Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 3
Open Access

Insurance Policy Summarization Using Natural Language Processing (NLP)

Sapana Kolambe[1] and Dr. Parminder Kaur[2]
[1] Research Scholar, MGM University, Chhatrapati Sambhajinagar, Maharashtra, India
[2] Professor, MGM University, Chhatrapati Sambhajinagar, Maharashtra, India
[1]sapanaborole07@gmail.com

Cite this paper as: Sapana Kolambe, Parminder Kaur (2024) Insurance Policy Summarization Using Natural Language Processing (NLP). *Frontiers in Health Informatics, 13* (3), 4774-4782

Abstract. *Policyholders may encounter challenges in understanding important terms and conditions of insurance policies due to their lengthiness and intricacy. This research investigates the usage of Natural Language Processing (NLP) methods for automating summaries of insurance policy documents. Our approach aims at producing concise and accurate summaries which highlights important aspects such as cover details, exclusions and premiums through a combination of extraction-based and abstractive summarization strategies. We elaborate on pre- processing procedures that consist of tokenization, part-of-speech tagging, text cleaning among others; moreover, we explain how modern deep learning models like transformers are employed for sentence creation and context understanding. To ensure that the summarized results are accurate and consistent, human evalu- ations were conducted in addition to standard metrics tests on the proposed sys- tem. Our findings demonstrate that NLP is able to enhance the understandability of insurance products thus facilitating better decision-making processes leading to higher levels of customer satisfaction.*

Keywords: *Insurance Policy Summarization, Natural Language Processing, In- surance Document Analysis.*

1 Introduction

Insurance policies are basic documents containing the provisions of insurance con- tracts, such as coverage limits, exclusions, premiums and claims procedures. Unfortu- nately, these papers tend to be long, intricate and full of legalese that may be difficult for policyholders to fully understand. Misunderstandings about coverage could result from this complexity leading to out-of-pocket expenses that people never anticipated; this results in customer dissatisfaction.

This has therefore called for tools that can simplify and make plain language insurance policy documents more understandable. Natural Language Processing (NLP), a sub- field of artificial intelligence (AI) dealing with how computers interact with human language shows promise in automating summaries of text. The use of NLP techniques in generating insurance summaries gives them a chance to understand their policies in short but precise terms.

This paper examines NLP techniques to summarize insurance policy documents. We investigate two types of summarizations namely extraction-based and abstractive meth- ods. For example, in extraction-based summarization the main task is to identify key sentences or phrases that are direct from the original text. This procedure mainly in- volves scoring sentences using their relevance and importance hence it uses metrics such

as statistical and linguistic features for determining the most critical areas in a given text. On the other hand, abstractive summarization creates new sentences that can convey what was in the original document but using fewer words and being more co- herent at the same time. Accordingly, deep learning models used in this type of sum- marizing have an understanding of context as well as meaning thereby making them sound more natural and human-like.

Our approach comprises several important stages:

- Text Preprocessing: Here, we remove redundant elements from texts such as stop words, punctuation marks, alignment features etc. while tokenization splits text into either individual words or phrases which are convenient when doing word analysis on a sentence, however part-of-speech tagging identifies and labels parts of speech such that they can establish structure and context for any given text
- Keyword Extraction and Sentence Scoring: Key terms and phrases that are crucial to the policy's meaning are identified. Sentences are scored based on the presence of important keywords and their overall relevance to the document's main themes.
- Application of Deep Learning Models: Advanced models such as transformers (e.g., BERT, GPT) are employed to grasp the context and meaning beyond indi- vidual sentences, allowing for the generation of concise and coherent summaries.
- Customization for Insurance Policies: Domain-specific models are trained specif- ically on insurance-related texts to improve accuracy and relevance. Legal and regulatory compliance is incorporated to ensure summaries accurately reflect pol- icy terms and avoid misinterpretation.
- Evaluation and Refinement: The performance of the summarization system is evaluated using standard metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and human reviews to ensure the quality and usability of the generated summaries.

The goal of this research is to demonstrate that NLP can be effectively utilized to make insurance policies more accessible and comprehensible for policyholders. By im- proving the clarity of these documents, we aim to enhance customer satisfaction and empower individuals to make better-informed decisions about their insurance coverage. This work contributes to the broader field of document summarization and has practical implications for improving the transparency and user-friendliness of insurance prod- ucts. Furthermore, it opens avenues for future research in applying NLP to other types of legal and regulatory documents, potentially transforming how individuals interact with and understand complex written information.

2 Literature Survey

Qaroush et al. (2019) proposes the summarization of the Arabic single document which produces a fairly informative extractive summary combining machine learning and score-based approaches that evaluate each sentence based on a combination of se- mantics and statistics. The results are superior in terms of precision metrics, memory, and F-scores, the disadvantage is that it has not optimized the weight of the features. Verma and Om (2019) minimized redundancy in multi-document summarizing by the Shark Smell Optimization (SSO) method and the performance results were far better than the previous summary method.

Furthermore, there are extractive summarizing studies using neural networks, which in recent years have achieved greater popularity than conventional approaches, some of these studies are (Mohsen et al., 2020; Anand and Wagh, 2019; Xu and Durrett, 2019; Chen et al., 2018a,b; Alami et al., 2019). Anand and Wagh (2019) conducted research by using a deep learning technique specifically Feed Forward Neural Network (FFNN) to summarize a single document in a authorized document that has the ad- vantage of producing an extractive summary without the need to create features or do- main knowledge and perform well as measured by the Rouge score and produces a coherent summary, will but weedy in terms of simplifying complex and extensive sen- tences.

Additionally, research on abstractive summarizing has been stimulated by the en- coder-decoder framework, as cutting-edge research conducted by Xu et al. (2020); Lee et al. (2020); Yao et al. (2018a); Iwasaki et al. (2019). In addition, existence thought that this model is smoother, the encoder-decoder framework is also convenient in ad- justing parameters automatically (Xu et al., 2020). Rank-biased precision-summariza- tion (RBP-SUM) by RodríguezVidal et al. (2019) which has advantages in overcoming redundancy by evaluating using rouge, but this method can only produce extractive summaries.

Text summarization is a formidable challenge in the field of Natural Language Pro- cessing (NLP) (Rane and Govilkar, 2019; Shabbir Moiyadi et al., 2016) because it re- quires precise text analysis such as semantic analysis and lexical analysis to produce a good summary. A good summary, in addition, must contain important information and must be concise but also must consider aspects such as non-redundancy, relevance, coverage, coherence, and readability (Verma et al., 2019). Where to get all these aspects in a summary is a great challenge. The review of papers on text summarization is im- portant because summarizing extractive techniques has become a very broad research topic and is heading towards maturity (Gupta and Gupta, 2019). Now research has shifted towards abstractive summarization (Gupta and Gupta, 2019) and real-time sum- marization. This is because abstractive summaries are more complex and complicated than extractive summaries.

Table 1: Abstractive Text Summarization methods

Methods	Description	Advantages	Limitation
Tree Based Method	It uses a dependency tree to represent the text which uses either a language generator or an algorithm for sum- mary generation.	It walks on units of the given document read and easy to sum- mary.	It lacks a com- plete model which would include an ab- stract representation for content selection.
Template Based Method	Linguistic patterns or extraction rules are matched to identify text snippets that will be mapped into template slots.	It generates sum- mary is highly co- herent relies on relevant infor- mation identified by IE system.	Requires designing of templates and generalization of template is too diffi- cult.
Ontology Based Method	Use ontology to im- prove the process of summarization. It ex- ploits fuzzy ontology to handle uncertain data that simple do- main ontology can- not.	Drawing rela- tion or context is easy due to on- tology Handles uncer- tainty at reason- able amount	This approach is lim- ited to Chinese news only. Creating Rule based system for handling uncertainty is a complex task.
Lead and Body Phrase Method	This method is based on the operations of phrases that have same syntactic head chunk in the lead and body sen- tences	Good for semanti- cally appropriate revisions for re- vising a lead sen- tence.	Parsing errors degrade sentential complete- ness such as grammat- icality and repetition. It focuses on rewriting techniques

Multi-modal semantic model	A semantic model which captures concepts and relationship among concepts is built to represent the contents of multimodal documents.	It produces abstract summary whose coverage includes salient textual and graphical content.	The limitation of this framework is that it is manually evaluated by humans.
Information Item Based Method	The contents of summary are generated from abstract representation of source documents, which is the smallest element of coherent information in a text.	The major strength of this approach is that it produces short, coherent, information rich and less redundant summary.	It rejected due to the difficulty of creating meaningful and grammatical sentences from them.
Semantic Graph Based Method	-This method is used to summarize a document by creating a semantic	It produces concise, coherent and less redun- dant and	This method is limited to single document abstractive summarization.
	graph called Rich Semantic Graph (RSG)	grammatically correct sen- tences.	

Summary of the text can be generated by two techniques which are Extractive Technique and Abstractive Technique. In abstractive text summarization, the concept of the original text is understood and it is retold changing the meaning of the original text. In Extractive summarization, the summary generated by the extractive technique contains those sentences which have highest scores among all the sentences means important sentences from the original document exactly as they appear in it which is decided on the basis of statistical and linguistic features of sentences in the document. This literature survey highlights the significant progress made in text summarization and the application of NLP in the insurance domain. However, the specific task of summarizing insurance policies presents unique challenges that require specialized approaches. Future research should focus on developing and fine-tuning models tailored to the complexities of insurance documents, ensuring that the generated summaries are both accurate and user-friendly. By addressing these challenges, NLP can play a crucial role in making insurance policies more accessible and comprehensible, ultimately benefiting both policyholders and insurers.

3 Text Summarization

The field of Natural Language Processing (NLP) has seen significant advancements in recent years, leading to various applications in text summarization. This literature sur- vey reviews key research efforts and methodologies relevant to summarizing insurance policies using NLP. Text summarization can be broadly categorized into extraction- based and abstractive methods. Extraction-based summarization selects key sentences or phrases directly from the source text, while abstractive summarization involves gen- erating new sentences that convey the core information.

3.1 Techniques:

Extraction-Based Summarization:
Early methods for extraction-based summarization focused on statistical techniques, such as TF-IDF (Term Frequency-Inverse Document Frequency) and lexical chains. More recent approaches utilize machine learning algorithms to rank sentences based on features like sentence position, length, and the presence of named entities.
Techniques such as TextRank, which applies the PageRank algorithm to identify the most important sentences in a text, have also been widely adopted.

Abstractive Summarization:
Abstractive summarization models often rely on sequence-to-sequence (Seq2Seq) ar- chitectures using recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and more recently, transformers.

The introduction of transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), has significantly improved the performance of abstractive summarization by enabling better context understanding and generation of coherent summaries.
Natural Language Processing solves the most important ones are:

- Machine translation is the first classic task assigned to the developers of NLP- technologies
- grammar and spell checking – as a conclusion of the first task;
- text classification – definitions of text semantics for further processing
- named-entity recognition (NER) – definition and selection of entities with a predefined meaning (used to filter text information and understand general se- mantics);
- summarization – the text generalization to a simplified version
- text generation – one of the tasks that are used to build AI-systems
- topic modeling – technique for extracting hidden topics from large text vol- umes.

Natural Language Processing use various metrics such as Edit distance, Cosine simi- larity, Vectorization, Bag of words, TF-IDF, Text normalization, Stemming and Lem- matization, Naive Bayes algorithm, Naive Bayes algorithm, long short-term memory (LSTM).

3.2 NLP Applications in the Insurance Domain:

The application of NLP in the insurance sector has focused on various tasks, including information extraction, claims processing, and document classification. However, sum- marization specifically for insurance policies remains a relatively underexplored area. **Information Extraction:** Techniques for extracting structured information from un- structured insurance documents have been developed using named entity recognition (NER) and template-based methods. Machine learning models have been employed to identify and extract relevant entities such as policyholder names, coverage amounts, and dates.
Claims Processing: NLP has been used to automate parts of the claims processing workflow, including the extraction of claim details and the categorization of claims based on their descriptions. Rule-based and machine learning approaches have been combined to enhance the accuracy of automated claims assessment.
Document Classification: Classification models have been developed to categorize insurance documents into predefined classes, such as policy types or claim status, using features extracted from text.

3.3 Challenges in Insurance Policy Summarization

Summarizing insurance policies poses unique challenges due to the legal and technical language used in these documents. Ensuring that the summaries are accurate, legally compliant, and comprehensible to non-experts is critical.
Legal and Technical Language: Insurance policies are written in a formal, legalistic style, which can be

difficult for summarization models to interpret accurately.

Abstractive summarization, while powerful, must be carefully designed to avoid mis- interpretation or omission of critical policy details.

Context and Coherence: Maintaining the context and coherence of the original docu- ment in the summary is a significant challenge, especially for long and complex texts. Advanced models like transformers have shown promise in addressing these issues by capturing long-range dependencies in text.

4 Methodology

We have used NLP, which seeks to summarize articles by picking a collection of words that hold the most essential information, can address this problem with the help of ex- tractive summarizer. This approach takes a significant portion of a phrase and utilizes it to create a summary. To define sentence verbs and subsequently rank them in terms of significance and similarity, a variety of algorithms and approaches are utilized.

There is a great need for text summary techniques to address the amount of text data available online to help people find the right information and use the right information quickly. In addition, the implementation of text summaries reduces reading time, speeds up the process of researching information, and increases the information that may not be in one field. This research paper focuses on the frequency-based approach for text summarization. The steps involved in text summarizer are Sentence and word tokeni- zation and then calculating sentence score on the basis of TF-IDF score which is being used to select the most important sentences to retain the information and merge it to form a summary.

STEP-1: Import all necessary libraries: NLTK (Natural Language toolkit) is a widely used library while we are working with text in python. Stop words contain a list of English stop words, which need to be removed during the pre-processing step.

STEP-2: Generate clean sentences Text processing is the most important step in achiev- ing a constant and positive approach result. The processing steps removes special digits, word, and characters.

STEP-3: Calculate TF-IDF and generate a matrix We'll find the TF and IDF for each word in a paragraph.

TF (t) = (Frequency of t from document) / (total_no.of t in the document) IDF (t) = log_e (total_no. Of documents / No. of documents with t it) [4]

Now, we will be generating a new matrix after multiplying the calculated TF and IDF values.

STEP-4: Score the sentences Here, we use TF-IDF word points in a sentence to give weight to a paragraph. However, Sentence scoring varies with different algorithms.

STEP-5: Generate the summary This is the last stage of text summarization. Top sen- tences are calculated based on the score and retention rate given to the user are included in the summary and finally, a summary is created.

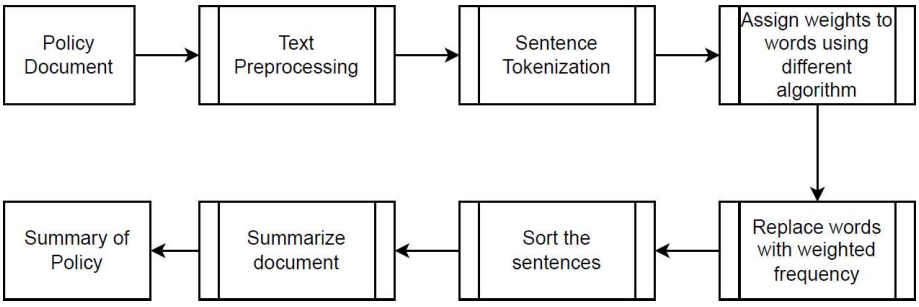


Fig. 2. System diagram of text summarization

5 Observations and Results

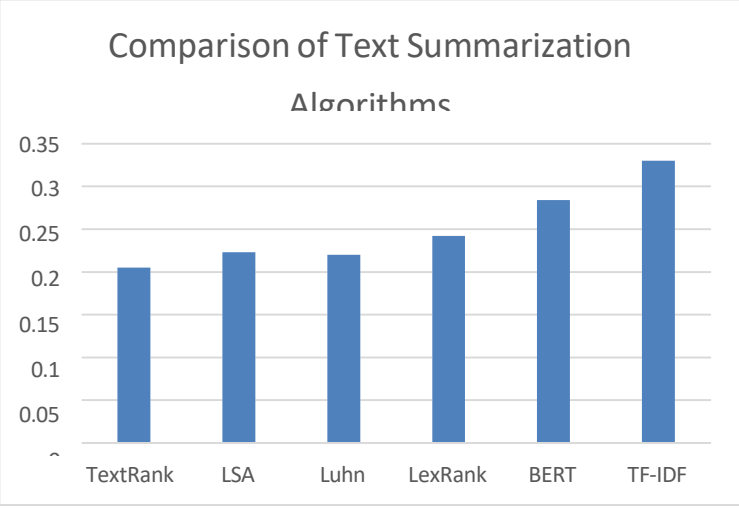
Insurance policy dataset is used to verify methodology. This dataset contains 152 health insurance policy documents. The assessment of text summarizing techniques has ex- perienced a dynamic transition from conventional metrics to metrics that are focused on semantics and human evaluations. Every stage of its development has unique traits, each characterized by advantages and disadvantages related to the techniques used. The basis for evaluation was established by traditional metrics, which used quantitative measurements to gauge how well summarizing techniques worked. Semantics-focused metrics arose as the area developed, demonstrating a greater comprehension of the com- plex interactions between meaning in summaries. Furthermore, human evaluations added the crucial component of subjective judgment, bringing evaluation procedures closer to the viewpoints of end users. Every assessment method offers a different per- spective in this complex environment, which emphasizes the necessity for a thorough and sophisticated method to evaluate the effectiveness of text summarizing strategies. Recent advancements in deep learning and NLP offer promising solutions for improv- ing the summarization of insurance policies. Key areas of ongoing research include: *Fine-Tuning Pre-trained Models:* Fine-tuning pre-trained transformer models on do- main-specific corpora can enhance their performance on insurance texts. Domain-spe- cific embeddings and transfer learning techniques are being explored to adapt general- purpose NLP models to specialized fields like insurance.

Hybrid Approaches: Combining extraction-based and abstractive methods can leverage the strengths of both approaches to produce more accurate and readable summaries.

Hybrid models that integrate rule-based components with machine learning techniques are also being developed to ensure compliance with legal standards.

Evaluation and Validation: Developing robust evaluation frameworks that include both automated metrics (e.g., ROUGE) and human assessment is crucial for validating the effectiveness of summarization models. Collaborations with domain experts are essen- tial for refining models and ensuring that summaries meet practical needs and regula- tory requirements.

We have compared our system with TextRank, Yake, KeyBert, LexRank, Luhn, LSA, BERT.



Conclusion

This paper introduces a method for summarization of insurance documents. The method implements the TF-

IDF technique with optimization for insurance related terms. The system is fast and modest baselines. The evaluation results show that eval- uating multi-word terms vs single-word ones improves the quality of the summaries and that extracting continuous sequence from the document provides the results. Future work may include modifying the current method to extract the most important sentences instead of extracting the whole sequence. In addition, combining the multi-term TF- IDF weighting scheme with machine learning algorithms may provide interesting re- sults.

References

1. Qaroush, A., Abu Farha, I., Ghanem, W., Washaha, M., Maali, E., 2019. An efficient single document Arabic text summarization using a combination of statistical and semantic fea- tures. J. King Saud Univ. - Comput. Inf. Sci. https://doi.org/ 10.1016/j.jksuci.2019.03.010.
2. Verma, P., Om, H., 2019. MCRMR: Maximum coverage and relevancy with minimal re- dundancy based multi-document summarization. Expert Syst. Appl. 120, 43– 56. https://doi.org/10.1016/j.eswa.2018.11.022.
3. Mohsen, F., Wang, J., Al-Sabahi, K., 2020. A hierarchical self-attentive neural extractive summarizer via reinforcement learning (HSASRL). Appl. Intell. 1–14. https://doi.org/10.1007/s10489-020-01669-5.
4. Anand, D., Wagh, R., 2019. Effective deep learning approaches for summarization of legal texts. J. King Saud Univ. - Comput. Inf. Sci. https://doi.org/10.1016/j. jksuci.2019.11.015.
5. Xu, J., Durrett, G., 2019. Neural Extractive Text Summarization with Syntactic Compres- sion, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3290–3301. https://doi.org/ 10.18653/v1/d19-1324
6. Alami, N., Meknassi, M., En-nahnahi, N., 2019. Enhancing unsupervised neural networks based text summarization with word emb e dding and ensemble learning. Expert Syst. Appl. 123, 195–211.
7. Xu, W., Li, C., Lee, M., Zhang, C., 2020. Multi-task learning for abstractive text summari- zation with key information guide network. EURASIP J. Adv. Signal Process. 2020.
8. Lee, H., Choi, Y., Lee, J.H., 2020. Attention history-based attention for abstractive text sum- marization, in: Proceedings of the ACM Symposium on Applied Computing. pp. 1075– 1081.
9. Iwasaki, Y., Yamashita, A., Konno, Y., Matsubayashi, K., 2019. Japanese abstractive text summarization using BERT, in: Proceedings - 2019 International Conference on Technolo- gies and Applications of Artificial Intelligence, TAAI 2019.
10. Rodríguez-Vidal, J., Jorge, C.-D.-A., Amigó, E., Plaza, L., Gonzalo, J., 2019. Automatic generation of Entity - oriented Summaries for Reputation Management. J. Ambient Intell. Humaniz. Comput.
11. Rane, N., Govilkar, S., 2019. Recent trends in deep learning based abstractive text summa- rization. Int. J. Recent Technol. Eng. 8, 3108–3115
12. Verma, P., Pal, S., Om, H., 2019. A comparative analysis on Hindi and English extractive text summarization. ACM Trans Asian Low-Resource Lang. Inf. Process. 18, 30–39.
13. S. Pattnaik and A. K. Nayak, "Summarization of Odia Text Document Using Cosine Simi- larity and Clustering," 2019 International Conference on Applied Machine Learning (ICAML), pp. 143-146, 2019.
14. Munot, Nikita & Govilkar, Sharvari, Comparative Study of Text Summarization Methods. International Journal of Computer Applications, 2014
15. Kotadiya, R., Bhatt, S., Chauhan, U. (2020). Advancement of Text Summarization Using Machine Learning and Deep Learning: A Review. In: Singh, P., Pawłowski, W., Tanwar, S., Kumar, N., Rodrigues, J., Obaidat, M. (eds) Proceedings of First International Confer- ence on Computing, Communications, and Cyber-Security (IC4S 2019). Lecture Notes in Networks and Systems, vol 121. Springer, Singapore.

16. Ramón-Hernández, A.; Simón-Cuevas, A.; Lorenzo, M.M.G.; Arco, L.; Serrano-Guerrero, J. Towards Context-Aware Opinion Summarization for Monitoring Social Impact of News. Information 2020, 11, 535.
17. Mallick, Chirantana & Das, Ajit & Dutta, Madhurima & Das, Asit & Sarkar, Apurba. (2018). Graph-Based Text Summarization Using Modified TextRank. Advances in Intelligent Sys- tems and Computing.
18. Yadav, Anurag & Kumar, Mukesh & Pathre, Ayonija. (2020). Implemented Text Rank based Automatic Text Summarization using Keyword Extraction. International Research Journal of Innovations in Engineering and Technology. 04. 20-25.