2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) | 978-1-6654-9400-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAECT57570.2023.10118150

A Novel Approach Using Extractive and Abstractive Summarization for the Genre Classification of Short Text

Chaitanyasuma Jain
Department of Computer Engineering
MKSSS' Cummins College of Engineering for Women, Pune
Pune, India

Rajashree Jain
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed University), Pune
Pune, India

Jatinderkumar R. Saini
Symbiosis Institute of Computer Studies and Research
Symbiosis International (Deemed University), Pune
Pune, India

***Abstract*—Genre classification for text documents is useful in media monitoring and detection of misinformation. Recent work in text classification for genre has shown that advanced algorithms such as neural networks and transformers are well-suited for the purpose. However, for shorter text documents, such as those obtained from social media or news articles, training of deep learning models becomes challenging since they require a large amount of input. Furthermore, genre classification of text summaries, such as headlines of news, is an important direction which has not been explored at large. In this work, the effect of Extractive and Abstractive Summarization on classification for genre of text documents was evaluated. Gensim summarizer was used to obtain extractive summaries and the Pegasus summarizer to obtain abstractive summaries. For classification, two classes of genres, Fiction and Non-fiction, were considered while the gold standard Brown Corpus was used for experimentation. The features used for genre classification were frequencies of various Part-of-Speech (PoS) tags derived from five Penn TreeBank annotated tags. Logistic Regression (LR) and Support Vector Machines (SVM) were used for classification purposes. The results of classification were better for summaries obtained using the extractive technique, indicating that the features of extractive summaries remain in agreement with the documents from which the summary is constructed as compared to abstractive summaries. Further, the SVM classifier performed better than the LR classifier. For exhaustive coverage of the research goal, further experimentation with the number of words of the output summaries of the extractive technique was performed to arrive at a threshold value of the length of summaries. The value indicated that summaries as short as 80 words can be successfully classified using this method.**

Keywords—abstractive summarization; Brown corpus; classification; extractive summarization; Gensim summarizer; Pegasus summarizer

I. INTRODUCTION

Enormous amount of textual information is available today in both online and offline mode. Mining textual information can present several useful insights for processing of natural language. Genre Classification of text is one such technique which refers to the grouping of text into categories, such as Fiction and Non-fiction. The genre of a text can be represented by multiple characteristics of text, such as the frequency of certain words, punctuation, part-of-speech tags, and richness of vocabulary [1]. Advanced algorithms [2] based on Neural Networks and Transformers are known to perform the best for this task. Such algorithms tend to require a large amount of input to train. This becomes a challenge for genre classification of shorter text documents. This is an increasingly relevant research scenario in the day of social media and web-based information, where genre classification can potentially be used for misinformation and fake news detection [3].

Methods are being developed for genre classification of short texts [4], but a similar problem for summaries of documents has not been explored at large. Summaries are representative of the underlying extended document that are intended to capture its essence in brief. Automatic summarization techniques [5] are rapidly being adopted for use in question-answering, search result retrieval and monitoring of media. Understanding the effect of summarization on the genre of a document can be useful in two ways - for effective genre classification of summaries and for effective formation of summaries. Herein lies the primary objective of our work.

In this work, we undertake a case study on the effect of summarization on the genre of a text document. We use the Brown Corpus for experimentation and consider Fiction and Non-fiction as the two genres of text for our purpose. To identify the genre of text and classify unseen data, we use syntactic features of the text, viz. the number of nouns, pronouns, verbs, adverbs and adjectives present in the text. We use two classification techniques, Logistic Regression (LR) and Support Vector Machines (SVM), both of which perform exceedingly well on the underlying extended documents.

We generate Extractive and Abstractive Summaries for the documents using the Gensim Summarizer and Google's Pegasus Summarizer respectively. We classify these summaries in the same manner as the extended documents and observe the variation in results of classification. The LR and SVM classifiers trained on the summaries perform better for Extractive Summaries than for Abstractive Summaries of the same length. Overall, the classifiers perform better for the documents than for the summaries. We also experiment with the number of words in the Extractive summaries to pinpoint a threshold for high-performing classification. We conclude that the genre classification technique is quite accurate for

Extractive summaries of length as small as 80 words, and its performance gradually decreases thereon.

In the first section, we describe the previous work around this problem statement. In the next section, we present the methodology used for this work, including a brief description of the dataset and the data pre-processing performed. In the final section, we describe our findings and conclude the work.

II. Related Work

Textual data contains two primary characteristics, content and style. Identification and grouping of these characteristics is used for analysis of textual data, and retrieval of information thereof [6]. Genre detection is one such information retrieval task which refers to the application of a disambiguation algorithm to find the category of the text in question [6]. Categories of text can be Informative, Imaginative, Fiction, Non-fiction, Narrative, and Poetry amongst others.

Text genre is influenced by external criteria such as the speaker's purpose and topic [7], and other observable properties of text such as linguistic features [8]. Linguistic features of textual data include phonological features, tense markers, questions, word count, punctuation count, and vocabulary richness. Based on these features, methods derived from common word frequencies [1], discriminant analysis techniques [9] and ensemble techniques [10] have been used for text genre classification. Syntactic features of text such as adjectives [11] and Part-of-Speech tags [12] and their ratios [13] are also considered successful indicators of genre. More recent work focuses on advanced algorithms for genre classification. Deep learning methods that use neural networks and transformers are gradually becoming state-of-the-art for classification tasks [2]. Notably, a considerable amount of literature focuses on syntactic features of text as an indication of genre. By using Part-of-Speech tags for text classification, our work corroborates the previous findings.

Genre Classification is a relevant natural language processing problem in that it is frequently incorporated in information retrieval systems. A developing application lies in the detection of misinformation based on the style of writing [3]. The main challenge in combating misinformation is the short length of the underlying text. It is difficult to classify short text documents due to their sparseness and non-standard format, among other reasons [14]. In [15], various challenges and opportunities for short text mining are detailed, including classification of the text. High-performing advanced algorithms are not optimal for text containing a lesser number of words since deep learning methods require a large amount of information for training [4]. In [4], an algorithm based in the semantic space for classification of short text documents significantly outperforms deep learning methods for data consisting of tweets and news.

While much of the work focuses on text documents that are short in length, genre classification for summaries of documents has not been worked upon at large. Representative of the source documents, summaries are expected to be a reliable alternative to the source document capturing the key elements from the source text [16]. Manually writing summaries requires effective construction of sentences such that they represent the text's meaning and inferences precisely; these are very hard tasks [17]. It stands to reason that manual summaries are difficult to write.

Automatic summarization techniques relieve much of this difficulty. There are mainly two kinds of summarization techniques - Extractive and Abstractive. Extractive techniques combine the most important sentences of the underlying document, condensing the size but preserving its context [5]. Abstractive summarization techniques, on the other hand, learn the context and reproduce it differently from the original text [5].

It is clear that Abstractive techniques may make certain concessions on the syntactic, semantic and lexical properties of text. It has been found that one of the limitations of automated summary generation is that it may implicitly assume that the summaries are similar to the source documents even when it is untrue [18]. In this work, we seek to examine such effects of Extractive and Abstractive summarization on the genre of the document.

Most of the previous work in summarization focuses on formulating robust techniques for construction of summaries [19], [20], [21], [22] or on recognizing summaries for filtering and media monitoring [23]. Our work seeks to evaluate the effect of the process of summarization in terms of the genre of the text. We also compare Extractive and Abstractive Summarization techniques using genre classification. By experimenting with the number of words in the output summaries, we wish to determine a threshold beyond which the genre of text may not be preserved during summarization. Through this work, we seek to address the question of whether summaries of documents must preserve all the characteristics of the underlying text, and if yes, why?

III. Methodology

The methodology can be divided into four main steps which we describe in the following subsections, and as depicted by the process diagram in Fig. 1. In the first step, we perform pre-processing on the Brown Corpus which serves as our dataset. Next, we summarize the text documents present in Brown Corpus using two techniques - Extractive and Abstractive Summarization. For Abstractive Summarization, we use the Pegasus summarizer and for Extractive Summarization, we use the Gensim Summarizer. For classification of genre into Fiction or Non-fiction, we create

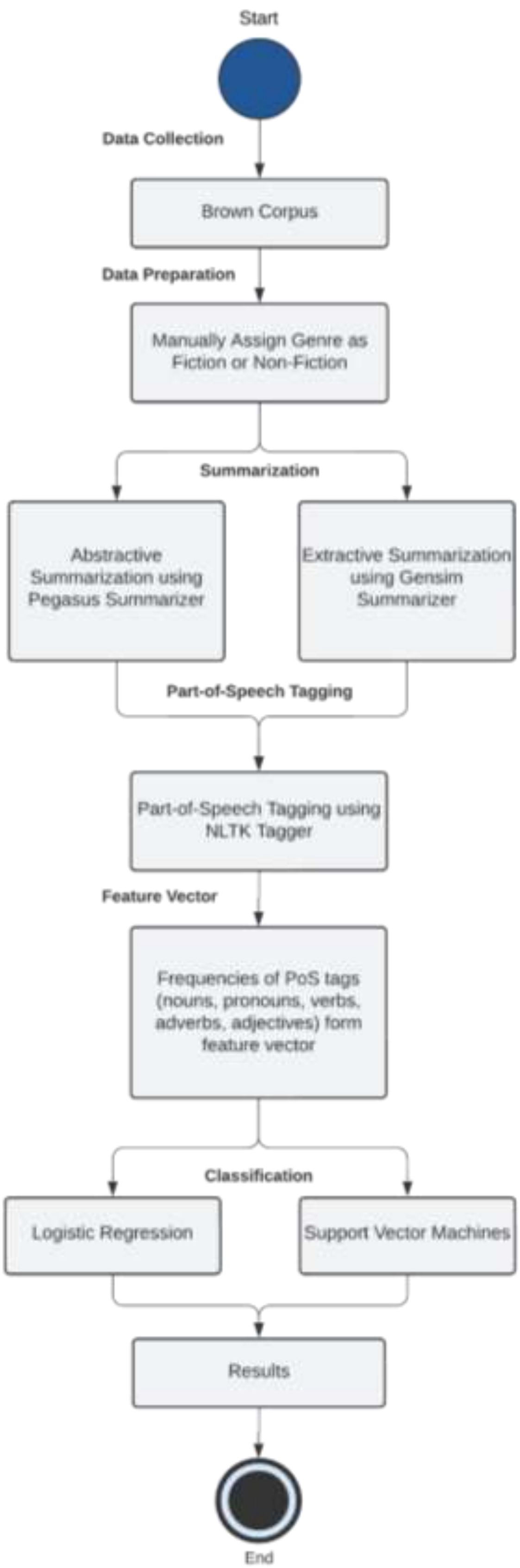


Fig. 1. Diagrammatic representation of methodology

feature vectors by tagging the documents and the summaries using the Natural Language Toolkit (NLTK) Part-of-Speech tagger. We use the frequencies of nouns, pronouns, verbs, adverbs and adjectives obtained from the tags as our features for classification.

A. Data Preparation

In this section, the data used for this work is described in detail. The genre classification of summaries of documents present in the Brown Corpus is performed. The two genres defined for classification are Fiction and Non-fiction, and the summaries are constructed using both Extractive and Abstractive Summarization techniques.

The Brown Corpus is one of the earliest computer-readable corpora of text documents. It consists of 500 text samples. The size of every text document is nearly 2000 words [24]. Each text document contains a `Filename`, the `Text` present in the document and a `Genre`. There are fifteen Genre values in the Brown Corpus.

A set of pre-processing steps on the Brown Corpus are performed prior to the analysis of the data. For the purpose of binary classification into Fiction or Non-fiction genres, a new attribute called 'main_genre' is created, having value as Fiction or Non-fiction. We manually assign the value of 'main_genre' to every document present in the Corpus in the following manner:

- The documents which have Genre values of 'Mystery', 'Adventure', 'Romance', 'Fiction', and 'Science fiction' are assigned the 'main_genre' of Fiction
- The documents which have the values of Genre as 'Learned', 'Government', 'Hobbies', 'News', and 'Reviews' are abstracted to the 'main_genre' of Non-fiction
- We exclude the genres of 'Humour', 'Lore', 'Editorial', 'Religion', 'Letters' and 'belles_lettres' since they are difficult to classify into Fiction or Non-fiction

Other pre-processing steps include cleaning of the text, and dropping rows with null values of `Text` or `genre`. Following the initial processing, the size of the dataset is reduced to 324 text documents. Out of the 324 documents, 207 belong to the Non-fiction genre, and 117 belong to the Fiction genre. We take the class imbalance into consideration while computing the metrics for comparison between genre classification for Extractive and Abstractive Summaries.

B. Summarization

In order to observe the effect of summarization on the genre of the documents, we use two techniques to summarize the documents. There are two types of text summarization techniques - Extractive and Abstractive. Extractive techniques summarize a document by selecting the most sentences and combining them [5]. Abstractive summarization techniques perform a deeper analysis of the underlying text and generate summaries by understanding the concepts and reproducing it in fewer words [5].

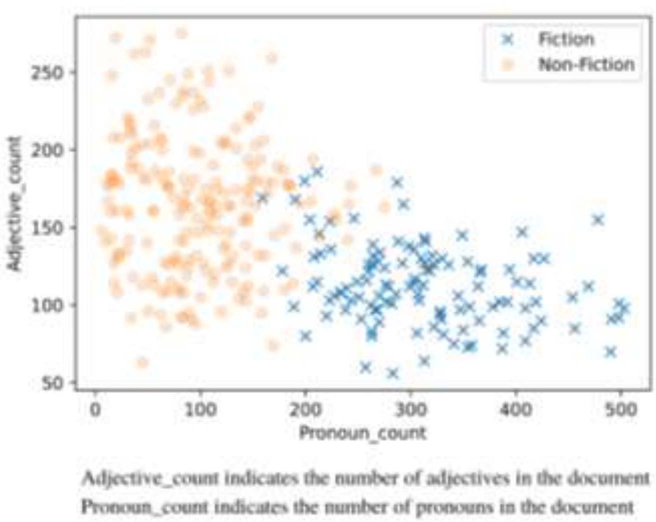
To form summaries using Extractive Summarization, we use the Gensim Summarizer. The summarization.summarizer module is used for this purpose. The algorithm used by the summarizer module is a variation of the TextRank algorithm. In TextRank, the sentences in the text are represented as nodes, and the edges, and their weights, are calculated based on their similarity to other sentences.

The Gensim summarizer uses the BM25 TextRank algorithm which uses a Term Frequency-Inverse Document Frequency (TF-IDF) function to calculate the similarity between sentences probabilistically [27]. The parameter word_count is used to determine the length of the output summaries. We specify different values of word_count for our experimentation purpose, such as 20, 50, 80 and 300 words.

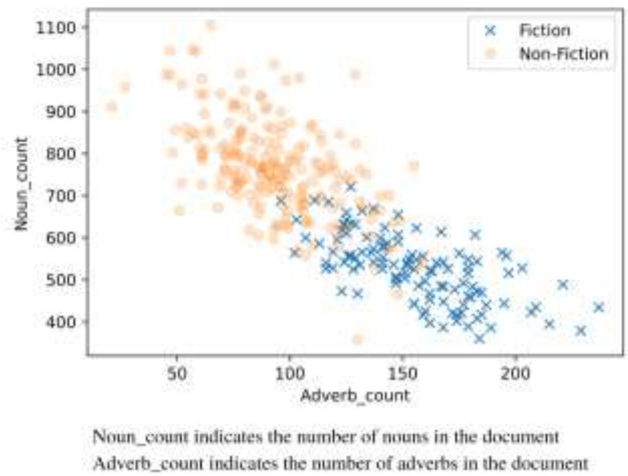
We perform Abstractive Summarization of the text documents using Google's Pegasus Summarizer. It is a transformer-based encoder-decoder model with a self-supervised pre-training phase. In the pre-training phase, the model is input with documents missing important sentences, and is required to recover the missing sentences. Since it is self-supervised, it requires less human input for training and contributes to fine-tuning of the transformer model [28]. The output summaries obtained using this technique have a mean word length of 21.65. The summary lengths lie in the range of 4 words to 62 words, and the median is 20 words.

C. Feature Computation

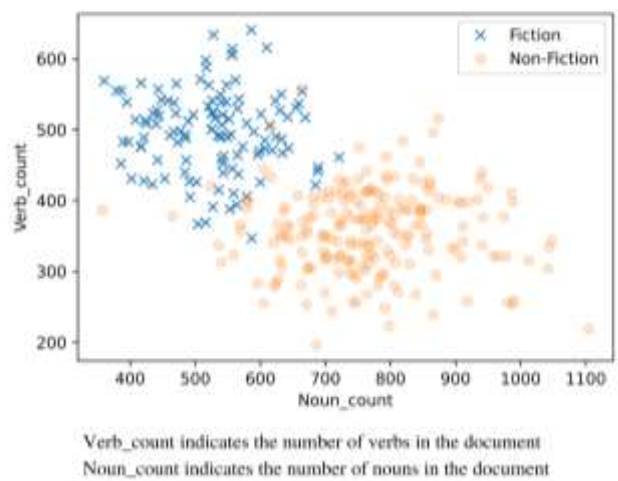
For the binary classification task of categorizing Extractive and Abstractive Summaries as Fiction or Non-fiction, we use five features of the text - the number of nouns, pronouns, verbs, adverbs and adjectives. These are known as Part-of-Speech tags of text. NLTK Tagger [25] is used to parse the text present for every document and tag it. The Penn TreeBank tag-set is used to calculate the number of nouns, pronouns, verbs, adverbs and adjectives using the tagged text. The Penn TreeBank provides a familiar Part-of-Speech for every annotation or tag in the text [26]. We pass every document through a set of rules to associate the corresponding Part-of-Speech to the tag given by NLTK.



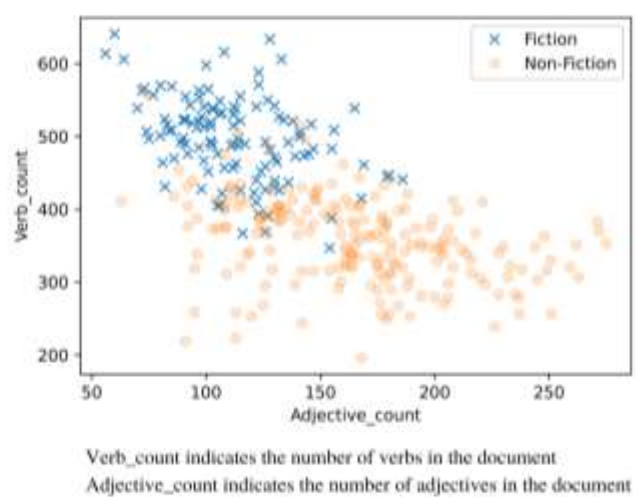
(a) Adjective count vs. Pronoun count



(b) Noun count vs. Adverb count



(c) Verb count vs. Noun count



(d) Verb count vs. Adjective count

Fig. 2. Feature graphs for Fiction and Non-fiction documents

The patterns in parts (a), (b), (c) and (d) of Fig. 2 are obtained when the graphs of PoS tags are generated for the Fiction and Non-fiction documents. The PoS tag categories used for generating the graphs are Adverbs, Adjectives, Pronouns, Nouns, and Verbs. It is clear from the graphs that the PoS tags are a good indicator of genre. Another notable observation regarding writing style is that Fiction documents tend to contain more pronouns, verbs and adverbs than Non-fiction documents.

D. Binary Classification

In this work, we observe the effect of Extractive and Abstractive summarization on the genre of text. The genre of text is identified by training a binary classifier on five features consisting of frequencies of Part-of-Speech tags in the text.

Classification is a supervised technique for categorizing data into two or more classes. There are several methods for classification, out of which we use LR and SVM for this work. In previous work, SVM is known to outperform other supervised classification techniques for textual analysis [29].

As a baseline for comparison, we perform classification on the unaltered Brown Corpus. We use the same parameters for training classifiers for datasets containing summaries, including the train-test split. These summaries are constructed by varying the technique or the number of words of the output summaries.

To compare the performance of classification, four metrics, namely Accuracy (A), Precision (P), Recall (R) and F1-score (F), are considered and analyzed. According to [30], Accuracy is calculated as the ratio of the number of instances correctly evaluated and the number of instances evaluated in total. Precision is calculated as the ratio of the number of true positives and the total number of positive predictions. Recall measures the fraction of positive or true data points that are correctly classified. The F1-score is calculated by taking the Harmonic Mean of Precision and Recall values.

Accuracy, as a metric, is considered to be less informative and more biased. Since we have an imbalanced training dataset, accuracy is not the most appropriate metric for measuring performance. Generally, F1-score is considered to be a better evaluator of the performance of a classifier than its accuracy [30]. The results and conclusions of the experimentation are described in the next section.

IV. RESULTS AND DISCUSSION

The experiments are conducted for the binary classification into Fiction or Non-fiction genres for the summaries of the documents constructed using Extractive and Abstractive Summarization Techniques. A baseline for the comparison is obtained by recording the results of classification on the entire document. Two classification algorithms (represented as 'Algo.' in Table 1, Table 2 and Table 3), viz. LR and SVM are used. The results of classification are shown in Table 1. The results in Table 1, Table 2 and Table 3 are presented in terms of percentage values of the four metrics: Accuracy (A), Precision (P), Recall (R), and F1-score (F). As can be seen, the method used for classification results in high values of Accuracy and F1-score for both classification techniques.

TABLE I. BASELINE SCORES FOR CLASSIFICATION ON 324 DOCUMENTS CONTAINING NEARLY 2000 WORDS

Algo.\Metric	A	P	R	F
LR	98	97	97	97
SVM	97	94	97	95

TABLE II. RESULTS OBTAINED FOR CLASSIFICATION OF 324 SUMMARIES CONSTRUCTED USING ABSTRACTIVE AND EXTRACTIVE TECHNIQUES

Summarization Type	Abstractive Summarization				Extractive Summarization			
Algo.\Metric	A	P	R	F	A	P	R	F
LR	68	53	25	34	76	70	44	54
SVM	71	73	31	44	74	69	51	59

To observe the effect of summarization on the genre, we perform genre classification for Extractive and Abstractive summaries having a median length of 20 words. The Extractive summaries are constructed using the Gensim Summarizer. The Abstractive summaries are constructed using the Pegasus summarizer. The results of genre classification are shown in Table 2.

TABLE III. RESULTS OBTAINED FOR CLASSIFICATION ON 324 SUMMARIES OF VARYING WORD LENGTHS CONSTRUCTED USING EXTRACTIVE TECHNIQUES

No. of words	20 words				50 words			
Algo.\ Metric	A	P	R	F	A	P	R	F
LR	76	70	44	54	77	66	59	62
SVM	74	69	51	59	82	79	66	72
No. of words	80 words				300 words			
Algo.\ Metric	A	P	R	F	A	P	R	F
LR	86	82	72	77	91	87	84	85
SVM	92	91	86	88	94	91	91	91

For experimentation purposes, we also perform genre classification on Extractive summaries of varying output lengths. We capture the classification results of summaries of length 20, 50, 80 and 300 words in Table 3. Fig. 3 shows the trend for Accuracy and F1-score values as the number of words of the output summaries increase.

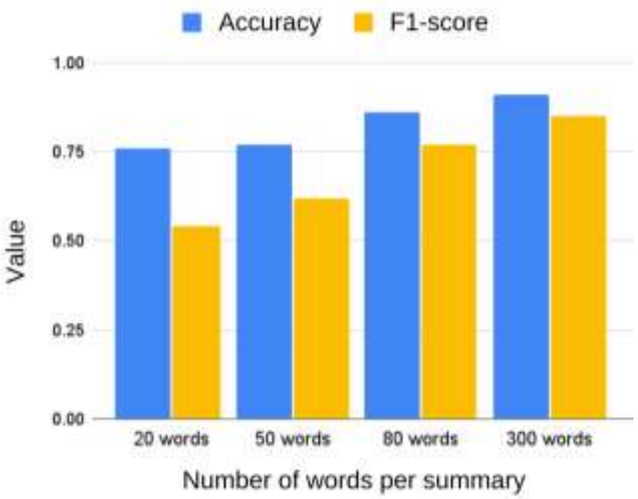


Fig. 3. Accuracy and F1-score values for LR for Extractive Summaries of varying lengths

V. CONCLUSIONS AND FUTURE WORK

In this paper, we performed a case study to observe the effect of Extractive and Abstractive Summarization on the genre of a text document. We categorized the text documents in the Brown Corpus as Fiction or Non-fiction based on syntactic features of the text, namely the number of nouns, pronouns, verbs, adverbs and adjectives. We performed Extractive Summarization on the documents using the Gensim Summarizer, and Abstractive Summarization using the Pegasus Summarizer. For classification, we used the techniques of LR and SVM.

The results show that the classification method using the five syntactic features of text performs well, with high values of Accuracy and F1-score. For the text summaries, classification using SVM performs better than LR for both Extractive and Abstractive. Since our dataset is imbalanced, the Recall values are lower than the Precision values. Overall, the classifiers perform better for Extractive summaries which are constructed by selecting most relevant sentences from the underlying documents, than for Abstractive summaries which are formed by learning the context of the document. Therefore, we can say that the features of the summaries remain in agreement with the text documents in case of Extractive Summarization.

It is important to note that one of the reasons for the low performance of classifiers on the summaries may be that the summaries themselves are not labeled. We used the genre labels of the underlying documents for classifying the genre of the summaries. Another possible reason is the length of the summaries. For experimentation, we also performed genre classification for Extractive summaries of varying lengths. For summaries as small as 80 words, the classification results are optimistic. As the number of words decreases, the performance of the classifiers decreases as well.

This work can be extended to formulate an algorithm for genre classification of short-length summaries. Other features such as punctuation count, feature ratios and vocabulary can be used for classification. Deep learning methods can be used for classification, although their performance may be dependent on the length of the summaries. Other Extractive and Abstractive techniques can be used for further comparison as well.

REFERENCES

[1] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, 'Text genre detection using common word frequencies', in COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics, 2000.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', arXiv preprint arXiv:1810. 04805, 2018.

[3] X. Zhou and R. Zafarani, 'A survey of fake news: Fundamental theories, detection methods, and opportunities', ACM Computing Surveys (CSUR), vol. 53, no. 5, pp. 1–40, 2020.

[4] B. Škrlj, M. Martinc, J. Kralj, N. Lavrač, and S. Pollak, 'tax2vec: Constructing interpretable features from taxonomies for short text classification', Computer Speech & Language, vol. 65, p. 101104, 2021.

[5] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, 'Automatic text summarization: A comprehensive survey', Expert Systems with Applications, vol. 165, p. 113679, 2021.

[6] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, 'Automatic text categorization in terms of genre and author', Computational linguistics, vol. 26, no. 4, pp. 471–495, 2000.

[7] D. Biber and E. Finegan, 'Adverbial stance types in English', Discourse processes, vol. 11, no. 1, pp. 1–34, 1988.

[8] D. Biber, Dimensions of register variation: A cross-linguistic comparison. Cambridge University Press, 1995.

[9] J. Karlgren and D. Cutting, 'Recognizing text genres with simple metrics using discriminant analysis', arXiv preprint cmp-lg/9410008, 1994.

[10] A. Onan, 'An ensemble scheme based on language function analysis and feature engineering for text genre classification', Journal of Information Science, vol. 44, no. 1, pp. 28–47, 2018.

[11] J. Cao and A. C. Fang, 'Investigating variations in adjective use across different text categories', Advances in Computational Linguistics, Journal of Research In Computing Science Vol 41, pp. 207–216, 2009.

[12] S. Feldman, M. A. Marin, M. Ostendorf, and M. R. Gupta, 'Part-of-speech histograms for genre classification of text', in 2009 IEEE international conference on acoustics, speech and signal processing, 2009, pp. 4781–4784.

[13] M. R. Qureshi, S. Ranjan, R. Rajkumar, and K. Shah, 'A simple approach to classify fictional and non-fictional genres', in Proceedings of the Second Workshop on Storytelling, 2019, pp. 81–89.

[14] G. Song, Y. Ye, X. Du, X. Huang, and S. Bie, 'Short text classification: a survey', Journal of multimedia, vol. 9, no. 5, 2014.

[15] M. Grida, H. Soliman, and M. Hassan, 'Short text mining: state of the art and research opportunities', J. Comput. Sci, vol. 15, no. 10, pp. 1450–1460, 2019.

[16] U. Hahn and I. Mani, 'The challenges of automatic summarization', Computer, vol. 33, no. 11, pp. 29–36, 2000.

[17] M. Hill, 'Writing summaries promotes thinking and learning across the curriculum: But why are they so difficult to write?', Journal of reading, vol. 34, no. 7, pp. 536–539, 1991.

[18] T. Copeck and S. Szpakowicz, 'Vocabulary agreement among model summaries and source documents', in ACL Text Summarization Workshop, 2004, vol. 18, p. 46.

[19] Y. Gong and X. Liu, 'Creating generic text summaries', in Proceedings of Sixth International Conference on Document Analysis and Recognition, 2001, pp. 903–907.

[20] R. M. Alguliyev, R. M. Aliguliyev, and N. R. Isazade, 'An unsupervised approach to generating generic summaries of documents', Applied Soft Computing, vol. 34, pp. 236–250, 2015.

[21] S. Narayan, S. B. Cohen, and M. Lapata, 'Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization', arXiv preprint arXiv:1808. 08745, 2018.

[22] K. Chawla, B. V. Srinivasan, and N. Chhaya, 'Generating formality-tuned summaries using input-dependent rewards', in Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), 2019, pp. 833–842.

[23] M. Fisher, D. Albakour, U. Kruschwitz, and M. Martinez, 'Recognising summary articles', in European Conference on Information Retrieval, 2019, pp. 69–85.

[24] W. N. Francis and H. Kucera, 'Brown corpus manual', Letters to the Editor, vol. 5, no. 2, p. 7, 1979.

[25] E. Loper and S. Bird, 'Nltk: The natural language toolkit', arXiv preprint cs/0205028, 2002.

[26] M. P. Marcus, B. Santorini, M. A. Marcinkiewicz, and A. Taylor, 'Treebank-3', Linguistic Data Consortium, Philadelphia, vol. 14, 1999.

[27] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, 'Variations of the similarity function of textrank for automated summarization', arXiv preprint arXiv:1602. 03606, 2016.

[28] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, 'Pegasus: Pre-training with extracted gap-sentences for abstractive summarization', in International Conference on Machine Learning, 2020, pp. 11328–11339.

[29] T. Pranckevičius and V. Marcinkevičius, 'Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification', Baltic Journal of Modern Computing, vol. 5, no. 2, p. 221, 2017.

[30] M. Hossin and M. N. Sulaiman, 'A review on evaluation metrics for data classification evaluations', International journal of data mining & knowledge management process, vol. 5, no. 2, p. 1, 2015.