2vec and Long Short term memory are used to generate the titles of talk shows. We have compared the performance of LSTM model with lda2vec model.The proposed technique is validated on Custom dataset and all the news dataset and reported good results.The LSTM model generate good title as compare to Lda2vec model. The LSTM model acheives higher accuarcy as compare to lda2vec model if we increase the number of epochs to train the model.

TABLE OF CONTENTS

AUTHOR'S DECLARATION ii

PLAGIARISM UNDERTAKING iii

DEDICATION iv

ACKNOWLEDGEMENTS v

ABSTRACT vi

LIST OF TABLES ix

LIST OF FIGURES x

LIST OF SYMBOLS xi

1 INTRODUCTION xii

1.1 Introduction xii
1.2 Problem Description and objectives 1
1.3 Research Contribution 1
1.4 Thesis Organization 1

2 RELATED WORK 2

2.1 Introduction 2
2.2 Text summarization 2
2.3 Title Generation Approaches 3
2.3.1 Naive Bayes approach with restricted vocabulary 3
2.3.2 Naive Bayes approach with full vocabulary 3
2.3.3 Term frequency (TF) and inverse document frequency (Idf) approach 3
2.3.4 K nearest neighbor approach 3
2.3.5 Iterative Expectation-Maximization approach 4
2.4 Title Generation Techniques 4
2.4.1 Machine learning techniques 4
2.4.2 Natural Language Processing techniques 7
2.4.3 Title generation using Implicit and explicit Relations 9
2.4.4 Title generation using Extracted keywords 9
2.4.5 Title generation using Neural network 10

3 METHODOLOGY **12**
3.1 Dataset description 12
3.1.1 Traning dataset 12
3.1.2 Testing dataset 13
3.2 Proposed Methodology 13
3.2.1 Data Preprocessing 13
3.2.2 Feature Extraction 14
3.2.3 Overview of LDA 14
3.2.4 Overview of Word2vec 15
3.2.5 Workflow of Title generation using lda2vec 16
3.2.6 Overview of RNN 16
3.2.7 Overview of LSTM 17
3.2.8 Workflow of Title generation using LSTM 18

4 ANALYSIS & RESULTS **19**
4.1 Experimental Protocol and Results 19
4.1.1 Title Generation using LDA 19
4.1.2 Title generation using LSTM 23
4.2 Performance Analysis and Discussion 24
4.2.1 Generated Titles with LDA+word2vec 25
4.2.2 Generated Titles with LSTM 25

5 CONCLUSION & FUTURE WORK **26**
5.1 Conclusion 26
5.2 Perspective 27

REFERENCES **27**

LIST OF TABLE

2.1 Summary of machine learning based techniques 5
2.2 Summary of Natural Language Processing based techniques 7
2.3 Summary of Implicit and explicit based techniques 9
2.4 Summary of Neural network based techniques 11

3.1 Statistics of All the news Dataset 13
3.2 Statistics of custom Dataset 13

4.1 All the news Dataset division 19
4.2 Accuracy rate of title generation using LDA+word2vec 24
4.3 Accuracy rate of title generation using LSTM 24

LIST OF FIGURE

3.1 A sample Image from All the news dataset 12
3.2 The architecture of Latent Dirichlet Allocation Model[31] 15
3.3 Workflow of Title generation using lda2vec 16
3.4 Simple architecture of RNN[8] 17
3.5 Architecture of LSTM[8] 18
3.6 Workflow of Title generation using LSTM 18

4.1 Word Frequency And Distribution 20
4.2 Representation of most discussed topic in the Document 20
4.3 Python library for interactive topic model visualization in the Document . 21
4.4 word vectors of frequent topics in the document 22
4.5 Accuracy rate of Generated title using lda2vec 23
4.6 No of parameter used for training LSTM 23
4.7 Accuracy rate of Title Generation using LSTM 24
4.8 Title Generated using LDA+word2vec 25
4.9 Generated Title using RNN-LSTM 25

Acronyms and Abbreviations

NLP	Natural Language Processing
ML	Machine Learning
LDA	Latent Dirichlet Allocation
POS	part of speech
RNN	Recurrent Neural network
LSTM	Long short term memory
TF	term frequency
IDF	inverse document frequency
KNN	K nearest neighnopur
CRF	Conditional random fields
SVM	support vector machine
DL	Deep learning

CHAPTER 1

INTRODUCTION

1.1 Introduction

Sentences of every language contains different part of speech according to their grammatical structure. English language sentences consist of eight parts of speech i.e. noun, pronouns, verbs, adverbs, adjectives, preposition, conjunction and interjection [1]. POS tagging is the process of assigning the tags to words with their respective part-of speech. Part of speech tagging is the main problem in natural language processing to tag the word correctly. POS tagger use in various problems like text generation, titling of text and speech recognition. Various research studies were conducted for POS tagging in offline documents as well as in YouTube videos where different techniques were employed in order to improve the accuracy. Due to lack of standard benchmark of talk shows this area could not gain the attention of research community. Discussion videos of talk shows do not provide the title, so if the listener wants to listen the specific news he has no idea whether it is included in the discussion or not. So to overcome this problem, our proposed method will generate title for that videos. The latent dirichlet allocation finds out the topics and word probability of every word from topic occurring in the document then we train the shallow neural network for word embedding. The LDA model considers the likelihood circulation of driving conduct in the driving example and accepts driving practices as free factor by utilizing the measurable technique for driving behavior words in driving information [2].We give the topic and highest probability words of every topic to one layer of shallow neural network and then they give the vector of that number. Finding the cosine similarity between the vectors of numbers to and sum of the mean of cosine similarity of ever topic vector and ranked them. The highest ranked topic vector is given as a title of that recorded videos.A large number of various news reports from various channels are introduced to us. In [3] they use LDA technique n to decide which news belong to huge news features headlines gathered from news destinations. Many researches use part of speech tagging to tag the words as noun, verb and adverb then use these words applying LDA and other techniques to generate the title. If POS do not tag the word correctly then ultimately the generated title is not correct.In [4]the issue of grammatical feature labeling in the area of ungrammatical expressions solved. They investigate the quality of existing grammatical feature taggers, at that point propose and assess conceivable improvements through the incorporation of ungrammatical POS tagged sentences into the preparation text corpus. Through part of speech tagging many approaches are used to improve accuracy[5]. The First approach is rule based approach which is based

on non-automatic approach. So, therefore its very time consuming approach. The other approach is stochastic approach, by using this approach the accuracy is lesser than rule based approach. In our proposed model we use the recurrent neural network word2vec technique to generate the title of videos. In this study, we will not use POS tagging because the accuracy of generating the title of videos is affected if the POS not tag the words correctly.In the space of text similarity, corpus-based methodology has overcome the fundamental issue in Natural language processing, accomplishing human-competitive accuracy.unfortunately, a large portion of the past text similarity didn't consider the implanting significance behind the words. Embedding meaning is amazingly helpful when managing documents have a similar length but use different words[6]. So to overcome this problem, we use latent Dirichlet allocation and word2vec to generate the title. We will compare the performance of traditional NLP techniques with deep learning technique (RNN-LSTM).

1.2 Problem Description and objectives

The Talk shows have manually generated titles and videos is suggested based on a title of a video. The problem is a mismatch of title with the contents of talks shows, after watching complete video may be the title of video is not related with the content of video. This research focuses an effective technique for title generation. This study will significantly contribute to resolve the problem. The objective of this research study covers two major parts i.e. custom development of data from recorded talk shows discussion as well to propose an effective methodology for title generation of the talk shows using LSTM and LDA2vec techniques. To highlight the effectiveness of LSTM for title generation over traditional approaches i.e. LDA2vec. We will employed LSTM model and will compare their performance with LDA2vec model.

1.3 Research Contribution

To develop a custom dataset from talk shows and will make it public for research community in order to extend the research in this domain. Secondly, we have used the LSTM a Deep Learning technique to generate the titles automatically after extracted the text from talk shows.

1.4 Thesis Organization

The thesis is organized as follows.Chapter 2 presents a discussion on the related work that has been done for title generation.Chapter 3 introduces the data set preparation and proposed methodology of the present study. Chapter 4 presents a comprehensive discussion of results. Conclusion and perspectives are presented in chapter 5.

CHAPTER 2

RELATED WORK

2.1 Introduction

A lot of work has been done in the field of titling the text. The work has been done by using part of speech tagging and latent dirichlet allocation and many other techniques. There are various methods to generate the title using part of speech tagger and Latent dirichlet allocation. In [7] they uses different techniques to tag the words using part of speech tagging but it requires more training and no one model generate the tag correctly. But the first article [8] using word2vec and recurrent neural network for Turkish language to generate the title. Our propose model generate the title of videos of talk shows in English language by using latent Dirichlet allocation+word2vec and RNN-LSTM to generate the title.

2.2 Text summarization

Text information are accessible as a large document. seeing huge content records and removing significant data out of now is the time consuming assignments. summarized text will have decreased size when contrasted with original one. In [9] they have attempted to highlight significant methods for removing significant data from a given text with the assistance of topic modeling, key expression extraction and summary generation.Automatic text summarization decrease the number of texts without losing any information and help you to perceive the meaning of the text. Automatic text summarization has two methodologies: extractive methodology and abstractive methodology. The extractive methodology is a strategy that removes significant words or significant sentences from text as indicated by a scoring system. The extractive methodology has limitations. Initially, created theoretical comprises of sentences in an original text and the sentence can't be short regardless of whether it incorporates repetitive articulations.Second, when the extractive methodology extract words from original text, it is difficult to build a characteristic sentence from the extracted words. Then again, the abstractive methodology is a technique that produces abstracts from original text.In Abstractive summarization, the first content gets changed over into another more justifiable semantic structure to get a shorter summary of original content document.In [10] examines abstractive content summarization procedures and features the parametric assessment of these procedures. The methodology can produce a dynamic abstract but it needs progressively normal language handling strategies content to develop a theoretical and it includes some unsolved

issues[11]. Now a days the automatic generative text summarization is acknowledged with a neural network motivated from machine translation.

2.3 Title Generation Approaches

To make a title for a document is a difficult task. To produce a title for a verbally expressed record turns out more challenging because we need to manage word errors and ambiguities created by speech recognition.[12] In [13] depicts a novel way to produce the title for Hindi stories, article or section consequently. They executed a framework which is sufficiently skilled to propose proper titles that are applicable to the story. By using POS tagging they generate the title of Hindi story.The six different title generation approaches are

2.3.1 Naive Bayes approach with restricted vocabulary

It attempts to capture the relationship between the words in the original document and the words of the title[12]. For every word of document, it tallies the event of title word same as document word and apply the measurements to the test reports for creating titles.In [14] document classification is achieved by Naive Bayes approach which is the method of machine learning. . These reports, which incorporated financial, health, sports,magazine news and political were basically exposed to explicit pre-handling steps and then trained. The model got from the preparation procedure was utilized in the testing procedure and successful results were achieved.

2.3.2 Naive Bayes approach with full vocabulary

It relaxes the limitation in the previous methodology and checks all the word-title-word sets.Document classification was utilizing by Naive Bayes approach which is one of the machine learning strategies [14]. A sum of 1150 archives were utilized for the training process and the n-gram used for feature extraction technique. Turkish documents were utilized during the preparation and testing stage and the results of the classification procedure was assessed by ascertaining the estimations of recall, precision,f-measure and accuracy. Then this measurements will be applied on creating titles for the test documents.

2.3.3 Term frequency (TF) and inverse document frequency (Idf) approach

TF is the recurrence of words happening in the document and IDF is logarithm of the complete number of documents partitioned by the quantity of documents containing this word. The archive words with highest TF [12].IDF were picked for the title word candidates.

2.3.4 K nearest neighbor approach

It scans the training dataset for the nearest related document and allot the training document title to the new document as title. In another study, [15] the KNN text classi-

fication is too unpleasant to even think about calculating content closeness, disregarding the relations inside the document and the connections between the reports.

2.3.5 Iterative Expectation-Maximization approach

It sees reports as written in a 'verbal' language and their titles as composed a 'concise' language. It constructs the interpretation model between the 'verbal' language and the 'brief' language from the documents and titles in the training corpus and 'make an interpretation of' each testing document into title[12].

2.4 Title Generation Techniques

2.4.1 Machine learning techniques

The machine learning approaches [12] is used to generate the title of broadcast news. They analyze the different techniques like KNN, Naïve Bayes etc. In [8] they analyze a different machine learning techniques to identify the tags of words. They use CRF, Tree tagger model, SVM to identify the tag using part of speech tagging. The tree tagger model is better than other models. The unigram, Bigram and other techniques was used for in [7] to tag the words but these models not tag the correct word. The Latent Dirichlet Allocation [16]is used to generate the title of songs based on their lyrics. They used estimation process to find out the probability of words.In [17] used LDA topic model form segment, utilizing the idea of the Bag Of-Word to show the word recurrence, so as to prepare the LDA topic model.The SVM topic classification model train area, the content will as highlight parameters and utilize the word recurrence table to assemble include vector, the topic classification model is prepared by feature vector.

Table 2.1: Summary of machine learning based techniques

Paper Ref.	**Technique**	**Problem statement**	**Results**	**Limitations**
[11]	Topic segmentation, title assignation, term weighting, similarity measures.	The videos on the news website have several latest news in one video so its time consuming to watch the full video if we only want to listen the particular news.A framework which associate the segments of the video to its related news article.	The video is converted into segments by topic segmentation algorithm and then they assign the segments of video to its related article by measuring the similarities.	when the title is not generated correctly the model cannot combine the topic segment with their related article.
[12]	Naive Bayes, Knn, Iterative Expectation Maximization, Term frequency, Inverse document frequency.	Title generation for broadcast videos is difficult because many error were generate when we convert video to text. In this paper they compare the results of different machine learning methods used for title generation.	The naive Bayes with using full vocabulary achieved 21.6% which is higher achieved accuracy as compare to other models.By observing the accuracy rate and limitations which method is good to generate the title.	By converting audio to text using speech recognition the generated text may have missing some information due to misses some chunks of audio.

Paper Ref.	Technique	Problem statement	Results	Limitations
[17]	Latent Dirichlet Allocation, SVM classifier	At this time, we can get the various of information and data on various ways, for example, science record, paper, blog, voice messages. It's significant issues that how to arrange this data and manage the shrouded data of the unstructured reports.	The LDA topic model form segment, utilizing the idea of the Bag Of-Word to show the word recurrence, so as to prepare the LDA topic model.The SVM topic classification model train area and utilize the word recurrence table.	The dialogue have numerous pointless words, these words consistently confound the subject order, how to locate the identical the word is significant.
[18]	Conditional random field, SVM, Tree Tagger system	Identify the same word by different tags is not easy for a machine.	The Amazigh corpus dataset is used. Three techniques is used on the same dataset to compare the accuracy of three techniques. The highest accuracy is achieved by tree tagger model.	The Model used in paper tag the words if you have the lexicon or manually tagged training corpus of that language .
[19]	Latent Dirichlet Allocation	A various number of electronic information, explicitly in the form of content report, result the high number of classification studies.	actualized an classification technique dependent on topic modeling idea.With the theme idea, an actual document is defined as a conveyance of points.	when the title is not generated correctly the model cannot the topic with their related article.

2.4.2 Natural Language Processing techniques

Natural Language Processing is an data extraction way to deal with consequently extract articraft from the textual depictions. In addition, NLP is frequently applied to produce the different component of concerns like fundamental terms, class models, experiments from the underlying Textual depictions. Be that as it may, it is generally needed to consider total section to extract applicable data from textual content that makes this cycle time consuming [20].Part of speech tagging is a main problem in natural language processing. Many techniques for use to identify the tag. Part of speech tagging is used to tag the words in the document which is used for further natural language processing. In [1] they use NLTK toolkit for automatic term extraction in the documents.Each client takes a reviews at the surveys before requesting anything on the web. In any case perusing those long reviews isn't simple for everybody. In this way, there must be something that can decrease the long reviews to short sentences of restricted words representing a similar importance. Text Summarization can come close by in this angle [21]. The language problems are handled by natural language processing [22] they developed a system which identify the words which is appropriate to send to other users or not. Many researchers used POS and natural language techniques to extract the text from the document, but the problem is that POS not tag the word correctly and then it cannot recognized the correct title.

Table 2.2: Summary of Natural Language Processing based techniques

Paper Ref.	**Technique**	**Problem statement**	**Results**	**Limitations**
[1]	Automatic Term Extraction, C programming Language, NLTK toolkit, tree tagger model	when writing any technical document by extracting words their synonym changes the meaning of the sentence.	In this paper they focus on all words not any terminology used to focus only nouns or verb etc.	The proper noun is tagged as common noun, this misclassification is still occur in this method.

Paper Ref.	Technique	Problem statement	Results	Limitations
[7]	Unigram, Bigram, Trigram, N-gram,Hidden Markov Model.	Many techniques works on part of speech tagging but many ambiguities still occur in Natural Language Processing.	Rule based approach, which cannot generate the correct tag for unknown words. The stochastic approach use different datasets the frequency and probability is not correct. The hybrid rule approach which model is better.	The drawback of this model is to tag the word correctly so this model requires more training then they classify the correct tag for the word.
[22]	Natural Language Processing.	The no mechanism is used to check the message is suitable to send to other end user.	If the message have these types of words which is not suitable to send to any user the application does not allow to send this message.	This application is not worked without internet. The message is not send when the user is offline.
[23]	POS Tagger, Link Grammar	The problem happen when the movement of verb does not exist before the preposition of location.	They use a standard POS tagging and LG parser to generate a set of these pattern which we extract from verb of movement to recognize a crucial pattern.	This model not recognize the location entity if preposition is not given in the document.

2.4.3 Title generation using Implicit and explicit Relations

In [24] they used Implicit and explicit relation among captions to generate the title of the videos. If you have a video with their subtitles or captions the morphological analyzer takings out verb or noun from the subtitles and then the stochastic matrix use TEXTRANK to set out the weights of the word which we analyze through kernel matrix.In another study, they consider the issue of video captioning, i.e creating one or various sentences to depict the substance of a video.[25] presents a hierarchical-RNN structure for video section captioning.The system models inter sentence reliance to create an arrangement of sentences of sentences given video information.

Table 2.3: Summary of Implicit and explicit based techniques

Paper Ref.	**Technique**	**Problem statement**	**Results**	**Limitations**
[24]	text rank, stochastic matrix, morphological analyzer	The title of the video are manually write. The content of the video is not same as the video title.	If you have a video with their subtitles or captions the morphological analyzer takings out verb or noun from the subtitles and then the stochastic matrix use TEXTRANK to set out the weights of the word which we analyze through kernel matrix.	The limitation of this model is they work when the video have closed caption or subtitles.

2.4.4 Title generation using Extracted keywords

It gathers word recurrence measurements on thing phrases that happen in the section, by utilizing on-line word references and records.They approached the records from the other course and attempted to produce titles by extracting keywords from the content measurably. The title generator created right now satisfactory titles for document that contain engaging, educational or substantive material with a rate of 70%[26].

2.4.5 Title generation using Neural network

The first study in 2018 which used recurrent neural network and LSTM for generating the title for Turkish language. In this paper they used stochastic approach which results is better than rule based approach. The term back propagation is used in the neural network learn earlier layers slowly because of error within the given weights, so to avoid this problem a recurrent neural network is used with long term short memory. The accuracy achieved by this model is 89%.RNN is the kind of neural network that includes cycles in memory. While the past yields are not utilized in future on standard sorts of neural systems, the past yields are utilized in cycles to get better outcomes [27]. However, evaporating and detonating gradient issues can't be handled appropriately on RNN. To have the option to manage this issue, LSTM offers new gates like forget and input gates [8]. By using these gates, slope flow control improved and conservancy of long-extend conditions are permitted. LSTM is a kind of RNN and qualified for learning long term conditions. The algorithm is specifically designed to maintain a strategic distance from the issue of RNN about long term dependencies. The RNN is fit for holding the old yield yet as the holes between grows. The RNN gets incapable to learn old yield and associate the yield to the necessary info. This issue is solved with LSTM.In another study, Yi Bin et al. [28] acquainted a grouping with succession way to deal with depict video cuts with natural language.The focal point of their method was applied two LSTM networks for the visual encoder and natural language generator fragment of model. Specifically, they encoded video arrangements with a bidirectional Long-Short Term Memory (BiLSTM) network, which could viably catch the bidirectional worldwide transient structure in video.

Table 2.4: Summary of Neural network based techniques

Paper Ref.	**Technique**	**Problem statement**	**Results**	**Limitations**
[8]	Recurrent Neural Network,Long-Short Term Memory, Neural network.	Traditional model used rule based approach for part of speech tagging which consumes more time and more expensive.	In this paper they used stochastic approach which results is better than rule based approach. A recurrent neural network is used with long term short memory. The accuracy achieved by this model is 89%.	When we increase the data both models performed well but on short dataset the LSTM is better than RNN.
[29]	Recurrent Neural Network,Long-Short Term Memory	On the planet brimming with sarcastic individuals, it's getting trying for the individuals of 21st century to distinguish sarcasm utilizing sentiment investigation productively.	sarcasm identification causes us to comprehend the severe truth under the glossed over sentences. It is generally utilized in different systems administration destinations for understanding the genuine issue.	A decent information makes a decent model, and till date, the datasets have their own constraints and predispositions.

CHAPTER 3

METHODOLOGY

This chapter presents the detail of proposed strategy carried out for title generation of Talk shows. We employed Latent Dirichlet Allocation and word2vec for classification of title. We employed LSTM for generation of title.We first present the detail of our custom developed dataset. Then we discussed each section of proposed methodology in detail i.e information preprocessing, include feature extraction,classification and generation.

3.1 Dataset description

3.1.1 Traning dataset

The training dataset is downloaded from kaggle.All the news have three CSV files and each file contains the 50,000 articles. All the news dataset which is used for for training contains 100,000 articles.For every distribution, they utilized archive.org to get the past 18 months of either home page important news or RSS channels and ran those connections through the scraper.That is, the articles are not the result of scratching a whole site, yet rather their all the more noticeably placed articles. The distributions incorpo-

	Unnamed: 0	id	title	publication	author	date	year	month	content
0	0	17283	House Republicans Fret About Winning Their Hea...	New York Times	Carl Hulse	2016-12-31	2016.0	12.0	WASHINGTON — Congressional Republicans have...
1	1	17284	Rift Between Officers and Residents as Killing...	New York Times	Benjamin Mueller and Al Baker	2017-06-19	2017.0	6.0	After the bullet shells get counted, the blood...
2	2	17285	Tyrus Wong, 'Bambi' Artist Thwarted by Racial ...	New York Times	Margalit Fox	2017-01-06	2017.0	1.0	When Walt Disney's "Bambi" opened in 1942, cri...
3	3	17286	Among Deaths in 2016, a Heavy Toll in Pop Musi...	New York Times	William McDonald	2017-04-10	2017.0	4.0	Death may be the great equalizer, but it isn't...
4	4	17287	Kim Jong-un Says North Korea Is Preparing to T...	New York Times	Choe Sang-Hun	2017-01-02	2017.0	1.0	SEOUL, South Korea — North Korea's leader, ...

Figure 3.1: A sample Image from All the news dataset

rate the Reuters, New York Times, Breitbart, Business Insider, the Guardian,Buzzfeed News, the Atlantic,New York Post, Fox News, NPR, Talking Points Memo, National Review,CNN,Vox, and the Washington Post. Inspecting wasn't exactly logical; I picked distributions dependent on my recognition of the space and attempted to get a scope of political arrangements, just as a blend of print and computerized productions. By check, the distributions separate in like manner: The information fundamentally falls between

the long periods of 2016 and July 2017, in spite of the fact that there is a no irrelevant number of articles.

Table 3.1: Statistics of All the news Dataset

Detail	**statistics**
Total no of articles	143,000
Total no of files	3
1 file contain articles	50,000
Total no of columns in 1 article	10
Column used in this research study	2

3.1.2 Testing dataset

To analyze the proposed methodology, two distinctive datasets are used.The motivation to utilize two datasets is that each dataset has its own challenges which assist us with analyzing the proposed methodology better.First we use All the news dataset [30] which is used to train the dataset and 70% dataset is used to train the dataset and 30% dataset is used to test the proposed scheme.Then we build custom dataset to test the proposed methodology.The custom dataset contains the text of 50 videos from BBC channel.We extract the text from the season of the The lead with Jake Tapper.The season of this programs contain manually generated title.With the proposed methodolgy we generate the title of these videos.

Table 3.2: Statistics of custom Dataset

Detail	**statistics**
Total no of articles	50
Total no of columns	2

3.2 Proposed Methodology

Today we have a lot of text data in internet and it's difficult to read and recommend that data is related to that which is actually we want to search.

3.2.1 Data Preprocessing

The data set contains 3 files which contains the 143,000 articles from 15 American publications.In this work we use the first file which contains the 50,000 news articles.The first step of preprocessing to clean text of websites, email addresess and any punctuation

and convert text into lower case. We use the function that removes all stopwords from the text.The last step is to stem words, so plural and singular are treated the same.Due to machine which we use for this work takes a lot of time to perform the first step so we use first 6000 columns clean and then use this columns for further processing.The dataset we used in this research contains 3 articles file and each article file contains 10 columns.we use the data of content column and title column because others columns are not useful for this research work.The next step is to tokenized the words and the time utilize our machine to clean and tokenize 6000 articles is 6.950782787799835 min.

3.2.2 Feature Extraction

Text extraction by LDA requires an intial procedure to decide the intial topic for each document. In the LDA, this is done by utilizing a multinonimal random function. The distinction among LDA and other topic based strategies is LDA removing text units into three levels. The most important level is extracting the significance of every content unit in each document by computing its probability and characterizing new topic for the content unit. This procedure is done consistently until it comes to a decided threshold value.The subsequent level is deciding topic distribution of each document dependent on the probability and topic of content unit.The outcome determines that each content have several topics with their particular probabilities. The external level determines the distribution of topics for the corpus (set of archives) dependent on content distribution topic. Consequently, it tends to be inferred that LDA isn't just valuable for separating the importance of the content unit and yet having the option to make soft clustering for reports dependent on topic.

3.2.3 Overview of LDA

The topic modeling helps us what is actually in text corpus. The Latent Dirichlet Allocation (LDA) organize a bag of words into different topics. The LDA works that which topic belongs to which word. Latent dirichlet allocation(LDA) is a probabilistic model and commonly used for topic modeling. The probabilistic model gives us the probability of per document topic distribution and topic word distribution. In a text corpus the model finds out the most significant words and theses words is used as a topic and the topic number is defined by a user.[17] A various number of electronic information, explicitly in the form of content report, result the high number of classification studies.In [19] actualized an classification technique dependent on topic modeling idea. With the theme idea, an actual document is defined as a conveyance of points and a topic is defined by a lot of words.LDA is a statistical probability model, it create a arbitrary text which establish by multi words. Through LDA model we can get the content distribution probability $P(w|z)$ in the theme, and the topic distribution probability $P(z|d)$ in the content[31]. We will clarify the LDA model by figure.

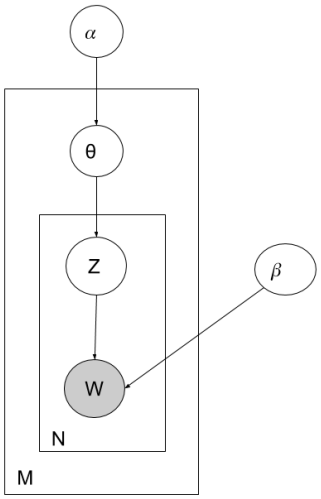

Figure 3.2: The architecture of Latent Dirichlet Allocation Model[31]

3.2.4 Overview of Word2vec

Another methodology, word embedding calculation named ”Word2Vec”, works by computing the word loads, semantic relationship, and the last weights of vectors [32].The word2vec model accepts a data as information and produces the word vectors as output. It first develops a vocabulary from the preparation text information and afterward generates vector representation of words .The achievement in the content handling, examination and characterization has been essentially upgraded by utilizing deep learning. This achievement is contributed by the nature of the word representations. Glove, FastText, Word2Vec and TFIDF are utilized for the word representation. They intended to improve word representation Word2Vec boundaries. The achievement of the representation was estimated by utilizing a profound learning order model [33]. The subsequent word vector record can be utilized as features in many natural language processing and AI applications. A basic method to research the educated representations is to locate the nearest words for a user determined word.The reason and value of Word2vec is to bunch the vectors of similar words in vectorspace. That is, it identifies similarities numerically. Word2vec makes vectors that are appropriated mathematical representations of word features for example the setting of individual words. It does as such without human intercession.

3.2.5 Workflow of Title generation using lda2vec

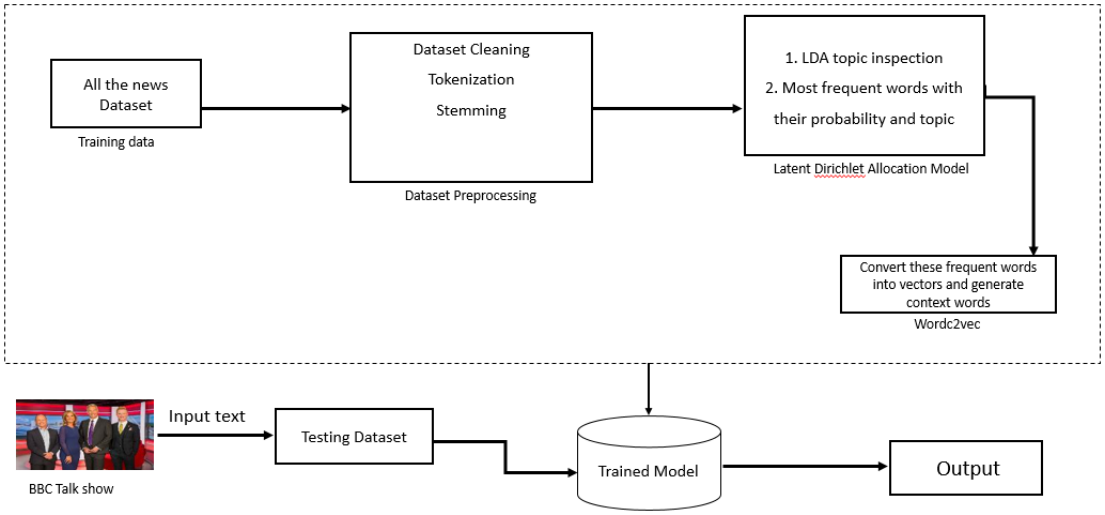

Figure 3.3: Workflow of Title generation using lda2vec

3.2.6 Overview of RNN

With the ascent of DL in the previous decade, RNNs have become functional and integral assets for large scope supervised learning of successions. This advancement has gotten generally obvious in Natural Language Processing where they have set a few new benchmarks beating methods that have been viewed as cutting edge for quite a while [34].Neural Networks are set of algorithms which intently look like the human mind and are intended to perceive patterns. They decipher sensory information through a machine discernment, naming or clustering raw data. They can perceive numerical examples, contained in vectors, into which all real world information (pictures, sound, text or time arrangement), must be deciphered. Artificial neural systems are made out of an enormous number of exceptionally interconnected preparing components (neuron) cooperating to tackle an issue. Recurrent Neural Network is a hypothesis of feedforward neural framework that has an internal memory. RNN is irregular in nature as it plays out a comparative limit with regards to every commitment of data while the yield of the current data depends upon the past one calculation. After handling the outcome, it is copied and sent go into the repetitive framework.For settling on a choice, it considers the current information and the output it has gained from the past information. Not in the slightest degree like feedforward neural frameworks, RNNs can use their inward state (memory) to handle groupings of sources of data. This makes them applicable to undertaking for instance, unsegmented and discourse acknowledgment. In other neural systems, all the information sources are autonomous of one another. In any case, in RNN, all the sources of info are identified with one another.RNN can demonstrate succession of information with the goal that each example can be thought to be dependent to past ones Recurrent neural framework are even used with convolutional layers to expand the successful pixel neighborhood.

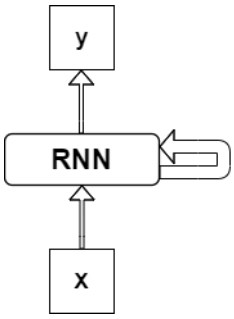

Figure 3.4: Simple architecture of RNN[8]

3.2.7 Overview of LSTM

Individuals don't start their deduction without any planning reliably. As you read this paper, you see each word subject to your understanding of past words. You don't dispose of everything and start thinking without any planning again. Your considerations have diligence. Traditional neural frameworks can't do this, and it has all the earmarks of being a huge shortcoming. For instance, imagine you need to group what sort of occasion is going on at each point in a film.Utilization of Recurrent Neural Network, LongShort Term Memory (LSTM) and Word Embeddings can make the sarcasm discovery productive and accordingly making the statements from twitter effectively classifiable [29]. It's hazy how a traditional neural system could utilize its thinking about past occasions in the film to advise later ones. Recurrent neural systems address this issue. They are systems with circles in them, permitting data to persevere.Long Short-Term Memory (LSTM) systems are a changed rendition of recurrent neural network, which makes it simpler to recollect past information in memory. The vanishing gradient issue of RNN is settled here. LSTM is appropriate to characterize, process and anticipate time arrangement given delays of unknown duration.It prepares the model by utilizing back-propagation.In another study S.Chakraborty et al. [35] discussed that the increment of LSTM cells the model beginnings performing better. This happens in light of the fact that with the expansion in the quantity of cells the model beginnings fitting better. At that point with additional expansion of LSTM cells, the model begins once again fitting as the performance begins deteriorating. The advantages of lstm are following:

1.The consistent error backpropagation inside memory cells brings about LSTM's capacity to connect long delays if there should be an occurrence of issues like those talked about above.

2.For long delay issues LSTM can deal with noise and persistent qualities. Rather than finite state automata or Hidden Markov models LSTM doesn't need a from the earlier decision of a finite number of states.

3.LSTM functions admirably over a wide range of boundaries, for example, learning rate, output gate bias and input gate bias.

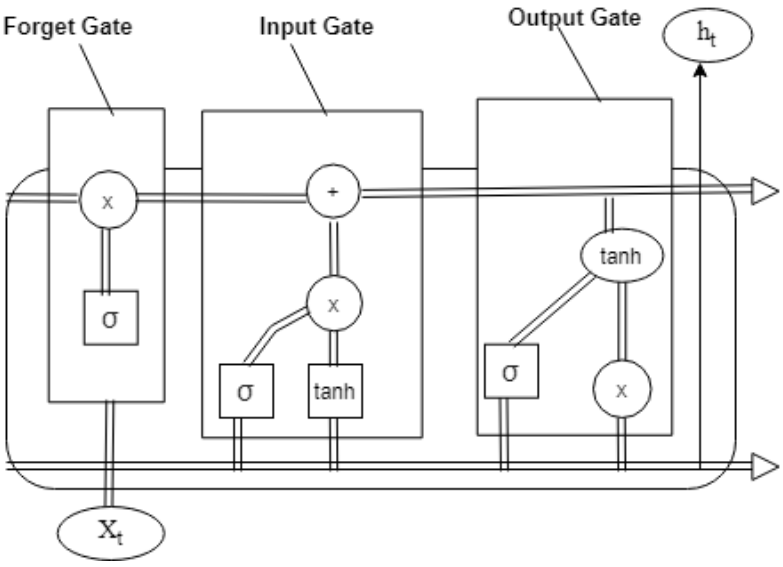

Figure 3.5: Architecture of LSTM[8]

3.2.8 Workflow of Title generation using LSTM

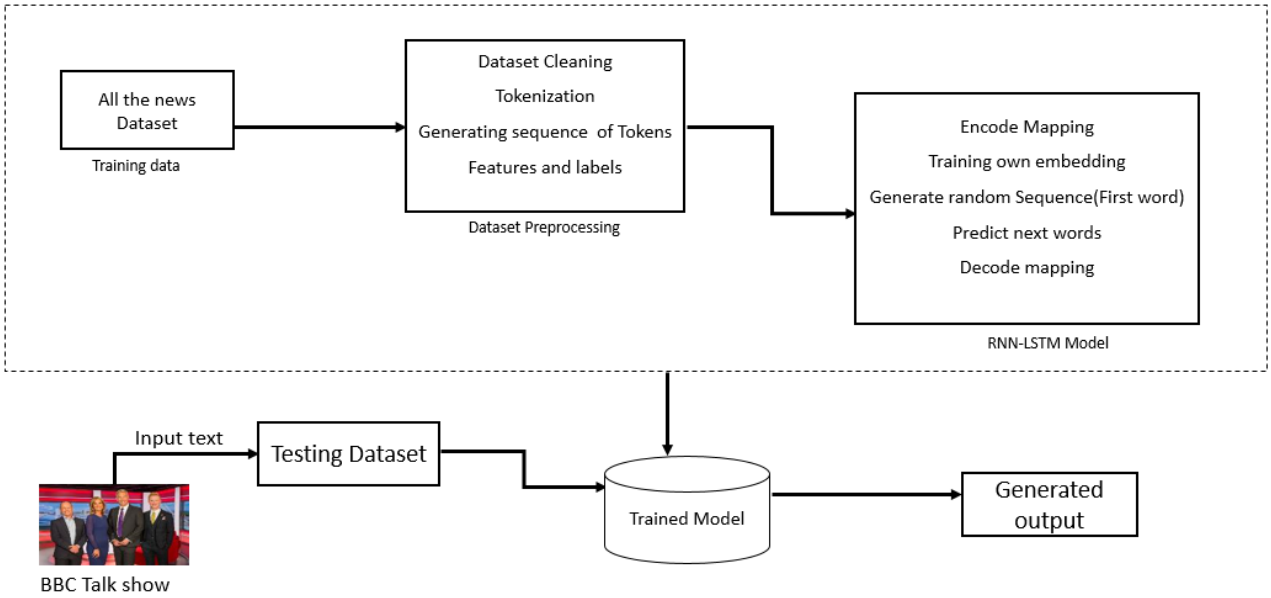

Figure 3.6: Workflow of Title generation using LSTM

CHAPTER 4

ANALYSIS & RESULTS

This section presents the details of experiments assess to evaluate the effectiveness of the proposed methodology. We utilized all the news dataset as well as custom dataset to investigate this study. We first present the experimental protocol and results. Later, we present different experimental scenarios used for title generation. Finally, we summarize a few of the recent studies of title generation for comparison purpose.

4.1 Experimental Protocol and Results

To assess the viability of title generation we utilized two distinctive dataset for example publicly available dataset and our custom developed dataset. As clarified before, the All the news dataset contains 50,000 news article however there are only 6000 articles used for one experiment.We acquired these content lines and apply different preprocessing steps (as of now chapter in section 3).we used two methodologies for two datasets.first we employed LDA and we use All the news dataset and for this experiment we used first article which contains 10,000 articles and for training we use 6000 columns due to reduce processing time of machine.The second methodology we used RNN(LSTM) we used second article which contains 10,000 articles.The All the news datset contain 10 colums but for this experiment we use only two columns title and content.The custom dataset which contains the text of 50 videos.In testing process we generate the title of one video.

Table 4.1: All the news Dataset division

Training set	6000 articles
Testing set	2000 articles

4.1.1 Title Generation using LDA

The first step is preprocessing of data.For training we use the first article from all the news dataset which contain 10,000 articles.Due to machine restrictions we only use first 6000 articles.The total time to clean and tokenize the 6000 articles is 6.950782787799835 min.

Word Frequency And Distribution

Using LDA, After preprocessing step the next step is used to choose the frequent words in the document.First we get list of all words in the document.By using use nltk fdist to get a frequency distribution of all words. The length of list in the document is 6000 and the average document length is 695.1946666666666.The minimum document length is 4 and the maximum document length 14384. The quantity of documents for every topic

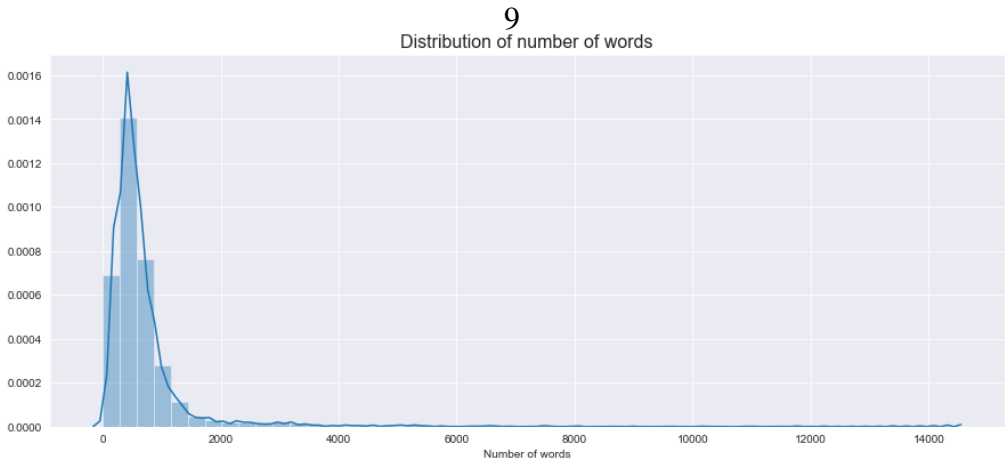

Figure 4.1: Word Frequency And Distribution

by allotting the document to the theme that has the most weight in that document.After cleaning and excluding short aticles, the dataframe now has 5932 articles.

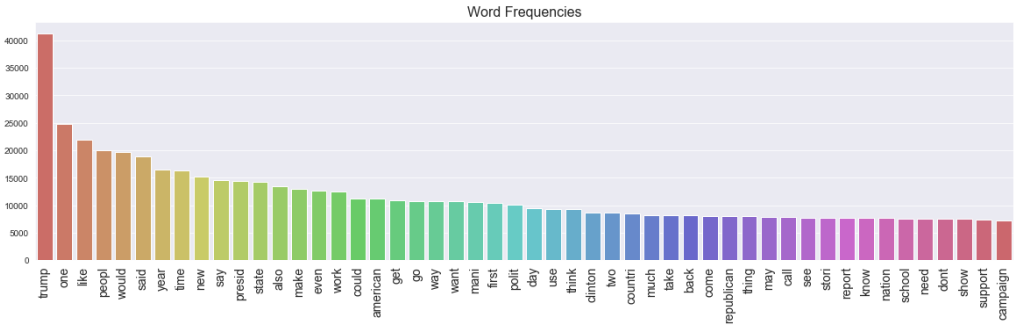

Figure 4.2: Representation of most discussed topic in the Document

pyLDAvis

pyLDAvis is intended to assist clients with interpretting the topics in a topic model that has been fit to a corpus of text data. The bundle separates data from a fitted LDA subject model to educate an interactive web-based visualization.Our representation (outlined in Figure 4.3) has two essential pieces. To begin with, the left board of our visualization presents a worldwide perspective of the topic model.furthermore, addresses 2 and 3. In this view, we plot the subjects as circles in the two-dimensional plane whose centers are dictated by calculating the distance among themes, and afterward by utilizing multidimensional scaling to extend the intertopic distances onto two measurements We encode

every point's generally speaking predominance utilizing the area of the circles, where we sort the points in diminishing order of prevalence.Second, the right board of our visualization portrays a horizontal barchart whose bars determines the individual terms that are the most helpful for deciphering the right now chosen topic on the left and permits clients to respond to address 1, "What is the significance of every topic?". A couple of overlaid bars depicts both the corpus-wide recurrence of a given term just as the topic specific recurrence of the term. The left and right boards of our visualization are connected with the end goal that choosing a subject (on the left) uncovers the most helpful terms (on the right) for deciphering the chosen topic. Also, choosing a term (on the right) uncovers the restrictive conveyance over topics (on the left) for the chosen term. This sort of connected choice permits clients to analyze an enormous number of topic term connections in a minimal way.

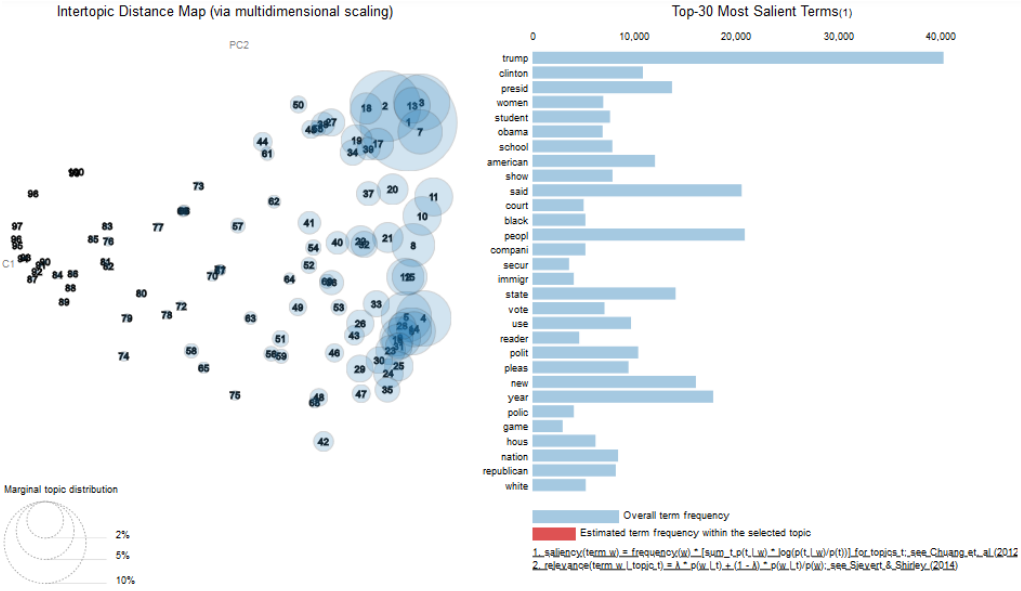

Figure 4.3: Python library for interactive topic model visualization in the Document

Word2Vec Model And Training

In extremely oversimplified terms, Word Embeddings are the texts changed over into numbers and there might be distinctive mathematical representation of a similar text. Word2Vec is a prescient word embedding procedure which changes over a word into a vector of numbers dependent on the setting of the target word. Vectors out of the words are created utilizing the encompassing words to represent the target words. Transform a group of sentence(containing multiple lists of words) into a feature vector. It averages out all the word vectors of the group of sentence.All frequent topics and topics per words convert into vectors and these vectors help us to find the similarity between the generated title by this methodology and actual title.The most frequent words in the document which we used in the next process we changed these topics into vectors by using word2vec.These vectors are helped us to choose the most similar words of their related topic in the document.The following figure shows us the vectors of most frequent words in the document.By using

these words we calculate the similarity of actual title and generated title.

```
(array([-3.9551e-01,  5.4660e-01,  5.0315e-01, -6.3682e-01, -4.5470e-01,
         3.0889e-01, -4.9240e-02,  2.7191e-01,  3.1562e-01, -3.2879e-01,
         2.5089e-01,  1.4508e-01,  3.5136e-01, -2.2793e-01, -1.5894e-01,
        -5.1527e-01, -2.7978e-01,  3.6470e-01, -3.9425e-01,  3.3299e-01,
         4.3051e-01,  1.8300e-01,  2.5095e-01, -1.8547e-01,  3.4698e-01,
         5.5137e-02, -4.5979e-01, -8.2963e-01, -1.8523e-02, -3.6772e-01,
         4.5566e-02,  7.1052e-01, -2.2782e-02, -8.0889e-02,  2.0685e-01,
         4.9855e-01, -5.9794e-02, -8.0048e-03, -2.3823e-01, -3.3759e-01,
        -2.4201e-01, -2.3788e-01, -1.1362e-03, -4.0395e-01, -4.4859e-01,
        -3.2189e-01,  4.8405e-01, -2.7999e-02,  1.0148e-01, -9.3585e-01,
        -8.7522e-02, -3.9959e-01,  3.6545e-01,  1.3726e+00, -3.0713e-01,
        -2.5940e+00,  2.2431e-01, -4.1168e-02,  1.7765e+00,  4.0010e-01,
        -1.0996e-01,  1.4178e+00, -2.6154e-01,  1.8617e-01,  7.9328e-01,
        -1.1709e-01,  8.7541e-01,  4.3911e-01,  3.4711e-01, -2.8515e-01,
         7.6269e-02, -6.3038e-01,  1.6408e-01, -3.7053e-01,  5.8485e-01,
        -1.5472e-01, -2.6382e-01, -1.8590e-01, -7.5228e-01, -1.5752e-01,
         7.8539e-01, -1.8846e-02, -8.0130e-01,  1.5561e-01, -1.8624e+00,
        -1.6969e-01,  1.9419e-01, -3.0683e-01, -7.8067e-01, -4.9689e-01,
        -1.8256e-01, -4.2016e-02, -2.6290e-01,  5.8531e-02, -4.4664e-01,
        -9.9765e-02, -4.3050e-01, -2.3693e-01, -1.4519e-02,  3.1981e-01]),
```

Figure 4.4: word vectors of frequent topics in the document

Word mover distance

Sentence similarity estimates assume an significant job in text-related exploration and applications in those areas, for example, text mining, Web page recovery and dialogue frameworks. Sentence similarity helps in recognizing the redundant information by estimating the occurrences of the comparable words in the content.

To do as such, it is commonly determined by first embedding the sentences and afterward taking the similarity between them. Sentence similarity is additionally utilized in text characterization and text summarization. How about we take a straightforward model, we need to basic sentences: I travel to my office utilizing my vehicle and: I travel to my office utilizing a taxi. Here the meaning of the sentence is the equivalent for example 'I travel to my office' however decision of vehicle contrasts. Here the sentence similarity needs to decide how 'close' two bits of text are both in surface closeness and significance.By using the word mover distance we find the distance between the words of actual title and the generated title by this method.The generated title contains those words in the title which distance is small.The recovery of records has entered a practically equivalent to arrangement, where each word is related with an exceptionally informative feature vector. Word Mover's distance is the first to make the association between top word embeddings and EMD retrieval methods. The WMD has a few intriguing properties:

1. It is hyper-boundary free and straight-forward to comprehend and utilize;

2. It is exceptionally interpretable as the distance between two archives can be separated and clarified as the meager separations between few individual words

3. It normally joins the information encoded in the word2vec/Glove space and prompts high retrieval accuracy. The accuracy rate of generated title using lda2vec is shown in fig: 4.5.The X-axis shows the title no and the y axis shows the accuracy rate.

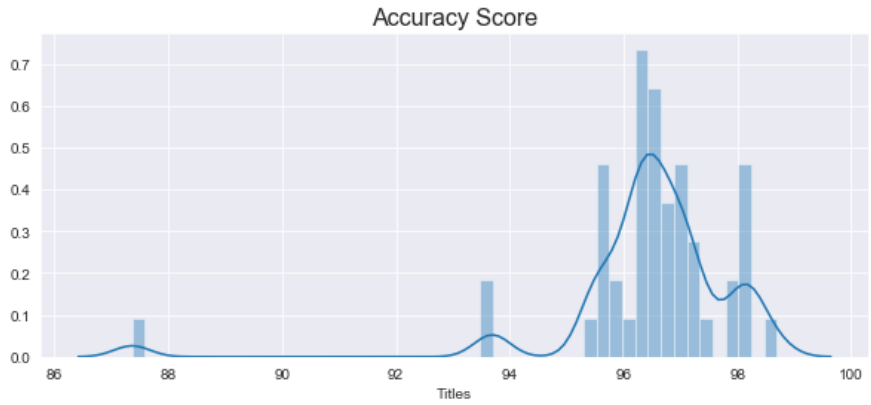

Figure 4.5: Accuracy rate of Generated title using lda2vec

4.1.2 Title generation using LSTM

Recurrent neural systems can likewise be utilized as generative models. This implies that it is being utilized for predictive models they can become familiar with the arrangements of an issue and afterward produce completely new conceivable groupings for the problem area. Generative models like this are valuable not exclusively to concentrate how well a model has taken in an issue, however to become familiar with the problem space itself. first we use All the news dataset for training the model.The article 1

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, None, 100)         5381200
_________________________________________________________________
lstm_1 (LSTM)                (None, 64)                42240
_________________________________________________________________
dense_2 (Dense)              (None, 128)               8320
_________________________________________________________________
dropout_1 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_3 (Dense)              (None, 53812)             6941748
=================================================================
Total params: 12,373,508
Trainable params: 12,373,508
Non-trainable params: 0
```

Figure 4.6: No of parameter used for training LSTM

from all the news dataset is used to train the model. For training this model we use the first articles which contains the 50,000 records.The total no of parameter used in the training is 12,373,508.The trainable parameter used in training is 12,373,508.After training the dataset we used the second articles from all the news dataset which contains the 50,000 articles and we randomly pick the aticle and generate the title. First we preprocess the data then train our model.The accuracy rate id better than traditional methodology if the number of epochs is increased ant trained the model more. The accuracy rate of generated title using Lstm is shown in fig: 4.7.The X-axis shows the title no and the y axis shows

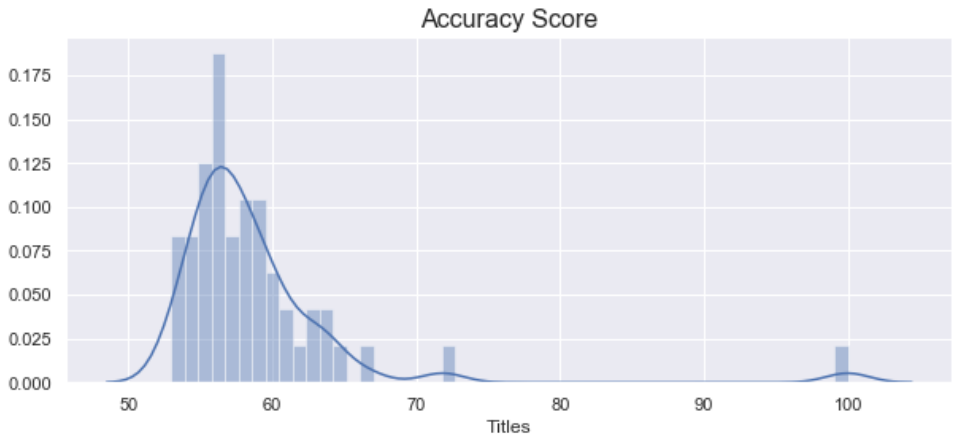

Figure 4.7: Accuracy rate of Title Generation using LSTM

the accuracy rate.

4.2 Performance Analysis and Discussion

We performed two experiments using two datasets to highlight the effectiveness of proposed methodology.we use All the news dataset and employed LDA and using the same dataset we employed the RNN-LSTM. Then using the custom dataset we analyze the accuracy and other parameters which one method is better for title generation.

Scenario-I: Title Generation of talk shows using LDA+word2vec on all the news dataset as well as custom dataset.

Table 4.2: Accuracy rate of title generation using LDA+word2vec

Dataset	**Recognition rate**
All the news datset	85% to 95%
custom dataset	80% to 93%

Scenario-II: Title Generation of talk shows using LSTM on all the news dataset as well as custom dataset.

Table 4.3: Accuracy rate of title generation using LSTM

Dataset	**Recognition rate**
All the news datset	70% to 80%
custom dataset	65% to 75%

4.2.1 Generated Titles with LDA+word2vec

	Actual Titles	Genrated Titles
0	What Happens If You Stick Your Head in a Parti...	What Betsy DeVos Did (and Didn't) Reveal About...
1	Resistance to the Antibiotic of Last Resort Is...	Donald Trump's Conflicts of Interest: A Crib S...
2	The Transnational Trolley and Doughnuts in Jua...	Soon, Bill Cosby Will Have His Day in Court
3	Black-ish's 'Lemons' Is Art for the Age of Trump	Donald Trump's Conflicts of Interest: A Crib S...
4	Trump, Remember You Are a Man	The Killing of a Western Hostage in the Philip...
5	Toni Erdmann Is a Comedy Experience Unlike Any...	Trump's Attack on a Judge For Staying His Trav...
6	How Cash Bail Keeps the Poor in Jail	Recognizing America's Nuclear Past in Japan
7	The Shadow Network of Anti-Vax Doctors	Libya, ISIS, and the Flow of Foreign Fighters
8	The Trump Promise Tracker	The Big Question: Reader Poll
9	Obama and the Limits of 'Fact-Based' Foreign P...	An Indictment in the Walter Scott Shooting

Figure 4.8: Title Generated using LDA+word2vec

4.2.2 Generated Titles with LSTM

	Actual Titles	Genrated Titles
0	what is marine le pen doing at trump tower?	not sought a meeting with trump "we did not
1	on the tonight show, michelle obama cements a ...	in the role it was also an endorsement not
2	could cancer drugs treat autism?	," sahin says there are no drugs approved to
3	resistance to the antibiotic of last resort is...	gut microbiota of healthy humans whenever an...
4	how to train a worker	it might fall from the sky , like a spaceship
5	the revolt of working parents	in the workplace was illegal gender discrimi...
6	the perfectly normal ways trump can enrich him...	our political system responds to the prefere...
7	the atlantic daily: insight and oversight	subjects have maintained their fondness for ...
8	the atlantic's week in culture	2017 golden globes — the editors discuss all...
9	ranked: the world's most unusual places to hol...	has been leading a rebellion against the wea...

Figure 4.9: Generated Title using RNN-LSTM

CHAPTER 5

CONCLUSION & FUTURE WORK

Titles are effective approaches to connect peoples for videos, it is a much significance factor for different services. In the proposed strategy, the title of a video is generated with the help of most important and repeated word used in the content of the video. The significance of a sentence is resolved with word weights. This reasearch study explores the chance of utilizing LDA for building up a title by analyzing the internal significance of the content of video. Applying LDA gives good results on the dataset which is defined for this work.In numerous automatic ordering or abstracting tasks, the keyword that demonstrate the content substance are gotten from the title, realizing that the title would incorporate the most significant topics covered in the data. In this investigation, we moved toward the document from the other course and attempted to produce titles by separating keywords from the content measurably.By using deep learning technique(LSTM) we generated a title on custom dataset and All the news datasets which achieve good results as compare to traditional techniques. To make a title for a document is a complicated task. To create a title for a document turns out to be considerably additionally challenging since we need to manage word mistakes produced by speech recognition. Generally, the title generation task is firmly associated with traditional summarization since it very well may be thought of amazingly short summarization. Customary summarization has stressed the extractive methodology, utilizing chosen sentences or passages from the document to give an outline. The shortcomings of this methodology are powerlessness of exploiting the training corpus and producing summarization with little proportion. In this way, it won't be reasonable for title generation problems. The KNN approach functions admirably for title generation particularly when cover in content between training dataset and test dataset is huge.A human created keyphrase was viewed as equivalent to a machine-produced keyphrase in the event that they had a similar sequences of stems.

5.1 Conclusion

Today, with the expanding utilization of the Internet, numerous archives have been moved to the web and opened to people in general. In any case, the way that these archives are not typically arranged significantly limits the capacity of clients to access and quest for data.The mostly talk shows content which we seen on internet does not match the title of the video. In this study we employed LSTM. A total no of 6000 records were utilized for the training procedure because the limit of machine.In principle, this limitation can be dodged by utilizing the full slope (maybe with extra customary hidden units getting

contribution from the memory cells). However, we don't suggest figuring the full gradient for the accompanying reason. It increments computational unpredictability.Each memory cell square needs two extra units (input and output gate). The proposed approach has been tried on a real dataset and accomplished a performance about 70% to 80% using LSTM. At the point when the outcomes acquired from the test strategies are analyzed, it has been seen that the methodology proposed inside the extent of the examination is profoundly productive and arranges the archives with great accuracy.By comparison the traditional approach LDA achieve good results in short dataset.if the dataset is long the traditional approach is not giving the appropriate results.The Lstm model acheives good results as compare to LDA+word2vec if we increase the number of epochs to train the model.

5.2 Perspective

This investigation essentially centered around the acknowledgment of English based content, the proposed method can likewise be executed for other language like Arabic, Pashto, and Persian etc.In our further investigation on this issue we generate the title of live stream videos.The proposed model of LSTM improves with more number of epochs.

REFERENCES

[1] N. I. Simon, V. Kešelj, Automatic term extraction in technical domain using part-of-speech and common-word features, in: Proceedings of the ACM Symposium on Document Engineering 2018, 2018, pp. 1–4.

[2] L. Xie, Y. Shi, Z. Li, Driving pattern recognition based on improved lda model, in: 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), 2018, pp. 320–324.

[3] Z. A. Guven, B. Diri, T. Cakaloglu, Classification of new titles by two stage latent dirichlet allocation, in: 2018 Innovations in Intelligent Systems and Applications Conference (ASYU), 2018, pp. 1–5.

[4] D. Ninomiya, M. Mozgovoy, Improving pos tagging for ungrammatical phrases, in: Proceedings of the 2012 Joint International Conference on Human-Centered Computer Environments, HCCE '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 28–31. doi:10.1145/2160749.2160756.
URL https://doi.org/10.1145/2160749.2160756

[5] C. Yi, An english pos tagging approach based on maximum entropy, in: 2015 International Conference on Intelligent Transportation, Big Data and Smart City, 2015, pp. 81–84.

[6] C. Xia, T. He, W. Li, Z. Qin, Z. Zou, Similarity analysis of law documents based on word2vec, in: 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C), 2019, pp. 354–357.

[7] S. G. Kanakaraddi, S. S. Nandyal, Survey on parts of speech tagger techniques, in: 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), IEEE, 2018, pp. 1–6.

[8] C. A. Bahcevan, E. Kutlu, T. Yildiz, Deep neural network architecture for part-of-speech tagging for turkish language, in: 2018 3rd International Conference on Computer Science and Engineering (UBMK), IEEE, 2018, pp. 235–238.

[9] A. R. Mishra, V. K. Panchal, P. Kumar, Extractive text summarization - an effective approach to extract information from text, in: 2019 International Conference on contemporary Computing and Informatics (IC3I), 2019, pp. 252–255.

[10] S. Modi, R. Oza, Review on abstractive text summarization techniques (atst) for single and multi documents, in: 2018 International Conference on Computing, Power and Communication Technologies (GUCON), 2018, pp. 1173–1176.

[11] A. Bouchekif, G. Damnati, D. Charlet, N. Camelin, Y. Esteve, assignment for automatic topic segments in tv broadcast news, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 6100–6104.

[12] R. Jin, A. G. Hauptmann, Automatic title generation for spoken broadcast news, in: Proceedings of the first international conference on Human language technology research, Association for Computational Linguistics, 2001, pp. 1–3.

[13] L. Jain, P. Agrawal, Sheershak: an automatic title generation tool for hindi short stories, in: 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 579–584.

[14] M. BAYGIN, Classification of text documents based on naive bayes using n-gram features, in: 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018, pp. 1–5.

[15] H. Li, H. Jiang, D. Wang, B. Han, An improved knn algorithm for text classification, in: 2018 Eighth International Conference on Instrumentation Measurement, Computer, Communication and Control (IMCCC), 2018, pp. 1081–1085.

[16] R. Hossain, M. R. K. R. Sarker, M. Mimo, A. Al Marouf, B. Pandey, Recommendation approach of english songs title based on latent dirichlet allocation applied on lyrics, in: 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), IEEE, 2019, pp. 1–4.

[17] J. Yeh, C. Lee, Y. Tan, L. Yu, Topic model allocation of conversational dialogue records by latent dirichlet allocation, in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, 2014, pp. 1–4.

[18] S. Amri, L. Zenkouar, R. Benkhouya, A comparative study on the efficiency of pos tagging techniques on amazigh corpus, in: Proceedings of the 2nd International Conference on Networking, Information Systems & Security, 2019, pp. 1–5.

[19] R. Kusumaningrum, M. I. A. Wiedjayanto, S. Adhy, Suryono, Classification of indonesian news articles based on latent dirichlet allocation, in: 2016 International Conference on Data and Software Engineering (ICoDSE), 2016, pp. 1–5.

[20] I. Ahsan, M. A. Ahmed, S. Rehman, M. Abbas, M. A. Khan, A novel nlp application to automatically generate text extraction concepts from textual descriptions, in: Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence, ICCAI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 55–58. doi:10.1145/3330482.3330506.
URL https://doi.org/10.1145/3330482.3330506

[21] R. Boorugu, G. Ramesh, A survey on nlp based text summarization for summarizing product reviews, in: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 352–356.

[22] S. Karthick, R. J. Victor, S. Manikandan, B. Goswami, Professional chat application based on natural language processing, in: 2018 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC), IEEE, 2018, pp. 1–4.

[23] Y. Sari, M. F. Hassan, N. Zamin, Creating extraction pattern by combining part of speech tagger and grammatical parser, in: 2009 International Conference on Computer Technology and Development, Vol. 1, 2009, pp. 515–519. doi:10.1109/ICCTD.2009.227.

[24] J. Son, W. Park, S. Lee, S. Kim, Video scene title generation based on explicit and implicit relations among caption words, in: 2018 20th International Conference on Advanced Communication Technology (ICACT), 2018, pp. 571–573.

[25] H. Yu, J. Wang, Z. Huang, Y. Yang, W. Xu, Video paragraph captioning using hierarchical recurrent neural networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4584–4593.

[26] D. Gokcay, E. Gokcay, Generating titles for paragraphs using statistically extracted keywords and phrases, in: 1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century, Vol. 4, 1995, pp. 3174–3179 vol.4.

[27] E. Balouji, I. Y. H. Gu, M. H. J. Bollen, A. Bagheri, M. Nazari, A lstm-based deep learning method with application to voltage dip classification, in: 2018 18th International Conference on Harmonics and Quality of Power (ICHQP), 2018, pp. 1–5.

[28] Y. Bin, Y. Yang, F. Shen, X. Xu, H. T. Shen, Bidirectional long-short term memory for video description, MM '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 436–440. doi:10.1145/2964284.2967258.
URL https://doi.org/10.1145/2964284.2967258

[29] S. S. Salim, A. Nidhi Ghanshyam, D. M. Ashok, D. Burhanuddin Mazahir, B. S. Thakare, Deep lstm-rnn with word embedding for sarcasm detection on twitter, in: 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1–4.

[30] T. A., All the news dataset, [Internet]. Kaggle.com. 2017, cited 12 April 2020] (2017).
URL https://www.kaggle.com/snapcrack/all-the-news

[31] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (Jan) (2003) 993–1022.

[32] I. U. Ogul, C. Ozcan, O. Hakdagli, Keyword extraction based on word synonyms using word2vec, in: 2019 27th Signal Processing and Communications Applications Conference (SIU), 2019, pp. 1–4.

[33] M. Tezgider, B. Yıldız, G. Aydın, Improving word representation by tuning word2vec parameters with deep learning model, in: 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018, pp. 1–7.

[34] T. Donkers, B. Loepp, J. Ziegler, Sequential user-based recurrent neural network recommendations, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 152–160. doi:10.1145/3109859.3109877.
URL https://doi.org/10.1145/3109859.3109877

[35] S. Chakraborty, J. Banik, S. Addhya, D. Chatterjee, Study of dependency on number of lstm units for character based text generation models, in: 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020, pp. 1–5.

MS Thesis

ORIGINALITY REPORT

15%	4%	9%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Jui-Feng Yeh, Chen-Hsien Lee, Yi-Shiuan Tan, Liang-Chih Yu. "Topic model allocation of conversational dialogue records by Latent Dirichlet Allocation", Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, 2014 Publication	1%
2	Submitted to Higher Education Commission Pakistan Student Paper	1%
3	Cenk Anil Bahcevan, Emirhan Kutlu, Tugba Yildiz. "Deep Neural Network Architecture for Part-of-Speech Tagging for Turkish Language", 2018 3rd International Conference on Computer Science and Engineering (UBMK), 2018 Publication	1%
4	Mehmet BAYGIN. "Classification of Text Documents based on Naive Bayes using N-Gram Features", 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018	1%

	Publication	
5	Submitted to Indian Institute of Technology, Kanpur Student Paper	1%
6	Submitted to International Black Sea University Student Paper	1%
7	Rong Jin. "Title Generation Using a Training Corpus", Lecture Notes in Computer Science, 2001 Publication	1%
8	www.mitpressjournals.org Internet Source	1%
9	Submitted to An-Najah National University Student Paper	1%
10	Submitted to Delhi Technological University Student Paper	<1%
11	Yi Bin, Yang Yang, Fumin Shen, Xing Xu, Heng Tao Shen. "Bidirectional Long-Short Term Memory for Video Description", Proceedings of the 2016 ACM on Multimedia Conference - MM '16, 2016 Publication	<1%
12	Sayed Saniya Salim, Agrawal Nidhi Ghanshyam, Darkunde Mayur Ashok, Dungarpur Burhanuddin Mazahir, Bhushan S. Thakare. "Deep LSTM-RNN with Word	<1%

Embedding for Sarcasm Detection on Twitter", 2020 International Conference for Emerging Technology (INCET), 2020
Publication

---

13 Jeong-Woo Son, Wonjoo Park, Sang-Yun Lee, Sun-Joong Kim. "Video scene title generation based on explicit and implicit relations among caption words", 2018 20th International Conference on Advanced Communication Technology (ICACT), 2018
Publication

<1%

---

14 Ying Wang, Hua-Xi Yu, Xian-Yong Xiao, Ming Ma, Yuan Zhou. "New method for calculating voltage dip/swell types based on six-dimensional vectors and Euclidean distance", IET Generation, Transmission & Distribution, 2019
Publication

<1%

---

15 Submitted to Queensland University of Technology
Student Paper

<1%

---

16 Submitted to University of San Francisco
Student Paper

<1%

---

17 Submitted to CSU, Fullerton
Student Paper

<1%

---

18 Ravali Boorugu, G. Ramesh. "A Survey on NLP

<1%

based Text Summarization for Summarizing Product Reviews", 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020
Publication

19	Alok Kumar Singh Kushwaha, Rajeev Srivastava. "Framework for dynamic background modeling and shadow suppression for moving object segmentation in complex wavelet domain", Journal of Electronic Imaging, 2015 Publication	<1%
20	Submitted to The Robert Gordon University Student Paper	<1%
21	open.library.ubc.ca Internet Source	<1%
22	psasir.upm.edu.my Internet Source	<1%
23	"Intelligent Technologies and Applications", Springer Science and Business Media LLC, 2020 Publication	<1%
24	avesis.hacettepe.edu.tr Internet Source	<1%
25	docplayer.net Internet Source	<1%

26	ieeexplore.ieee.org Internet Source	<1%
27	"Frontier Computing", Springer Science and Business Media LLC, 2020 Publication	<1%
28	Submitted to uva Student Paper	<1%
29	scholarcommons.usf.edu Internet Source	<1%
30	doras.dcu.ie Internet Source	<1%
31	hdl.handle.net Internet Source	<1%
32	link.springer.com Internet Source	<1%
33	"Information, Communication and Computing Technology", Springer Science and Business Media LLC, 2019 Publication	<1%
34	"Evolutionary Computing and Mobile Sustainable Networks", Springer Science and Business Media LLC, 2021 Publication	<1%
35	researchspace.ukzn.ac.za Internet Source	<1%

36	"Proceedings of Fifth International Congress on Information and Communication Technology", Springer Science and Business Media LLC, 2021 Publication	<1%
37	Rushali Dhumal Deshmukh, Arvind Kiwelekar. "Deep Learning Techniques for Part of Speech Tagging by Natural Language Processing", 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020 Publication	<1%
38	academic.oup.com Internet Source	<1%
39	repository.bilkent.edu.tr Internet Source	<1%
40	scholar.uwindsor.ca Internet Source	<1%
41	creativecommons.org Internet Source	<1%
42	repository.nwu.ac.za Internet Source	<1%
43	www.semanticscholar.org Internet Source	<1%
44	Pratanwanich, Naruemon, and Pietro Lio.	

	"Exploring the complexity of pathway–drug relationships using latent Dirichlet allocation", Computational Biology and Chemistry, 2014. Publication	<1%
45	www.simplilearn.com Internet Source	<1%
46	www.ijrte.org Internet Source	<1%
47	www.mdpi.com Internet Source	<1%
48	bmcmedresmethodol.biomedcentral.com Internet Source	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		