International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Comparative Analysis of Lexical Chain, Bidirectional Encoder Representations from Transformers (BERT) and Graph based Approaches for Condensation of Low Resource Language Documents

Pranjali Deshpande * [1 2], Dr. Sunita Jahirabadkar[1 2]

Submitted: 05/12/2023 **Revised:** 14/01/2024 **Accepted:** 28/01/2024

Abstract: Humans use language as their primary and exclusive means of communication. There are around 7000 different languages spoken in the world. Among them, Low Resource Languages (LRL) are ones that do not have the linguistic resources required to create statistical NLP applications. The most common way that people express and store their thoughts is through writing. Technological developments are making the world smaller by making distant communication more accessible. Owing to the rise in internet usage, fresh textual content is created every second. Not all of the information in this text is helpful. In light of this, document condensation or summarization is becoming a more important responsibility. There are two methods for creating summaries: extractive and abstractive. While essential phrases and sentences from the original document are kept in an extractive summary, an abstractive summary is created by reworking the main sentences. When it comes to LRL materials, summarizing becomes more difficult. The studies for condensation or automatic summarization of LRL documents using BERT, lexical chain and Graph based approach are the main topic of this study.

Keywords: *Automatic Summarization, BERT, Extractive Summarization, Graph based approach, Low Resource Languages, Lexical Chain.*

1. Introduction

The ability to communicate oneself through an assortment of communication avenues sets humans apart from other animal groups. The most prevalent means of communication are body language, spoken words, written words, etc. The COVID-19 pandemic recently caused difficulty for everyone on the planet. It was impossible for people to physically communicate with one another. A range of digital media is proving to be useful for distant communication in these conditions, supporting the well-being of individuals. The world has become closer as a result of developments in digital technology. Written communication has become the most preferred form of digital communication among other forms. The rationale is that written papers provide a handy means of expressing, sharing, and preserving ideas. Electronically preserving the documents is also an option. A more socially and economically diverse community of people has been using electronic communication in their mother language in recent years, thanks to the digital transition. Many industries, including healthcare and education, are quickly moving to electronic communication. The support for the native language spoken by the residents of a particular place has grown as a result of the use of numerous tools and techniques in the field of natural language processing.

Over seven thousand languages are spoken by humans worldwide. English has the greatest online presence even though it is not the most spoken language in the world. There are a lot of English-language electronic text resources available. The Sanskrit language is the source of numerous Indo-Aryan languages, including Marathi, Hindi, and Kannada, among many others. These languages are not very well-represented on the internet. Low Resource Languages (LRL) are those languages that are devoid from extensive corpora. There is a sharp rise in the quantity of e-text from different domains in regional languages due to widespread internet use. Finding the relevant and needed information inside the lengthy documents created by different processes in fields like law, medicine, etc. is a difficult endeavour. A document is made up of several coherent, connected sentences. The true context and main ideas of any document are contained in a select few sentences. The sentences that follow are all supporting sentences. Automatic document summarizing is quite helpful in obtaining the main points of the text.

A summary is a text that has been condensed from one or more documents. Contextually significant information from the original document is included in the summary. A document's summary shouldn't be longer than half of the original. There are two methods for automatically summarising information: extractive and abstractive. The process of creating an extractive summary involves retaining the essential words or sentences from the original content. Rewriting the document's context-bearing words yields an abstractive summary. Natural Language Understanding (NLU) and Natural Language Generation (NLG) are the two core tasks that comprise the Natural Language Processing area. An extractive summary can be produced only by applying NLU principles and methods. It is necessary to have both NLU and NLG knowledge while creating abstractive summaries. Therefore, creating an extractive summary is easier than creating an abstractive summary. When the original documents are written in LRL, the condensing process is more difficult. Each language is distinct in its own way. Ambiguities of different kinds appear at every step of NLP, from phonetics to pragmatics. This linguistic variability makes it impossible to create generic summarization models. Within this framework, the research focuses on the experiments conducted using three different approaches: Lexical

[1] *Department of Computer Engineering, MKSSS's Cummins College of Engg. for Women, Pune-52, India*
ORCID ID: 0000-0003-2927-1376
[2] *Department of Computer and Information Technology (CI), Department of Technology, SPPU, Pune-07, India*
** Corresponding Author Email: Pranjali.deshpande@cumminscollege.in*

Chaining technique, Bidirectional Encoder Representations from Transformers (BERT) approach and the Graph-based approach. The investigation, experimentation, and analysis are covered in Section 2, which is followed by the conclusion.

2. Analysis of approaches for text condensation

The task of extractive summarization has been approached from a variety of angles by the researchers [1]. Lexical chaining, BERT, Semantic Triplet, Zero Shot Learning, WorldNet-based approach, and graph-based models are a few of these methods. The BERT approach, lexical chaining approach and graph-based approach are the three main topics of this work. The lexical chain technique is a more traditional method than BERT and graph-based condensation among these approaches. The comparative analysis of the tests conducted using these three methodologies is covered in this section. Initially, extractive summarization implementation is broken down into several smaller tasks. Among these smaller tasks are preprocessing, creating semantic connections between the phrases, identifying potential sentences for a summary by grading them, then create an extractive summary. Three ways are used in experiments to carry out these tasks. 1. The use of lexical chains 2. Transformer-Based Bidirectional Encoder Representations (BERT) and 3. Graph based approach.

1. Lexical Chaining approach:

The lexical cohesiveness notion and WordNet are utilized in the lexical chain [2] [3] technique to construct extractive summarizer. By using this method, the extractive summary is produced by first analyzing the input content and selecting the terms that occur as noun entries. These terms have been designated as candidates, and every candidate word has been selected in stage one. Words are added to the chain according to their senses once the input text has been preprocessed. This creates the chain structure. Let's take a look at this chain creation example to better grasp the idea: Mr. Atul is the inventor of a device that regulates the rate at which medication is injected into the bloodstream using tiny computers. The figures below illustrate two possible interpretations:

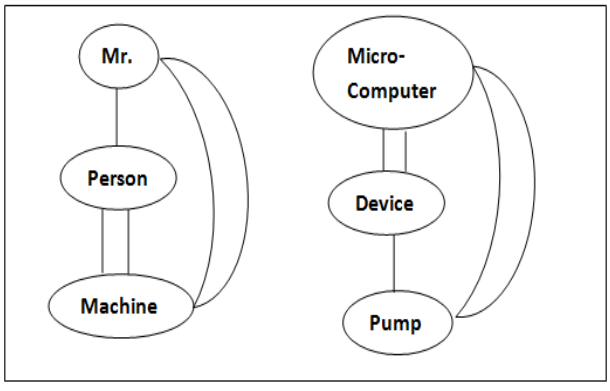

Fig. 1. Illustration 1 [2]

\

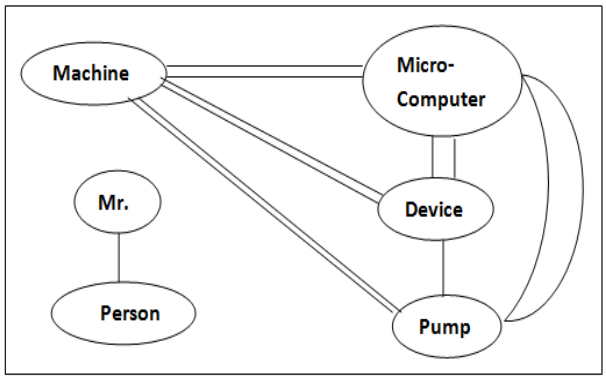

Fig. 2. Illustration 2 [2]

The chains between the cohesive texts are established, as Fig. 1. and Fig. 2. demonstrate. For this, WordNet [4] is employed. If there are more connectors, the interpretation is thought to be more cohesive. The length and homogeneity index are the two criteria used to generate the chain scores, which add up to the interpretation score. The length of the chain represents the number of times each component has occurred. The formula for calculating the Homogeneity Index is float (float (length-Len (temp))/length). Thus, the following formula is used to determine the final score:

Score of Chain=float (length*homogeneity Index) (1)

There are three sorts of chains that can be distinguished: weak, medium, and strong. This classification is based on the chain's strength, which can be determined using the following parameter:

Strength Criterion = Score (Chain) > Average (Scores)

+2*Standard Deviation (Scores) (2)

This table provides an overview of the lexical chaining methodology:

Table 1. Lexical chain approach

Methodology	Technique used	Remarks
Lexical chains for text summarization (1999) [2]	Step 1: Preprocessing: POS tagging is completed initially. The noun entries are known to be in WordNet. A list of potential summary terms is chosen from the retrieved entries.	1. WordNet contains a limited number of SynSets. 2.The sentence's ranking is influenced by its length. The summary's length is uncontrollable.
	Step 2: A calculation of the words' relatedness is made. WordNet word distance and occurrence are factors that determine relatedness.	
	Step 3: Sentence grading is based on relatedness criteria. The homogeneity index and sentence length are examples of parameters that are used to create summaries.	

Using Python, the experiment is conducted on a dataset of documents written in Devnagri Script [5]. The results of implementation can be seen in the Fig.3.

Original document:

कुलकर्णी, सलील श्रीनिवास संगीतकार ६ ऑक्टोबर १९७२. गेल्या दहा-बारा वर्षांत युवा पिढीची स्पंदने आपल्या सुरावटीतून नेमकेपणाने व्यक्त करणारा संगीतकार ही डॉ सलील कुलकर्णी यांची ओळख आहे. सुरावट आणि वाद्यमेळ यात कविता हरवून जाऊ नये, यासाठी गीतातील शब्दांकडे सूक्ष्मतेने आणि संवेदनशीलतेने पाहणे, हे सलील यांचे वैशिष्ट्य मानले जाते.मुंबईच्या दादर परिसरात जन्मलेले सलील हे श्रीनिवास कुलकर्णी आणि रेखा कुलकर्णी या दांपत्याचे चिरंजीव. लहानपणापासून संगीताकडे त्यांचा कल असला तरी तो छंद जपत त्यांनी भारती विद्यापीठाच्या वैद्यक महाविद्यालयाची पदवी प्राप्त केली. एम.बी.बी.एस. झाल्यावर काही वर्षे वैद्यकीय व्यवसायाचा अनुभव घेतल्यावर आपला कल संगीताकडेच आहे, हे उमगून वैद्यकीय पेशा बाजूला ठेवून पूर्णविळ संगीतकार म्हणून काम करण्याचा निर्णय घेतला.दरम्यान त्यांनी पं. गंगाधरबुवा पिंपळखरे, जयमाला शिलेदार आणि प्रमोद मराठे यांच्याकडून संगीताचे मार्गदर्शन घेतले. या मार्गदर्शनाला डॉ. सलील यांनी स्वत:च्या एकलव्य वृत्तीची जोड देऊन संगीताची जाण समृद्ध केली. साहित्याविषयीचा जिव्हाळाही प्रथमपासून असल्याने, संगीतरचना करण्याकडे सलील यांचा अधिक कल राहिला.उत्तमोत्तम कवितांना चाली लावण्याचा प्रयत्न सुरू असतानाच, त्यांची 'तरीही वसंत फुलतो', ही पहिली ध्वनिफीत १९९९ मध्ये प्रकाशित झाली. त्यापाठोपाठ 'स्वप्नांत पाहिली राणीची बाग', 'संधिप्रकाशात', 'अबोली', 'सांग स'या रे','आनंदपहाट', 'स्वरभाव', 'आयुष्यावर बोलू काही', 'अग्गोबाई ढग्गोबाई'... अशा अनेक ध्वनिफिती प्रकाशित झाल्या. 'आयुष्यावर बोलू काही' या ध्वनिफीतीप्रमाणेच त्याचे रंगमंचीय कार्यक'मही लोकप्रिय झाले. कवी संदीप खरे यांच्या कवितांना सलील यांनी दिलेल्या चाली युवा वर्गात विलक्षण गाजल्या. समकालीन कलाकार ध्वनिमुद्रित गाण्यांचे कार्यक'म करत असताना फक्त स्वत:च्या रचनांचे कार्यक'म लोकप्रिय करून स्वत:चे श्रोते निर्माण करण्याची किमया नव्या पिढीत डॉ. सलील यांनी संदीप खरे यांच्या साथीने करून दाखवली आहे. ध्वनिफिती, सीडी यानंतरचा टप्पा चित्रपटांचा होता. २००३ मध्ये सलील यांचे संगीत असलेला 'विठ्ठल विठ्ठल' हा चित्रपट प्रदर्शित झाला. त्यानंतर 'पांढर', 'आनंदीआनंद', 'बदाम राणी गुलाम चोर', 'चिंटू' आदी सुमारे २५ चित्रपटांना सलील यांनी संगीत दिले आहे. छोट्या पडद्यावरचा 'नक्षत्रांचे देणे' हा कार्यक'म विशेष लक्षणीय ठरला. तसेच सारेगमपचे परीक्षणही त्यांनी केले. 'मधली सुट्टी' हा कार्यक'मही ते छोट्या पडद्यावर सादर करतात. त्यांनी वृत्तपत्रीय सदरलेखनही केले असून 'लपवलेल्या काचा' हे पुस्तकही प्रकाशित झाले आहे. झी गौरव, मटा सन्मान, राज्य पुरस्काराचेही ते मानकरी आहेत. संगीतकार म्हणून काम करताना पार्श्वगायक म्हणूनही त्यांनी मोजकेच काम करून ठसा उमटवला आहे.कविता, गाणी आणि आठवणी, तसेच संगीतरचनेची प्रकि'या याविषयी सलील नेहमीच भरभरून बोलतात आणि अनेक रंगमंचीय कार्यक'मात ते स्वत:च सूत्रधाराची भूमिका निभावतात. सलील हे ज्येष्ठ संगीतकार-गायक पं. हृदयनाथ मंगेशकर यांच्यासह 'मैत्र जिवांचे' हा कार्यक'मही सादर करतात. त्यांनी पुण्यात सलील कुलकर्णी म्युझिक स्कूल ही सुरू केले असून याद्वारे ते युवा कलाकारांना मार्गदर्शन करतात.सलील कुलकर्णी यांची पत्नी अंजली राज्य पुरस्कार प्राप्त गायक असून मुलगा शुभंकरही या क्षेत्रात आपले स्थान निश्चिंत करण्याच्या मार्गावर आहे.

Total words: 397

Generated Summary:

६ ऑक्टोबर १९७२गेल्या दहा-बारा वर्षांत युवा पिढीची स्पंदने आपल्या सुरावटीतून नेमकेपणाने व्यक्त करणारा संगीतकार ही डॉ सलील कुलकर्णी यांची ओळख आहे. सुरावट आणि वाद्यमेळ यात कविता हरवून जाऊ नये, यासाठी गीतातील शब्दांकडे सूक्ष्मतेने आणि संवेदनशीलतेने पाहणे, हे सलील यांचे वैशिष्ट्य मानले जाते. मुंबईच्या दादर परिसरात जन्मलेले सलील हे श्रीनिवास कुलकर्णी आणि रेखा कुलकर्णी या दांपत्याचे चिरंजीव. लहानपणापासून संगीताकडे त्यांचा कल असला तरी तो छंद जपत त्यांनी भारती विद्यापीठाच्या वैद्यक महाविद्यालयाची पदवी प्राप्त केली. झाल्यावर काही वर्षे वैद्यकीय व्यवसायाचा अनुभव घेतल्यावर आपला कल संगीताकडेच आहे, हे उमगून वैद्यकीय पेशा बाजूला ठेवून पूर्णविळ संगीतकार म्हणून काम करण्याचा निर्णय घेतला. उत्तमोत्तम कवितांना चाली लावण्याचा प्रयत्न सुरू असतानाच, त्यांची 'तरीही वसंत फुलतो', ही पहिली ध्वनिफीत १९९९ मध्ये प्रकाशित झाली. त्यापाठोपाठ 'स्वप्नांत पाहिली राणीची बाग', 'संधिप्रकाशात', 'अबोली', 'सांग स'या रे', 'आनंदपहाट', 'स्वरभाव', 'आयुष्यावर बोलू काही', 'अग्गोबाई ढग्गोबाई'. कवी संदीप खरे यांच्या कवितांना सलील यांनी दिलेल्या चाली युवा वर्गात विलक्षण गाजल्या. समकालीन कलाकार ध्वनिमुद्रित गाण्यांचे कार्यक'म करत असताना फक्त स्वत:च्या रचनांचे कार्यक'म लोकप्रिय करून स्वत:चे श्रोते निर्माण करण्याची किमया नव्या पिढीत डॉ. त्यानंतर 'पांढर', 'आनंदीआनंद', 'बदाम राणी गुलाम चोर', 'चिंटू' आदी सुमारे २५ चित्रपटांना सलील यांनी संगीत दिले आहे. संगीतकार म्हणून काम करताना पार्श्वगायक म्हणूनही त्यांनी मोजकेच काम करून ठसा उमटवला आहे. कविता, गाणी आणि आठवणी, तसेच संगीतरचनेची प्रकि'या याविषयी सलील नेहमीच भरभरून बोलतात आणि अनेक रंगमंचीय कार्यक'मात ते स्वत:च सूत्रधाराची भूमिका निभावतात. त्यांनी पुण्यात सलील कुलकर्णी म्युझिक स्कूल ही सुरू केले असून याद्वारे ते युवा कलाकारांना मार्गदर्शन करतात. सलील कुलकर्णी यांची पत्नी अंजली राज्य पुरस्कार प्राप्त गायक असून मुलगा शुभंकरही या क्षेत्रात आपले स्थान निश्चिंत करण्याच्या मार्गावर आहे.

Total words:251

Fig. 3. Output of the lexical chain approach implementation [2]

2. BERT approach:

In 2019, researchers at Google Brains created the BERT algorithm [6] [3]. Transformer Neural Networks, which were first created to address the issue of language translation, are utilised by BERT. In the text processing problem, and more especially in the summarization problem, the following factors show that transformer architecture is more effective than previous Long Short-Term Memory (LSTM): The first limitation is that, in comparison to Transformer Architecture, LSTM Networks train more slowly. The sequential processing that the LSTM Network does is its second disadvantage. Sequential processing of the words may cause the text's meaning to be lost. The context is learned both left to right and right to left even in bidirectional LSTM, and the outputs are concatenated, which is ineffective. Transformer architecture processes words concurrently, resulting in faster processing than LSTM. Simultaneous learning from both sides also improves context learning. The encoder and decoder are the two halves of the BERT architecture. The words in the text are accepted as input by the encoder. Word embeddings are created as the words are accepted concurrently. Word embeddings capture a word's meaning. The Decoder is the second part. Along with previously created words, these embeddings are taken by the decoder. These encoders are stacked to form BERT, which uses them to comprehend language. It also fine tunes them to learn specific tasks. The Transformer architecture is shown in Fig. 4. [7]

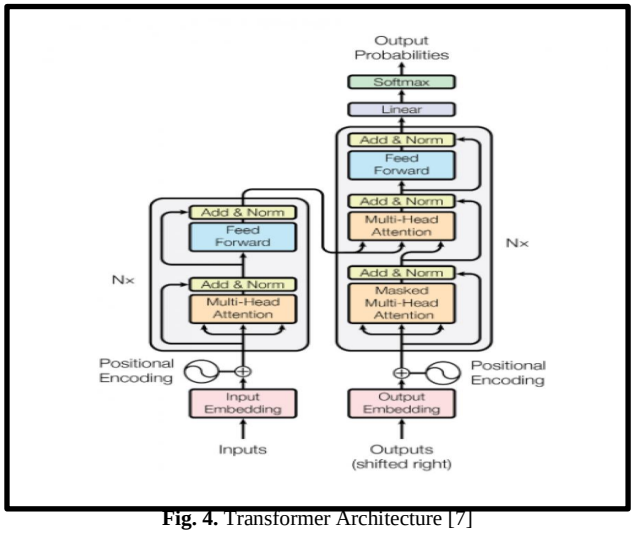

Fig. 4. Transformer Architecture [7]

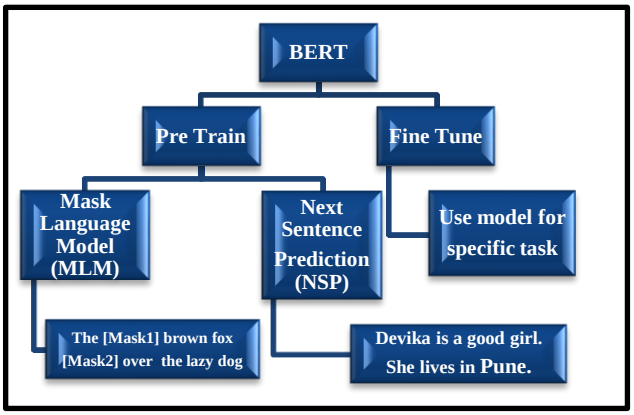

Fig. 5. Phases of BERT [7]

The Fig. 5. illustrates the two phases: pretraining, which involves learning the language, and fine tuning, which involves solving a particular problem.

The two subphases that comprise the pre-training phase are as follows:

1. Mask Language Model (MSM): BERT uses sentences that contain masks in this phase, and its output is intended to be masked tokens.

2. Next Sentence Prediction (NSP): BERT ascertains if the second sentence comes after the first in this step.

BERT uses the subsequent procedure to handle text:

Tokenizing the text is completed first. By using the Word Piece Tokenizer, which BERT employs, words can also be divided into tokens according to prefixes or suffixes. The word "melodically," for instance, is divided into melodic and ally. Additionally, punctuation is kept since it adds meaning to the sentence. Each token is given a vector in the following stage, which is referred to as embedding. Token meaning is contained in embedding vectors. No matter the context, the same token will always receive the same embedding according to dictionary lookout. Here, focus is really important. By examining the context in which they appear, it modifies the default embedding.

To further understand this, let's look at an example text: "Walk by river bank." In this sentence, each token is replaced with the default embedding, which is a vector of 768 components. The scalar product between embedding pairs is then computed. Words with comparable embeddings have greater scalar product values, indicating a strong link between them. Each and every feasible pair of embedding vectors has their scalar product computed. To improve the numerical behavior, the values are square root of 768-scaled down. The Soft Max activation function is applied to the scaled values. Soft max removes the small values and increases high values exponentially. Additionally, normalization is carried out. New embeddings are produced by utilizing the soft max function. Given that they comprise a portion of each input embedding, these are contextual embeddings. A significant portion of the new embedding comprises related embeddings if there is a substantial correlation between the token and another. The new embedding is the same as the input embedding if there is little correlation between them. The tokens "bank" and "river" ought to have significant values. As a result, the new embeddings of "river" and "bank" combine the two embeddings equally. Since "by" is neutral, there is little relationship between any of its embeddings and the others. However, we don't have to use the input embedding vectors exactly as they are; instead, we need to project them using the Key, Query, and Value vectors, three linear projections.

Each of these vectors consists of 64 elements. Every projection in this set focuses on a distinct vector space direction that corresponds to a different semantic feature. A projection of a preposition and a location query is what makes up a key. For instance, since "by" is a strong component towards prepositions and all other tokens have strong components towards places, the key of the token "by" in this instance should have a strong relationship with every other question. The values could alternatively represent a different projection, such as the direction of a physical site. Contextualized embeddings are produced by combining all these values together. The same procedure can be carried out repeatedly using various key, value, and query projections. This creates multi-head attention. Each head is capable of concentrating on a distinct input embedding projection. Preposition and location relationship calculations, for instance, are done by one head. The subject-verb relationship can be calculated by another person. Each head's output is combined to create a big vector. Twelve heads are used by BERT, meaning that the final output has 1768 contextualized embedded tokens that are the same length as the input vectors. Next, positional embeddings are coupled with input embeddings. Vectors with positional information in sequence, even before attention is applied, are known as positional embeddings. As a result, given the token order, attention may compute the relationships. Because the soft max function is non-linear, attention can be given repeatedly.

Table 2 summarizes the approach of BERT:

Table 2. BERT approach

Methodology	Technique used	Remarks
BERT for text summarization (2019) [7]	Step 1: Preprocessing is performed. The text input is tokenized. The sentences are extracted using NLTK. Sentences that are excessively brief or lengthy, as well as those that begin with a conjunction, are eliminated.	Sentences that have been eliminated influence centroid selection.
	Step 2: The BERT model receives the tokenized sentences and uses them to create embeddings. Between 10% and 15% of the words are masked. The formula N (number of sentences) * W (tokenized words) * E (embeddings) is used to build the matrix. The pre-trained BERT library for Pytorch is utilised.	The requirements for memory and computing are higher.
	Step 3 : The K-means approach is used to cluster the embeddings. Sentences that are closest to the centroid are selected.	More clusters are required for large texts in order to keep context.

The experiment is carried out on the dataset of documents written in Devnagri Script [5], using Python language. The results of implementation can be seen in the Fig. 6.

Original document:

कुलकर्णी, सलील श्रीनिवास संगीतकार ६ ऑक्टोबर १९७२. गेल्या दहा-बारा वर्षांत युवा पिढीची स्पंदने आपल्या सुरावटीतून नेमकेपणाने व्यक्त करणारा संगीतकार ही डॉ सलील कुलकर्णी यांची ओळख आहे. सुरावट आणि वाद्यमेळ यात कविता हरवून जाऊ नये, यासाठी गीतातील शब्दांकडे सूक्ष्मतेने आणि संवेदनशीलतेने पाहणे, हे सलील यांचे वैशिष्ट्य मानले जाते.मुंबईच्या दादर परिसरात जन्मलेले सलील हे श्रीनिवास कुलकर्णी आणि रेखा कुलकर्णी या दांपत्याचे चिरंजीव. लहानपणापासून संगीताकडे त्यांचा कल असला तरी तो छंद जपत त्यांनी भारती विद्यापीठाच्या वैद्यक महाविद्यालयाची पदवी प्राप्त केली. एम.बी.बी.एस. झाल्यावर काही वर्षे वैद्यकीय व्यवसायाचा अनुभव घेतल्यावर आपला कल संगीताकडेच आहे, हे उमगून वैद्यकीय पेशा बाजूला ठेवून पूर्णवेळ संगीतकार म्हणून काम करण्याचा निर्णय घेतला.दरम्यान त्यांनी पं. गंगाधरबुवा पिंपळखरे, जयमाला शिलेदार आणि प्रमोद मराठे यांच्याकडून संगीताचे मार्गदर्शन घेतले. या मार्गदर्शनाला डॉ. सलील यांनी स्वत:च्या एकलव्य वृत्तीची जोड देऊन संगीताची जाण समृद्ध केली. साहित्याविषयीचा जिव्हाळाही प्रथमपासून असल्याने, संगीतरचना करण्याकडे सलील यांचा अधिक कल राहिला.उत्तमोत्तम कवितांना चाली लावण्याचा प्रयत्न सुरू असतानाच, त्यांची 'तरीही वसंत फुलतो', ही पहिली ध्वनिफीत १९९९ मध्ये प्रकाशित झाली. त्यापाठोपाठ 'स्वप्नांत पाहिली राणीची बाग', 'संधिप्रकाशात', 'अबोली', 'सांग स'या रे','आनंदपहाट', 'स्वरभाव', 'आयुष्यावर बोलू काही', 'अग्गोबाई ढग्गोबाई'... अशा अनेक ध्वनिफिती प्रकाशित झाल्या. 'आयुष्यावर बोलू काही' या ध्वनिफीतीप्रमाणेच त्याचे रंगमंचीय कार्यक'मही लोकप्रिय झाले. कवी संदीप खरे यांच्या कवितांना सलील यांनी दिलेल्या चाली युवा वर्गात विलक्षण गाजल्या. समकालीन कलाकार ध्वनिमुद्रित गाण्यांचे कार्यक'म करत असताना फक्त स्वत:च्या रचनांचे कार्यक'म लोकप्रिय करून स्वत:चे श्रोते निर्माण करण्याची किमया नव्या पिढीत डॉ. सलील यांनी संदीप खरे यांच्या साथीने करून दाखवली आहे. ध्वनिफिती, सीडी यानंतरचा टप्पा चित्रपटांचा होता. २००३ मध्ये सलील यांचे संगीत असलेला 'विठ्ठल विठ्ठल' हा चित्रपट प्रदर्शित झाला. त्यानंतर 'पांढर', 'आनंदीआनंद', 'बदाम राणी गुलाम चोर', 'चिंटू' आदी सुमारे २५ चित्रपटांना सलील यांनी संगीत दिले आहे. छोट्या पडद्यावरचा 'नक्षत्रांचे देणे' हा कार्यक'म विशेष लक्षणीय ठरला. तसेच सारेगमपचे परीक्षणही त्यांनी केले. 'मधली सुट्टी' हा कार्यक'मही ते छोट्या पडद्यावर सादर करतात. त्यांनी वृत्तपत्रीय सदरलेखनही केले असून 'लपवलेल्या काचा' हे पुस्तकही प्रकाशित झाले आहे. झी गौरव, मटा सन्मान, राज्य पुरस्काराचेही ते मानकरी आहेत. संगीतकार म्हणून काम करताना पार्श्वगायक म्हणूनही त्यांनी मोजकेच काम करून ठसा उमटवला आहे.कविता, गाणी आणि आठवणी, तसेच संगीतरचनेची प्रकि'या याविषयी सलील नेहमीच भरभरून बोलतात आणि अनेक रंगमंचीय कार्यक'मात ते स्वत:च सूत्रधाराची भूमिका निभावतात. सलील हे ज्येष्ठ संगीतकार-गायक पं. हृदयनाथ मंगेशकर यांच्यासह 'मैत्र जिवांचे' हा कार्यक'मही सादर करतात. त्यांनी पुण्यात सलील कुलकर्णी म्युझिक स्कूल ही सुरू केले असून याद्वारे ते युवा कलाकारांना मार्गदर्शन करतात.सलील कुलकर्णी यांची पत्नी अंजली राज्य पुरस्कार प्राप्त गायक असून मुलगा शुभंकरही या क्षेत्रात आपले स्थान निश्चित करण्याच्या मार्गावर आहे.

Total words: 397

Generated Summary:

कुलकर्णी, सलीलश्रीनिवास संगीतकार ६ ऑक्टोबर १९७२ गेल्या दहा-बारा वर्षांत युवा पिढीची स्पंदने आपल्या सुरावटीतून नेमकेपणाने व्यक्त करणारा संगीतकार ही डॉ सलील कुलकर्णी यांची ओळख आहे. लहानपणापासून संगीताकडे त्यांचा कल असला तरी तो छंद जपत त्यांनी भारती विद्यापीठाच्या वैद्यक महाविद्यालयाची पदवी प्राप्त केली. दरम्यान त्यांनी पं. साहित्याविषयीचा जिव्हाळाही प्रथमपासून असल्याने, संगीतरचना करण्याकडे सलील यांचा अधिक कल राहिला. उत्तमोत्तम कवितांना चाली लावण्याचा प्रयत्न सुरू असतानाच, त्यांची 'तरीही वसंत फुलतो', ही पहिली ध्वनिफीत १९९९ मध्ये प्रकाशित झाली. समकालीन कलाकार ध्वनिमुद्रित गाण्यांचे कार्यक'म करत असताना फक्त स्वत:च्या रचनांचे कार्यक'म लोकप्रिय करून स्वत:चे श्रोते निर्माण करण्याची किमया नव्या पिढीत डॉ. संगीतकार म्हणून काम करताना पार्श्वगायक म्हणूनही त्यांनी मोजकेच काम करून ठसा उमटवला आहे.

Total words: 108

Fig. 6. Output of the BERT approach implementation [7]

3. Graph-based approach:

The graph-based framework serves as foundation of graph-based approach [8]. Four tasks comprise this approach. The first task creates a text graph model based on input text. The generated text graph is searched for sentence selection in the candidate summary during the second and third phases. The user has the option to select how long the summary is. The fourth algorithm chooses the most significant sentences if the summary is longer than the allowed number. Graph-based, statistical-based, semantic-based, and centrality-based approaches are the four methodologies that are integrated. In Fig. 7. the framework is shown.

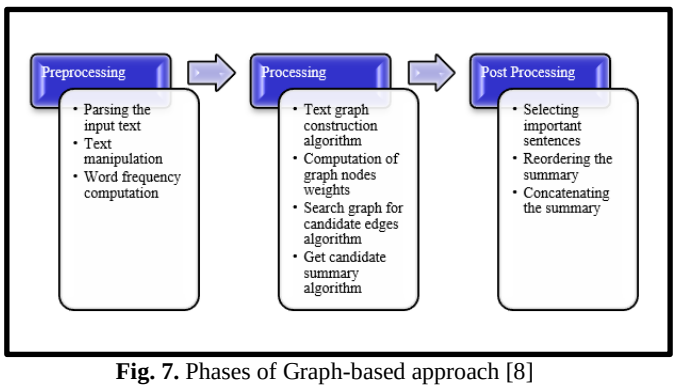

Fig. 7. Phases of Graph-based approach [8]

The preprocessing stage is the first. Sentence segmentation, word tokenization, Part of Speech (PoS) tagging, stemming, and other tasks are completed during this phase. The hyphen is eliminated.

Only one word is retained for each synonym after they have been identified. That word takes the place of every synonym. Using word2vec function, all the words are transformed to vector form [11]. Furthermore, the frequency of every word is calculated. Every word's frequency (W_f) is calculated by the below formula:

W_f (node) = word frequency in t + word frequency in K +

word frequency in S. (3)

The processing step is the second stage.

There are four tasks in this phase. 1. Text graph construction.2. Weight computation for graph nodes 3. Look for candidate edges in the graph. 4. Obtain candidate summary.

Using the Trigrams'n'Tags (TnT) tagger [9], Marathi language data is handled. TnT is a statistical tagger that can be trained using Python's Natural Language Tool Kit (NLTK) [10] on a variety of languages. It is a statistical tagger that applies to Markov models of second order. It is an extremely effective PoS tagger that can be trained on any dataset and in a variety of languages. After creation, it can also be trained using the train method. 25 million words from Marathi corpus are used in this project. The text graph creation approach is used in the first Preprocessing task. A directed graph G is constructed using the text graph creation algorithm, where

$$G = (V, E) \forall \{V_i \in V \text{ are nouns}\} \quad (4)$$

An edge (V_i, V_j) connecting two vertices, V_i and V_j, is labelled with non-noun terms. To denote the beginning and ending as S# and E#, two additional nodes are inserted.

Compute the weights of the graph nodes as the second task. Every node in a graph is assigned a weight once it has been constructed. The weight of a node W_n is determined by its word frequency, W_f, where:

$$W_n = W_f + F_t * W_t + F_k * W_k \quad (5)$$

W_t is the title's weight, while F_t is the title's frequency. Word frequency in the keyword list is F_k.

The third algorithm is Search graph for candidate edges algorithm

It takes as an input G, W_t and out criteria and selects source and destination node.

In the next step Get candidate summary algorithm is executed. In this algorithm from Candidate edges list the sentences are selected based on the criteria: Edges > Edge count threshold

In the Post Processing phase, the important sentences are selected for the generation of final summary. The ranking of the sentences based on following ranking criteria:

$$Sentence\ order = \frac{Count\ of\ sentences\ in\ the\ candidate\ summary - Sentence\ index\ in\ the\ summary}{Count\ of\ sentences\ in\ the\ candidate\ summary} \quad (6)$$

After ranking the sentences, they are sorted based on the calculated ranks. At the end similar sentences are clustered using K-means clustering algorithm [12] and summary is generated. Table 3 summarizes the graph-based approach.

Table 3. Graph based approach

Methodology	Technique used	Remarks
Graph-based approach for text summarization (2020) [8]	Step 1: Preprocessing is performed. Sentence segmentation, word tokenization, Part of Speech (PoS) tagging, stemming is done. Frequency of every word in the document is computed.	All the PoS are not considered.
	Step 2: Processing is performed. Four tasks are carried out, namely : 1. Text graph construction.2. Weight computation for graph nodes 3. Look for candidate edges in the graph. 4. Obtain candidate summary.	The nodes does not cover all the PoS.
	Step 3 : In Post processing phase sentences are ranked based on the sentence ranking criteria. Later they are ordered based on computed ranks and final summary is generated.	Covering complete context is crucial.

The experiment is carried out on the dataset of documents written in Devnagri Script [5], using Python language. Fig. 8. shows the results of implementation:

Original document:

Title: कुलकर्णी, सलील श्रीनिवास

संगीतकार ६ ऑक्टोबर १९७२. गेल्या दहा-बारा वर्षांत युवा पिढीची स्पंदने आपल्या सुरावटीतून नेमकेपणाने व्यक्त करणारा संगीतकार ही डॉ सलील कुलकर्णी यांची ओळख आहे. सुरावट आणि वाद्यमेळ यात कविता हरवून जाऊ नये, यासाठी गीतातील शब्दांकडे सूक्ष्मतेने आणि संवेदनशीलतेने पाहणे, हे सलील यांचे वैशिष्ट्य मानले जाते.मुंबईच्या दादर परिसरात जन्मलेले सलील हे श्रीनिवास कुलकर्णी आणि रेखा कुलकर्णी या दांपत्याचे चिरंजीव. लहानपणापासून संगीताकडे त्यांचा कल असला तरी तो छंद जपत त्यांनी भारती विद्यापीठाच्या वैद्यक महाविद्यालयाची पदवी प्राप्त केली. एम.बी.बी.एस. झाल्यावर काही वर्षे वैद्यकीय व्यवसायाचा अनुभव घेतल्यावर आपला कल संगीताकडेच आहे, हे उमगून वैद्यकीय पेशा बाजूला ठेवून पूर्णवेळ संगीतकार म्हणून काम करण्याचा निर्णय घेतला.दरम्यान त्यांनी पं. गंगाधरबुवा पिंपळखरे, जयमाला शिलेदार आणि प्रमोद मराठे यांच्याकडून संगीताचे मार्गदर्शन घेतले. या मार्गदर्शनाला डॉ. सलील यांनी स्वत:च्या एकलव्य वृत्तीची जोड देऊन संगीताची जाण समृद्ध केली. साहित्याविषयीचा जिव्हाळाही प्रथमपासून असल्याने, संगीतरचना करण्याकडे सलील यांचा अधिक कल राहिला.उत्तमोत्तम कवितांना चाली लावण्याचा प्रयत्न सुरू असतानाच, त्यांची ‘तरीही वसंत फुलतो’, ही पहिली ध्वनिफीत १९९९ मध्ये प्रकाशित झाली. त्यापाठोपाठ ‘स्वप्नांत पाहिली राणीची बाग’, ‘संधिप्रकाशात’, ‘अबोली’, ‘सांग स‘या रे’,‘आनंदपहाट’, ‘स्वरभाव’, ‘आयुष्यावर बोलू काही’, ‘अग्गोबाई ढग्गोबाई’... अशा अनेक ध्वनिफिती प्रकाशित झाल्या. ‘आयुष्यावर बोलू काही’ या ध्वनिफीतीप्रमाणेच त्याचे रंगमंचीय कार्यक‘मही लोकप्रिय झाले. कवी संदीप खरे यांच्या कवितांना सलील यांनी दिलेल्या चाली युवा वर्गात विलक्षण गाजल्या. समकालीन कलाकार ध्वनिमुद्रित गाण्यांचे कार्यक‘म करत असताना फक्त स्वत:च्या रचनांचे कार्यक‘म लोकप्रिय करून स्वत:चे श्रोते निर्माण करण्याची किमया नव्या पिढीत डॉ. सलील यांनी संदीप खरे यांच्या साथीने करून दाखवली आहे. ध्वनिफिती, सीडी यानंतरचा टप्पा चित्रपटांचा होता. २००३ मध्ये सलील यांचे संगीत असलेला ‘विठ्ठल विठ्ठल’ हा चित्रपट प्रदर्शित झाला. त्यानंतर ‘पांढर’, ‘आनंदीआनंद’, ‘बदाम राणी गुलाम चोर’, ‘चिंटू’ आदी सुमारे २५ चित्रपटांना सलील यांनी संगीत दिले आहे. छोट्या पडद्यावरचा ‘नक्षत्रांचे देणे’ हा

कार्यक‘म विशेष लक्षणीय ठरला. तसेच सारेगमपचे परीक्षणही त्यांनी केले. ‘मधली सुट्टी’ हा कार्यक‘मही ते छोट्या पडद्यावर सादर करतात. त्यांनी वृत्तपत्रीय सदरलेखनही केले असून ‘लपवलेल्या काचा’ हे पुस्तकही प्रकाशित झाले आहे. झी गौरव, मटा सन्मान, राज्य पुरस्काराचेही ते मानकरी आहेत. संगीतकार म्हणून काम करताना पार्श्वगायक म्हणूनही त्यांनी मोजकेच काम करून ठसा उमटवला आहे.कविता, गाणी आणि आठवणी, तसेच संगीतरचनेची प्रकि‘या याविषयी सलील नेहमीच भरभरून बोलतात आणि अनेक रंगमंचीय कार्यक‘मात ते स्वत:च सूत्रधाराची भूमिका निभावतात. सलील हे ज्येष्ठ संगीतकार-गायक पं. हृदयनाथ मंगेशकर यांच्यासह ‘मैत्र जिवांचे’ हा कार्यक‘मही सादर करतात. त्यांनी पुण्यात सलील कुलकर्णी म्युझिक स्कूल ही सुरू केले असून याद्वारे ते युवा कलाकारांना मार्गदर्शन करतात.सलील कुलकर्णी यांची पत्नी अंजली राज्य पुरस्कार प्राप्त गायक असून मुलगा शुभंकरही या क्षेत्रात आपले स्थान निश्चित करण्याच्या मार्गावर आहे.

Total words: 397

Generated Summary:

गेल्या दहा-बारा वर्षांत युवा पिढीची स्पंदने आपल्या सुरावटीतून नेमकेपणाने व्यक्त करणारा संगीतकार ही डॉ सलील कुलकर्णी यांची ओळख आहे. सुरावट आणि वाद्यमेळ यात कविता हरवून जाऊ नये यासाठी गीतातील शब्दांकडे सूक्ष्मतेने आणि संवेदनशीलतेने पाहणे हे सलील यांचे वैशिष्ट्य मानले जाते.मुंबईच्या दादर परिसरात जन्मलेले सलील हे श्रीनिवास कुलकर्णी आणि रेखा कुलकर्णी या दांपत्याचे चिरंजीव. लहानपणापासून संगीताकडे त्यांचा कल असला तरी तो छंद जपत त्यांनी भारती विद्यापीठाच्या वैद्यक महाविद्यालयाची पदवी प्राप्त केली. झाल्यावर काही वर्षे वैद्यकीय व्यवसायाचा अनुभव घेतल्यावर आपला कल संगीताकडेच आहे हे उमगून वैद्यकीय पेशा बाजूला ठेवून पूर्णवेळ संगीतकार म्हणून काम करण्याचा निर्णय घेतला.दरम्यान त्यांनी पं. सलील यांनी स्वत च्या एकलव्य वृत्तीची जोड देऊन संगीताची जाण समृद्ध केली. साहित्याविषयीचा जिव्हाळाही प्रथमपासून असल्याने संगीतरचना करण्याकडे सलील यांचा अधिक कल राहिला.उत्तमोत्तम कवितांना चाली लावण्याचा प्रयत्न सुरू असतानाच त्यांची ‘तरीही वसंत फुलतो’ ही पहिली ध्वनिफीत १९९९ मध्ये प्रकाशित झाली. त्यापाठोपाठ ‘स्वप्नांत पाहिली राणीची बाग’, ‘संधिप्रकाशात’, ‘अबोली’, ‘सांग सया रे’,‘आनंदपहाट’, ‘स्वरभाव’, ‘आयुष्यावर बोलू काही’, ‘ अग्गोबाई ढग्गोबाई’ ... अशा अनेक ध्वनिफिती प्रकाशित झाल्या. ‘आयुष्यावर बोलू काही’ या ध्वनिफीतीप्रमाणेच त्याचे रंगमंचीय कार्यकमही लोकप्रिय झाले. कवी संदीप खरे यांच्या कवितांना सलील यांनी दिलेल्या चाली युवा वर्गात विलक्षण गाजल्या. समकालीन कलाकार ध्वनिमुद्रित गाण्यांचे कार्यकम करत असताना फक्त स्वत च्या रचनांचे कार्यकम लोकप्रिय करून स्वत चे श्रोते निर्माण करण्याची किमया नव्या पिढीत डॉ. २००३ मध्ये सलील यांचे संगीत असलेला ‘विठ्ठल विठ्ठल ’ हा चित्रपट प्रदर्शित झाला. त्यानंतर ‘पांढर’, ‘आनंदीआनंद’,‘ बदाम राणी गुलाम चोर’, ‘चिंटू’ आदी सुमारे २५ चित्रपटांना सलील यांनी संगीत दिले आहे. छोट्या पडद्यावरचा ‘नक्षत्रांचे देणे’ हा कार्यकम विशेष लक्षणीय ठरला. संगीतकार म्हणून काम करताना पार्श्वगायक म्हणूनही त्यांनी मोजकेच काम करून ठसा उमटवला आहे.कविता गाणी आणि आठवणी तसेच संगीतरचनेची प्रकिया याविषयी सलील नेहमीच भरभरून बोलतात आणि अनेक रंगमंचीय कार्यकमात ते स्वत च सूत्रधाराची भूमिका निभावतात. हृदयनाथ मंगेशकर यांच्यासह ‘मैत्र जिवांचे’ हा कार्यकमही सादर करतात. त्यांनी पुण्यात सलील कुलकर्णी म्युझिक स्कूल ही सुरू केले असून याद्वारे ते युवा कलाकारांना मार्गदर्शन करतात.सलील कुलकर्णी यांची पत्नी अंजली राज्य पुरस्कार प्राप्त गायक असून मुलगा शुभंकरही या क्षेत्रात आपले स्थान निश्चित करण्याच्या मार्गावर आहे.

Total no of words: 322

Fig. 8. Output of the graph-based approach implementation [8]

The studies conducted using the lexical chain approach, BERT and Graph based approach yielded table, which displays the analysis conducted on different parameters:

Table 4. Analysis of all the approaches

Lexical Chaining Approach	BERT Approach	Graph Based Approach
Uses WordNet [4]	Uses Transformer model [6]	Uses Graph based visual representation of the sentences [8]
The words are processed sequentially in a sentence	All the words are passed parallelly and positional encoding is done	The sentences revolve around the context bearing words
Less accuracy is observed as compared to manual summary	More accuracy is observed as compared to manual summary	More accurate contextual referencing is observed
Context is retained for long documents	Result is affected for long documents	All the linguistically important PoS are not considered
Less Flexible	Adaptive model in terms of candidate sentences	Robust model retaining most of the context

3. Conclusion

Automatic summarization of low-resource language documents is a challenging endeavor. The growing number of native languages being used in digital communication is generating enormous amounts of content every day from different domains. In today's world, proper utilization of this data is imperative. Automatic summarization is critical in this scenario. There are many languages spoken throughout the world, but only a small number of them are widely available online. These languages are referred to as Low Resource Languages. Developing statistical applications with LRL data is a difficult undertaking. Automatic summary is a significant application for determining the essence of an article. Summarization is a crucial task for numerous applications across multiple diverse fields. The study focuses on three methods for automatic summarization or condensation of LRL documents: lexical chaining, BERT and graph-based approach. The detailed experiments were performed and analysis was presented. In future the more comprehensive approach can be designed to address the contextually complex linguistic documents.

References

[1] Deshpande, Pranjali, and Sunita Jahirabadkar. "A Survey on Statistical Approaches for Abstractive Summarization of Low Resource Language Documents." Smart Trends in Computing and Communications. Springer, Singapore, 2022. 729-738.

[2] Barzilay, Regina, and Michael Elhadad. "Using lexical chains for text summarization." Advances in automatic text summarization (1999): 111-121.

[3] P. Deshpande and S. Jahirabadkar, "Study of Low Resource Language Document Extractive Summarization using Lexical chain and Bidirectional Encoder Representations from Transformers (BERT)," International Conference on Computational Performance Evaluation (ComPE), 2021, pp. 457-461.

[4] MarathiWordNet: https://www.cfilt.iitb.ac.in/~wordnetbeta/marathiwn/wn.php

[5] www.maharastnayak.in, an initiative by Vivek Magazine.
[6] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton and Toutanova, Kristina. "BERT: Pre- training of Deep Bidirectional Transformers for Language Understanding." Paper Presented at the meeting of the NAACL-HLT (1), 2019.
[7] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz and Polosukhin, Illia "Attention Is All You Need". (2017). cite arxiv: 1706.03762
[8] EI-Kassas, Wafaa S., et al. "EdgeSumm: Graph-based framework for automatic text summarization." Information Processing & Management 57.6 (2020): 102264.
[9] Brants, Thorsten. "TnT-a statistical part-of-speech tagger." arXiv preprint cs/0003055 (2000).
[10] www.nltk.org
[11] Mandelbaum, Amit, and Adi Shalev. "Word embeddings and their use in sentence classification tasks." *arXiv preprint* arXiv:1610.08229 (2016).
[12] M. M. Haider, M. A. Hossin, H. R. Mahi and H. Arif, "Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm," 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 283-286, doi: 10.1109/TENSYMP50017.2020.9230670.