Dijlah Journal of Engineering Science (DJES) Vol. 3, No. 3, August, 2026, pp. 171-182
ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 195

A BERT-Based Approach for Hedges Detection in Scientific Texts with Cue-Based Labeling

Omar Hamid Flayyih

University Of Samarra, College Of Applied Sciences, Department Of Computer Science, Samarra, Iraq

Article Info

Article history:

Received Apr. 02,2026
Revised May,30,2026
Accepted Aug.,10,2026

Keywords:

Natural Language Processing (NLP)
Text Classification
Hedge Detection
Scientific Text Analysis
Deep Learning
BERT

ABSTRACT

The use of hedging phrases is vital in scientific papers because it allows expressing uncertainty and reducing the certainty of claims. An accurate recognition of such phrases is of great importance for effective data retrieval and decision making in specific areas such as biomedical science. In this work, we develop a BERT-based framework for automated detection of hedging phrases in scientific text. For this purpose, a custom dataset based on XML documents of scientific publications is compiled and annotated automatically on the basis of hedge phrases. With the use of the deep contextual embeddings, the model has been able to understand the features of uncertainty both semantically and syntactically. The experiments conducted have shown promising results with an accuracy of about 96.9% and an F1-score of 94.7%. This clearly shows the strength of the model. The experiments conducted prove the efficiency of the proposed model in dealing with difficult phenomena such as hedging. Future research will focus on using the proposed model in a multitasking framework which will include other components such as understanding context, classifying intents and analyzing attitudes.

Corresponding Author:

Omar Hamid Flayyih
University Of Samarra, College Of Applied Sciences, Department Of Computer Science, Samarra, Iraq
Email: omarhamid@uosamarra.edu.iq

1. INTRODUCTION

Language is regarded as the main means of communication among people, serving an essential function in comprehending and analyzing the surrounding environment [1]. An important component that facilitates successful communication is language competence. This notion was first put forward by Chomsky, who made the distinction between competence, understood as the knowledge of language, and performance, which refers to the application of language in real-life situations. Under this classification, pragmatic competence refers to the capability to utilize language in a proper manner in varying contexts, involving vague and imprecise utterances for different communicative intentions like mitigation, politeness, and tentativeness.

One of the most significant tools for conveying vagueness is hedges. Hedges are frequently examined in pragmatics and discourse analysis since they serve as means of making the utterance less assertive, less forceful, and uncertain. The notion of hedges had gone through development during the past few decades. The earlier ideas concerning this linguistic phenomenon could be seen in fuzzy logic theories developed by Zadeh [2] according to which many objects and concepts in reality have no definite borders. Afterward, the idea of linguistic hedges was further

advanced by Lakoff who described them as “words or expressions that made the meaning fuzzier or less fuzzy” [3]. Hedges can thus be said to be language constructs that indicate doubt or partial credibility about the information being relayed. Such constructs are extensively used in political language and strategic analysis, especially in diplomatic discourse, in presenting hypotheses, interpretations, or provisional explanations of international affairs and policy alterations. From statistical studies carried out on corpora of political speeches, it is evident that a sizable proportion of assertions made in official statements and policy documents feature speculative elements. Distinguishing between actual policy and speculation in political prediction has hence become vital in the process of political information extraction. In tackling this problem, scholars concentrate on identifying uncertainties in political texts, which include identifying hedge markers and assessing their scopes in political statements. Let us examine an example in diplomatic reports when a strategist says, “The recent legislative changes might imply a move towards isolationism in foreign policy.” Although the assertion makes a forecast, the usage of the word “might” shows uncertainty instead of an executive order.

In this case, within the domain of biomedical information extraction, the ability to distinguish between reliable and unreliable pieces of information has taken on new importance. As an answer to this problem, the CoNLL-2010 Shared Task [4], aimed at uncertainty detection, has been created. The Shared Task consists of two main subtasks; the first one involves identification of hedging phrases, whereas the second is devoted to their scope within the sentence. Take for example a situation from a doctor’s report: “It might seem like that the symptoms point to a certain neurological illness”. In spite of a possible suggestion about the patient’s disease, the word “might” show that it is not certain and should not be interpreted as a definite diagnosis.

The identification of uncertainties is thus an important task, not only in the medical domain but also in other critical domains like finance, engineering, and safety-related systems. In these areas, differentiating between the facts verified and the speculation is of great significance. The expressions used to convey uncertainty include modal verbs (can, might, etc.), vague quantifiers (some, many, etc.), and subjective expressions (it is likely, it seems that, etc.). These expressions are, however, very context-based, and their occurrence does not necessarily mean that there is uncertainty. This makes the detection of hedges in automatic ways a very difficult task. Additionally, according to Prince et al. [5], linguistic uncertainties can occur either due to the information provided or the attitude of the person providing the information.

However, due to the nuances in the language that can be found in the scientific literature, the automatic identification of hedges remains a challenging problem despite all the progress made in NLP research. Current approaches to this problem often experience difficulties with recognizing whether the word used functions as a hedge or a factual expression in complex sentences. In addition, classical methods for tackling this issue do not generalize well across various scientific domains.

To overcome the constraints associated with this approach, the current research presents an advanced architecture based on BERT, aiming to improve the accuracy and resilience of the method used. The key value of this research is represented by the fine-tuning of the architecture used for generating contextual representations of uncertainty, something not achieved by other existing methods. By applying an efficient binary classifier on top of the domain-based transformer model, the authors seek to outperform other existing solutions.

2. RELATED WORK

The initial work on hedging identification mainly focused on straightforward rule-based approaches, where sentences were considered hedged if they contained specific cue words like may, suggest, and likely [6]. Such rules were subsequently generalized via weak supervision methods, in which seed words were extended via co-occurrence modeling with SVMs and MaxEnt models [7], [8]. With the development of machine learning technologies, the hedge detection task became a classification problem. Many studies applied standard classification algorithms like SVM, k-NN, and DT to gain state-of-the-art performance in various datasets. In particular, the results generated from SVM-based approaches were superior and used extensively in early hedge detectors. Morante and Daelemans

[9] obtained an F1-score of 84.7% using k-NN with lexico-syntactic information such as lemma, POS tag, and neighbor tokens. Another turning point in hedging research was the advent of the CoNLL-2010 Shared Task [10], which offered a benchmark platform for identifying hedge indicators and their scopes. Most winning systems used sequence labeling schemes, such as CRF or SVM-HMM, to address this challenge.

Such approaches made use of a variety of linguistic information, including tokens, POS tags, chunking data, and syntax [11]. For example, previous research showed that traditional machine learning algorithms could be employed, since CRF models could provide competitive results. Further enhancement of the approach could be done via feature selection and the incorporation of other linguistic sources such as WordNet. However, the direction was then moved toward the employment of deep learning methods in order to solve the problems associated with manual feature engineering. Specifically, CNN models for ambiguity and hedging detection were found to have better performance than other types of machine learning approaches such as SVM [12], [13]. Additionally, the implementation of attention made the model able to pay more attention to the context of the word, which resulted in even greater performance [14].

More recently, models based on transformers have contributed significantly to enhancing the performance of the state-of-the-art systems for various natural language processing applications, including hedging. The pre-trained language models like BERT have been found to exhibit high levels of proficiency in understanding the context and semantics of the text. Research has revealed that fine-tuning the BERT model outperforms traditional methods such as Tree-LSTM and CNN-based models, especially when applied to datasets like BioScope [16].

Other than cue detection, the detection of hedge scope has received much attention and can be classified into rule-based and machine learning-based techniques. The rule-based technique uses manually or automatically generated syntactic rules to detect the scopes, which yield high accuracy in specialized domains but fail to generalize [17]–[20]. On the other hand, the machine learning-based technique views the hedge scope detection task as token classification, in which each token is tagged with a label indicating its function in the scope [21]. Although this technique exploits contextual information, it suffers from class imbalance and many negative examples that adversely affect its performance. To overcome these problems, recent research works have considered other approaches such as phrase-level classification and shallow semantic parsing [22], [23], [24]. But there remain some issues related to the proper description of such dependencies among tokens that are close and similar linguistically.

Therefore, even considering considerable progress that has been achieved, it cannot be claimed that detecting hedges is an easy task, especially because it depends on context. This is why transformer models such as BERT can be used for that purpose.

3. METHODOLOGY

Detecting hedging is a key issue in natural language processing especially in the domain of science and medicine because writers often indicate their uncertainty, speculation, and conjectures within their texts. Detecting this kind of hedging is important for the development of information extraction tools and knowledge representation systems. In our research paper, we propose using BERT for this problem, along with an automatic labeling method using cues based on an annotated XML corpus.

3.1 Dataset Description

The dataset for this study was extracted from the BioScope Corpus, which is a well-known benchmark corpus for the task of negation and speculation detection in biomedical literature. The BioScope corpus comprises manually annotated sentences regarding hedges and scopes in various fields, such as:

- Biomedical abstracts (from the GENIA corpus)
- Full scientific articles

- Clinical free-text records

The corpus was originally introduced by Veronika Vincze et al. (2008) [25] and has been extensively used in the CoNLL-2010 Shared Task for hedge detection and scope resolution. One of the primary advantages of the current dataset lies in its excellent annotation quality, obtained thanks to dual annotation and validation by experts, where inter-annotator agreement was calculated via F-measure at different levels (cue detection and scope identification). Regarding the size of the dataset, the abstracts subset includes 11,871 sentences, whereas the scientific articles subset has 2,670 sentences. Combining the two subsets results in a dataset of 14,541 sentences in total. Out of those, 10,386 sentences were annotated as non-hedge sentences, while 4,155 sentences were annotated as hedge sentences, representing around 71% of non-hedge sentences and 29% of hedge sentences. Moreover, the distribution of the sentences' lengths for both hedge and non-hedge sentences is represented on Figure 1. As can be seen from the chart, most sentences have a relatively moderate length, with both classes sharing a similar distribution of length. On the other hand, hedge sentences demonstrate some variability compared to the other class, especially concerning long sentences, which should be considered when choosing the maximum sentence length for the BERT model.

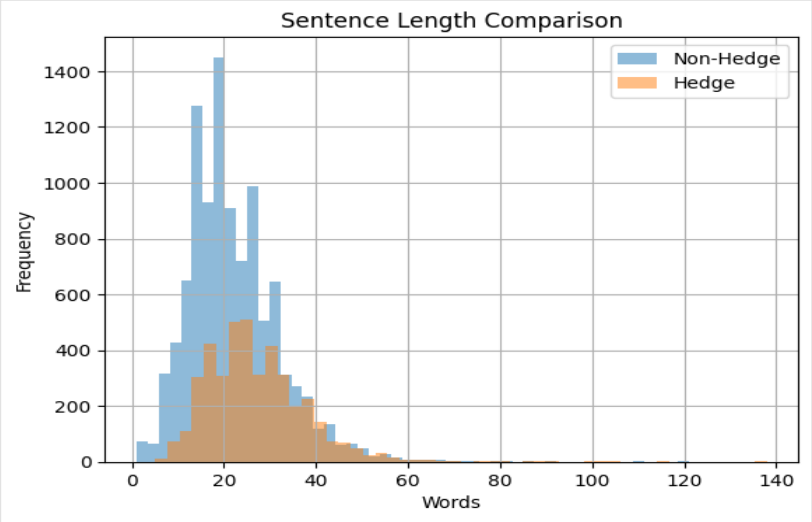

Figure 1. Distribution of sentence lengths for hedge and non-hedge classes

3.2 Data Processing and Splitting

For data quality enhancement and to improve the performance of the models built upon them, the following preprocessing was conducted on the extracted sentences. Firstly, the text of the sentences was lowered for uniformity and to keep the vocabulary relatively small. The next step involved removing newlines from the data, which were then substituted by whitespaces. After that, the data were cleaned by removing all the non-alphanumeric characters in the sentences, thus keeping only the meaningful information. All empty sentences produced in this way were deleted. Besides, duplicate sentences were removed from the data set to prevent duplication biases. After preprocessing, the dataset size was slightly reduced from 14,541 to 14,444 sentences. The final distribution consisted of 10,291 non-hedge sentences and 4,153 hedge sentences, preserving the original class imbalance characteristics of the dataset. The dataset was split into training and testing sets using a stratified sample technique in order to assess the effectiveness of the suggested model. Specifically, 80% of the data was allocated for training, while the remaining 20% was reserved for testing. Stratification was applied based on class labels to ensure that the distribution of hedge and non-hedge instances remained consistent across both subsets.

3.3 BERT Model Architecture and Fine-Tuning

To perform hedge detection, a transformer-based classification model based on BERT was employed. Specifically, the pre-trained bert-base-uncased model was fine-tuned for binary classification. BERT is a bidirectional language model that captures contextual relationships between words by processing the entire sentence simultaneously. Given an input sentence, BERT produces contextual embeddings for each token, where the special classification token *[CLS]* is used as the aggregate representation of the entire sequence. The final hidden representation of the *[CLS]* token is passed to a fully connected layer followed by a *softmax* function to obtain class probabilities. Formally, the prediction can be expressed as [26]:

$$\hat{y} = \text{softmax}\big(W \cdot h_{[CLS]} + b\big) \qquad (1)$$

In formulation (1), $h_{[CLS]}$ represents the contextual embedding of the special classification token produced by the BERT model, which encodes the overall semantic representation of the entire input sentence. The weight matrix and bias vector of the fully connected classification layer are represented by the parameters W and b, respectively, which are discovered during the fine-tuning procedure.

The function $\text{softmax}(\cdot)$is applied to transform the linear output into a probability distribution over the predefined classes, where each value in $\hat{y}$ indicates the predicted likelihood of the sentence belonging to a specific class (hedge or non-hedge).

Using supervised learning on the labeled dataset, the model was refined. The training process was implemented using the Hugging Face Transformers Trainer API with the PyTorch backend. The objective of the model is to minimize the cross-entropy loss between the predicted labels and the true labels, which is defined as:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (2)$$

Where y_i represents the ground truth label, $\hat{y}_i$ denotes the predicted probability for the corresponding class, and N is the total number of samples. The model parameters were optimized using the Adam optimizer with a learning rate of 2×10^{-5}, this is because the optimization algorithm has proven its efficiency and stability in handling the extensive language representations. The following table gives an overview of the training parameters employed in this experiment.

Table 1. Configuration of the BERT-Based Hedge Detection Model

Parameter	Value
Model	BERT-base
Max Length	128
Batch Size	16
Learning Rate	2×10^{-5}
Epochs	3
Classes	2

The process pipeline is shown in Figure 2 for a more intuitive representation of the proposed architecture. As shown in Figure 2, this figure shows how the process goes step by step. Starting with the BioScope data set, the process moves through the XML parser, preprocessing, and data split. This data is further tokenized with the BERT tokenizer and input into the BERT classifier algorithm to predict whether the sentence belongs to hedge or not-hedge category.

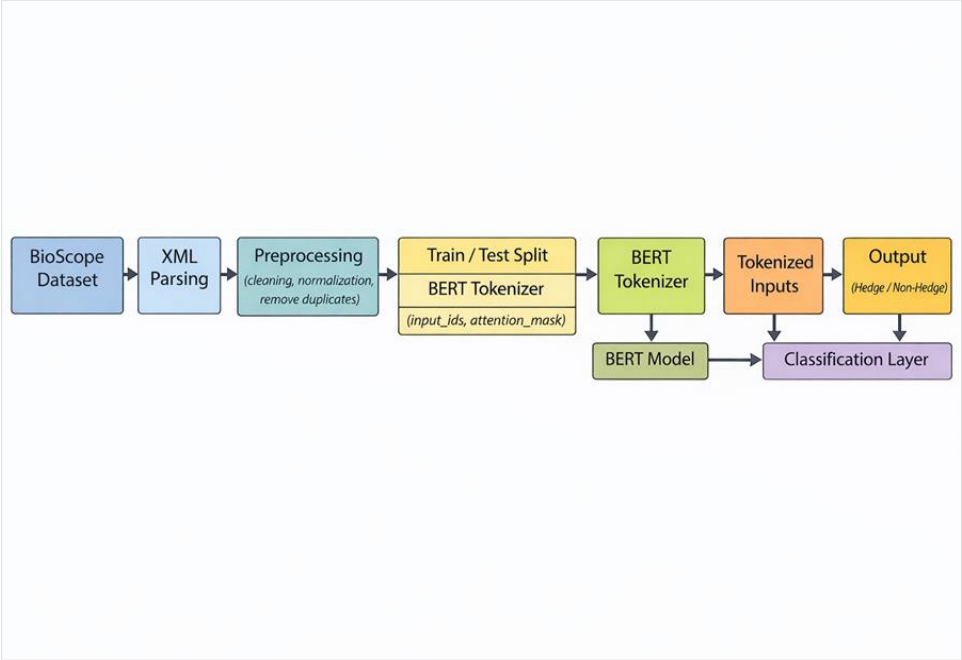

Figure 2. Block diagram of the proposed BERT-based hedge detection framework.

3.4 Evaluation Metrics

The performance of the suggested approach was measured using several metrics, including accuracy, precision, recall, and F1 score. All these metrics are used to evaluate the quality of the model in terms of differentiating between hedges and non-hedges. The accuracy metric refers to the correct functioning of the model. The formula for measuring the accuracy is [27]:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Precision evaluates the correctness of positive (hedge) predictions:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

Recall reflects the model's ability to correctly identify hedge sentences:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

The F1-score provides a balance between precision and recall and is defined as:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

where TP, TN, FP, and FN stand for true positives, true negatives, false positives, and false negatives, respectively.

4. RESULTS

In this part, we demonstrate the results of the experiments carried out using the developed BERT-based hedge detection model. These experiments were conducted using the test data set. The accuracy, precision, recall, and F1 score have been used as performance measures to evaluate the model. The created model's performance measure after tweaking is high for every metric, as can be seen below.

Table 2. Performance Evaluation Results of the Proposed BERT-Based Model

Metric	Value
Accuracy	96.92%
Precision	93.34%
Recall	96.15%
F1-score	94.72%

The results indicate that the proposed model has an accuracy rate of 96.92%, which is indicative of its robustness for classifying hedge sentences. On the other hand, the precision rate of 93.34% signifies that the majority of the hedge sentences predicted by the model are accurately detected.

Additionally, the high F1 score of 94.72% indicates the balance between the precision and recall scores, which is very important bearing in mind the class imbalance in the data. These results clearly show that the approach used for this task was highly effective.

As an additional measure for evaluating the effectiveness of classification, the confusion matrix is provided in Figure 3. From the results, it can be noted that 799 hedge sentences and 2001 non-hedge sentences were accurately detected by the classifier. In addition, few mistakes were made, where only 57 non-hedge sentences were wrongly classified as hedges, and 32 hedge sentences were wrongly classified as non-hedges.

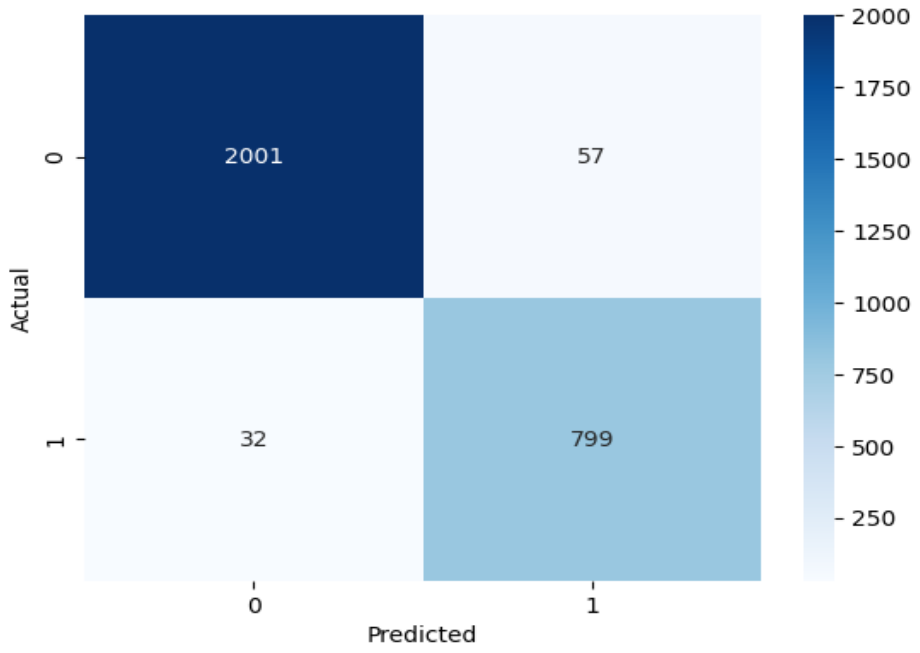

Figure 3. Confusion matrix of the proposed BERT-based hedge detection model.

In light of the fact that the number of false negatives is relatively small, these findings clearly indicate that the model exhibits a good trade-off between sensitivity and specificity, an attribute that is extremely important in hedge detection applications, where failure to recognize ambiguous sentences might be detrimental to the overall application. From the standpoint of evaluating the model's ability to discriminate, the next figure presents the ROC curve for this case (Figure 4). As can be seen from this figure, the TPR and FPR for different classification levels are presented by the ROC curve.

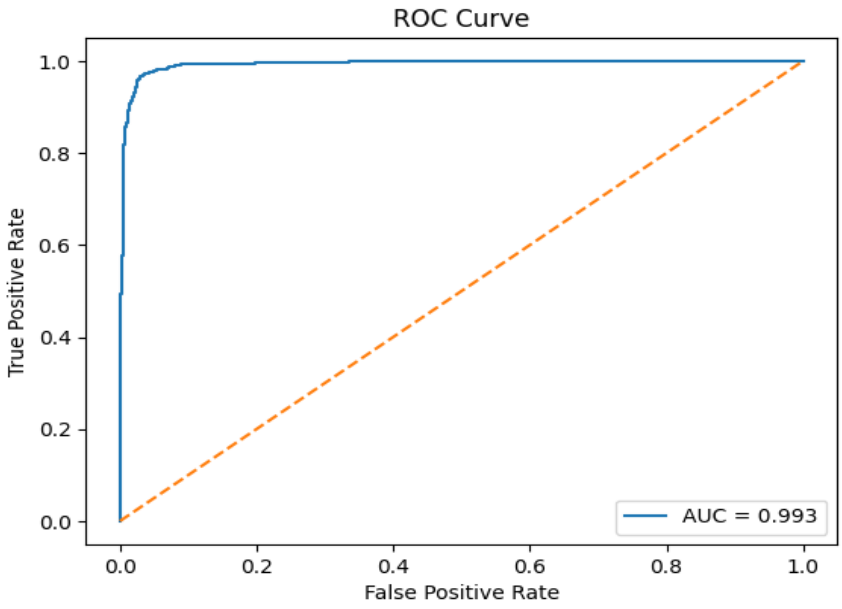

Figure 4. ROC curve of the proposed BERT-based hedge detection model (AUC = 0.993).

As illustrated in the above graph, the AUC of the model is 0.993, which implies that the model has performed very well in terms of classification. The graph lies near the top left-hand corner, which means that the true positive rate is very high while the false positive rate is relatively low. In general, the ROC curve once again proves that the proposed BERT-based model is effective and reliable for the task of detecting hedges.

To prove the applicability of the proposed model, a qualitative assessment was performed using a few sentences. This qualitative assessment was based on the usage of sample sentences that include hedge phrases as well as non-hedge phrases.

The proposed model successfully classified sentences into either the category of hedge or non-hedge. It should be noted that sentences containing uncertainty phrases like "may" and "might" were correctly classified as hedge phrases. The results of qualitative assessment can be seen in Table 3.

Table 3. Example Predictions of the Proposed Model

Sentence	Prediction
The results may indicate a possible improvement.	Hedge detected
The experiment confirms the hypothesis.	No hedge detected
It seems that the treatment might be effective.	Hedge detected
The algorithm achieves high accuracy.	No hedge detected

Qualitative analysis shows that the proposed model is robust and has the ability to generalize in dealing with unknown datasets. It has demonstrated the capacity to recognize indicators of uncertainty and differentiate them from facts. Taken together with the excellent quantitative results and ROC curve analysis, this implies that hedging in scientific documents is efficiently identified using the proposed BERT technique.

5. CONCLUSION

In the current investigation, a BERT model was introduced to automate the identification of hedge structures used in scientific articles. Specifically, by utilizing contextual features, the developed model successfully recognized the difference between hedges and non-hedge sentences. Results obtained through experiments indicate that the proposed method yields a good performance level based on several evaluation criteria, demonstrating its efficiency in recognizing linguistic structures associated with uncertainty. These results confirm the advantage of using transformer models in recognizing speculative language, especially considering its importance in various applications related to information retrieval, knowledge discovery, and decision-support systems. In spite of class imbalance present in the dataset, high levels of model generalization were achieved based on both quantitative and qualitative analysis. In conclusion, this paper proves that deep contextual models can be effectively used for the purpose of hedge detection.

6. FUTURE WORK

Further work will concentrate on developing the proposed methodology into a full-fledged multi-task framework for analyzing scientific text data. Although the present study focuses on addressing hedge detection as an independent problem, future studies will try to introduce other linguistic and semantic analysis tasks to construct a complete model. Specifically, future studies will try to introduce context-based analysis and named entity recognition in order to gain insight into the domain-specific information in the text. Additionally, intent classification will be addressed as a way to determine the functional aspect of sentences used in scientific contexts. Furthermore, future research may consider introducing sentiment analysis tools in order to get a sense of the underlying tone and communicative intent of the text data.

REFERENCES

[1] Canale, M., & Swain, M. (1980) Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1 (1), 1-47.
[2] Zadeh, L. A. (1972). A fuzzy-set-theoretic interpretation of linguistic hedges. *Journal of Cybernetics*, 2, 4-34.
[3] Lakoff, G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Concepts*, 2, 458-508.
[4] Farkas R, Vincze V, Móra G, Csirik J, Szarvas G (2010) The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. Proceedings of the Fourteenth Conference on Computational Natural Language Learning—Shared Task. Uppsala, Sweden: 1–12.
[5] Prince, E., Frader, J., & Bosk, C. (1982). On hedging in physicianphysician discourse. In R. J. Di Pietro (Ed.) Linguistics and the professions. Norwood: Ablex. 83-97.
[6] Viola Ganter and Michael Strube, "Finding hedges by chasing weasels: Hedge detection using wikipedia tags and shallow linguistic features," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2009, pp. 173–176.

[7] Maria Georgescul, "A hedgehop over a max-margin framework using hedge cues," in *Proceedings of the 14th International Conference on Computational Natural Language Learning: Shared Task*, 2010, pp. 26–31.

[8] Vladimir Vapnik, "The support vector method of function estimation," in *Nonlinear Modeling*, pp. 55–85. Springer, 1998.

[9] Morante R, DaelemansW(2009) Learning the scope of hedge cues in biomedical texts. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. Boulder, Colorado: 28–36.

[10] Velldal E, Øvrelid L, Read J, Oepen S (2012) Speculation and Negation: Rules, Rankers, and the Role of Syntax. Computational Linguistics 38: 369–410. doi: 10.1162/COLI_a_00126.

[11] Eunsol Choi, Chenhao Tan, Lillian Lee, Cristian Danescu-Niculescu-Mizil, and Jennifer Spindel, "Hedge detection as a lens on framing in the gmo debates: a position paper," pp. 70–79, 2012.

[12] Roma Patel and A. Nenkova, "Modeling ambiguity in text: A corpus of legal literature," 2019.

[13] Yoon Kim, "Convolutional neural networks for sentence classification," pp. 1746–1751, Oct. 2014.

[14] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[15] Heike Adel and Hinrich Schutze,¨ "Exploring different dimensions of attention for uncertainty detection," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 22–34.

[16] Manjira Sinha, Nilesh Agarwal, and Tirthankar Dasgupta, "Relation aware attention model for uncertainty detection in text," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 2020, pp. 437–440.

[17] Özgür A, Radev D R (2009) Detecting Speculations and their Scopes in Scientific Text. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Singapore: 1398–1407.

[18] Øvrelid L, Velldal E, Oepen S (2010) Syntactic Scope Resolution in Uncertainty Analysis. Proceedings of the 23rd international conference on computational linguistics. Beijing: 1379–1387.

[19] Apostolova E, Tomuro N, Demner-Fushman D (2011) Automatic Extraction of Lexico-Syntactic Patterns for Detection of Negation and Speculation Scopes. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon: 283–287.

[20] Zhou HW, Yang H, Huang DG, Li Y, Li LS (2013) Hedge Scope Detection Based on Syntactic Structural Constraints. Journal of Chinese Infromation Processing 27: 137–141.

[21] Morante R, Van Asch V, DaelemansW(2010) Memory-Based Resolution of In-Sentence Scopes of Hedge Cues. Proceedings of the Fourteenth Conference on Computational Natural Language Learning— Shared Task. Uppsala, Sweden: 40–47.

[22] Collins M, Duffy N (2002) Convolution Kernels for Natural Language. Advances in neural information processing systems 14. Cambridge, MA: MIT Press, 625–632.

[23] Zhou HW, Huang DG, Li XY, Yang YS (2011) Combining Structured and Flat Features by a Composite Kernel to Detect Hedges Scope in Biological Texts. Chinese Journal of Electronics 20: 476–482.

[24] Zou BW, Zhou GD, Zhu QM (2013) Tree Kernel-based Negation and Speculation Scope Detection with Structured Syntactic Parse Features. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: 968–976. A Effective Method for Hedge Scope Detection PLOS.

[25] Vincze, V., Szarvas, G., Farkas, R., Móra, G., & Csirik, J. (2008). "The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes." *BMC Bioinformatics*, 9(11), 1-9.

[26] Gonzalez-Carvajal, S., & Garrido-Merchán, E. C. (2023). "Comparing BERT against traditional machine learning text classification." *IEEE Access*, 11, 25887-25893. doi: 10.1109/ACCESS.2023.3256423.

[27] Grandini, M., Bagli, E., & Visani, G. (2020). "A metrics suite for network-based multi-label classification." *arXiv preprint arXiv:2008.05756*.